Incorporation of Protein Flexibility in
Structure-Based Drug Discovery of
HIV-1 Protease and Bacterial Hsp70 Chaperone

by

Man-Un Ung

A dissertation submitted in partial fulfillment
Of the requirements for the degree of
Doctor of Philosophy
(Medicinal Chemistry)
In The University of Michigan

2013

Doctoral Committee:
    Professor Heather A. Carlson, Chair
    Professor Hashim M. Al-Hashimi
    Associate Professor Jason E. Gestwicki
    Assistant Professor Oleg V. Tsodikov
    Professor Shaomeng Wang

DEDICATION

  I cannot be more grateful to have the support of my mother, Linda Wong. Even though we were never financially well off, my lovely mother encouraged me to take the opportunity to advance my education in the United States. My mother has worked tirelessly to support my education, and has endured the separation of twelve long years because of it. I would never be who I am today without my mother, who has sacrificed so much and asked for nothing in return. There will never be enough "Thank You"s to express my gratitude and love to my mother.

ACKNOWLEDGMENTS

my best years at the Cosumnes River College. I also want to give my gratitude to Richard J. Seifert for hosting me during my early years in Sacramento, CA. There are too many people I have met along the way, and too many people I am indebted to and want to thank.

From the day I arrived in the United States – with only a passport, a few hundred dollars, a backpack, and two pieces of luggage – to earning a degree in the Doctor of Philosophy in Medicinal Chemistry, a goal I have since high school, it has been a very long journey. This thesis will mark an end to the chapter as a student at the university. However, as a student of science in pursuit of scientific knowledge, this will only be the first chapter; there are many more exciting sciences I look forward to learning, experiencing, contributing to, and writing about in the next chapter.

TABLE OF CONTENTS

                    Small Molecules and Drug Discovery
                    Problem: "Static" Structure-Based Drug Design
                    Protein Flexibility
                    Allostery and Protein Flexibility

                    Sampling Ensemble of Conformations
                    Accounting for Protein Flexibility in Structure-Based Drug
                        Discovery

                    Clinical Importance and Genome Plasticity of HIV
                    HIV-1 Protease: A Principle Target of AIDS Treatment
                    Resistance and New Mode of Inhibition

                    Biological Significance of Hsp70 Chaperone
                    Structure of *Escherichia coli* DnaK
                    Regulation of DnaK
                  Modulation of Hsp70

                    Molecular Mechanism
                    Molecular Dynamics

LIST OF FIGURES

xii

xiv

LIST OF APPENDICES

ABSTRACT

Dynamics of a protein structure can be obtained from several sources, such as multiple crystal structures, NMR data, and computational simulations. The Carlson group has demonstrated that molecular dynamics simulations can provide robust protein flexibility information for dynamic SBDD. Using multiple protein structures (MPS) generated from dynamics simulations, protein conformations can be described as an ensemble of protein structures. This ensemble of protein states can be used in the identification of binding hot spots on protein surface and in the screening of potential small molecules binders.

Here, we have documented the application of MPS-based method in the search of small molecule inhibitor of HIV-1 protease (HIV-1p). Damm *et al.* utilized a MPS-based method and discovery an allosteric site of HIV-1p that had not been described before. Based on this new site, a MPS-based pharmacophore model was built to screen for potential HIV-1p allosteric binder, in which several new scaffolds were identified. Markush searches based on successful inhibitors were used to extend the chemical space of lead compounds. Subsequently biochemical and kinetics experiments were carried out to fully describe the mode of action of the allosteric inhibitors.

An alternative MPS-based method is mixed-solvent molecular dynamics (MixMD), which can be used to identify binding hot-spots on protein surfaces. MixMD fully samples the conformational space of the protein while uses small organic solvents

ABSTRACT

Dynamics of a protein structure can be obtained from several sources, such as multiple crystal structures, NMR data, and computational simulations. The Carlson group has demonstrated that molecular dynamics simulations can provide robust protein flexibility information for dynamic SBDD. Using multiple protein structures (MPS) generated from dynamics simulations, protein conformations can be described as an ensemble of protein structures. This ensemble of protein states can be used in the identification of binding hot spots on protein surface and in the screening of potential small molecules binders.

Here, we have documented the application of MPS-based method in the search of small molecule inhibitor of HIV-1 protease (HIV-1p). Damm *et al.* utilized a MPS-based method and discovery an allosteric site of HIV-1p that had not been described before. Based on this new site, a MPS-based pharmacophore model was built to screen for potential HIV-1p allosteric binder, in which several new scaffolds were identified. Markush searches based on successful inhibitors were used to extend the chemical space of lead compounds. Subsequently biochemical and kinetics experiments were carried out to fully describe the mode of action of the allosteric inhibitors.

An alternative MPS-based method is mixed-solvent molecular dynamics (MixMD), which can be used to identify binding hot-spots on protein surfaces. MixMD fully samples the conformational space of the protein while uses small organic solvents

as probes to compete with water on the protein surface. Using HIV-1p as a test case, MixMD successfully identifies the allosteric site described by Damm *et al*. Other known and proposed sites were also identified. Unusual "false positives" were actually identified as the common protein-protein interfaces utilized in crystal packing.

MPS has also been applied to heat shock protein 70 (Hsp70), a chaperone that is highly conserved among different species. Utilizing the ensemble of conformations generated through molecular dynamics simulations, potential hinge residues of Hsp70 were identified. Mutagenesis of these hinge residues has been carried out and mutation of a key residue, Gly228, is shown to decouple the allosteric communication between its catalytic activity and substrate binding activity. Hence, this demonstrates the strength of MPS methods in identifying key structural elements in a protein.

Lastly, efforts have been made to evaluate existing SBDD software through the Community Structure Activity Resource (CSAR). Many scoring approaches were assessed and their correlation to experimental binding affinity ranged $R^2 = 0.14$-0.51, with no particular scoring approach having an overwhelming advantage over others. There is a need for CSAR to develop data sets of congeneric series with a range of different physical characteristics for improved training of computational methods.

CHAPTER 1

Introduction

1.1 Background

Small Molecules and Drug Discovery

In the past several decades, progress in genomics, proteomics, molecular biology, and chemical biology have accelerated the discovery and characterization of new biological macromolecules, such as proteins and nucleic acids. Some of these macromolecules may have therapeutic potential in the cure of certain disease states, such as bacterial infections, acquired immunodeficiency syndrome, Alzheimer's disease and cancers.[1-3]

With the knowledge of the biological target and structural information provided through structural biology, development of therapeutic agents for the diseases with rational drug designs, such as structure-based drug design (**SBDD**), become possible and they are now routinely applied to pharmaceutical and medicinal chemistry research.[4,5] Instead of screening for biologically active chemicals and optimizing them to become suitable pharmaceutics through trial-and-error,[6] as has been done in the past century

when structural information or even the biological target was unknown, SBDD pinpoints the target of research and allows the discovery and design of chemicals that have high affinity and selectivity against specific targets. In addition, with a comprehensive understanding of the three-dimensional coordinates of the macromolecules, SBDD provides an opportunity to target sites outside of the traditional binding site (i.e., orthosteric site), termed allosteric sites, thus expanding the possible pathway to modulate the functions of the targets.[7]

The main goals of SBDD are to identify the relevant regions of the target that are linked to its functions, searching for ligands that specifically bind to the said regions of the receptor and induce functional changes in the receptor, and optimizing the ligand-receptor interactions. To accomplish these goals, my research has encompassed several aspects of SBDD: (*1*) methodology development for receptor binding hot spot identification, (*2*) study of the effects of ligand-receptor interaction on receptor structure and functionality, (*3*) experimental testing and characterization of ligand that induces functional changes in the targeted protein through allosteric site, and (*4*) optimizing computationally predicted ligand-receptor affinity to experimentally determine ligand-receptor interactions.

Problem: "Static" Structure-Based Drug Design

Traditionally, researchers would take a single X-ray crystal structure or a homology model of the target protein to conduct SBDD to design ligands with high affinity and specificity against the target. This approach has seen some successes,[8,9] such as the design of inhibitors against the MDM2-p53 complex.[10] However, the use of a

single "static" structure that accounts for only one conformation may not be appropriate as most proteins have a rather dynamic structure, altering from one conformer into another.[11,12] As the late Dr. Jeremy Knowles remarked, "making a model of a horse from photographs does not necessarily tell us how fast it can run";[13] neither does a snapshot of a protein tell us the intrinsic flexibility of the protein. Using only one or a few protein conformations in SBDD may result in unsuccessful campaign to identify ligands with strong affinity and selectivity against the protein target. Hence, it is desirable to advance from "static" SBDD to "dynamic" SBDD by incorporating protein flexibility into SBDD.

Protein Flexibility

Proteins in solution are intrinsically flexible and can adapt a wide range of conformations at both the local and global level. These can be localized motions, such as side chain and backbone rotation of the residue, or global motions such as movements of subdomains, as shown in the top panel of Figure 1.1. As our knowledge of protein flexibility grows, the model of protein-ligand interactions has moved away from the lock-and-key and the induced-fit models.[14,15] The current model proposes that the protein exists in an ensemble of conformational states, in which the population of each of these conformers can be described by the thermodynamic distribution, as illustrated in the bottom panel of Figure 1.1. In the event of ligand binding, the ligand selectively binds to certain conformers of the ensemble and shifts the equilibrium, which repopulates the conformational states according to the amount of protein that has been "trapped" by the ligand.[16]
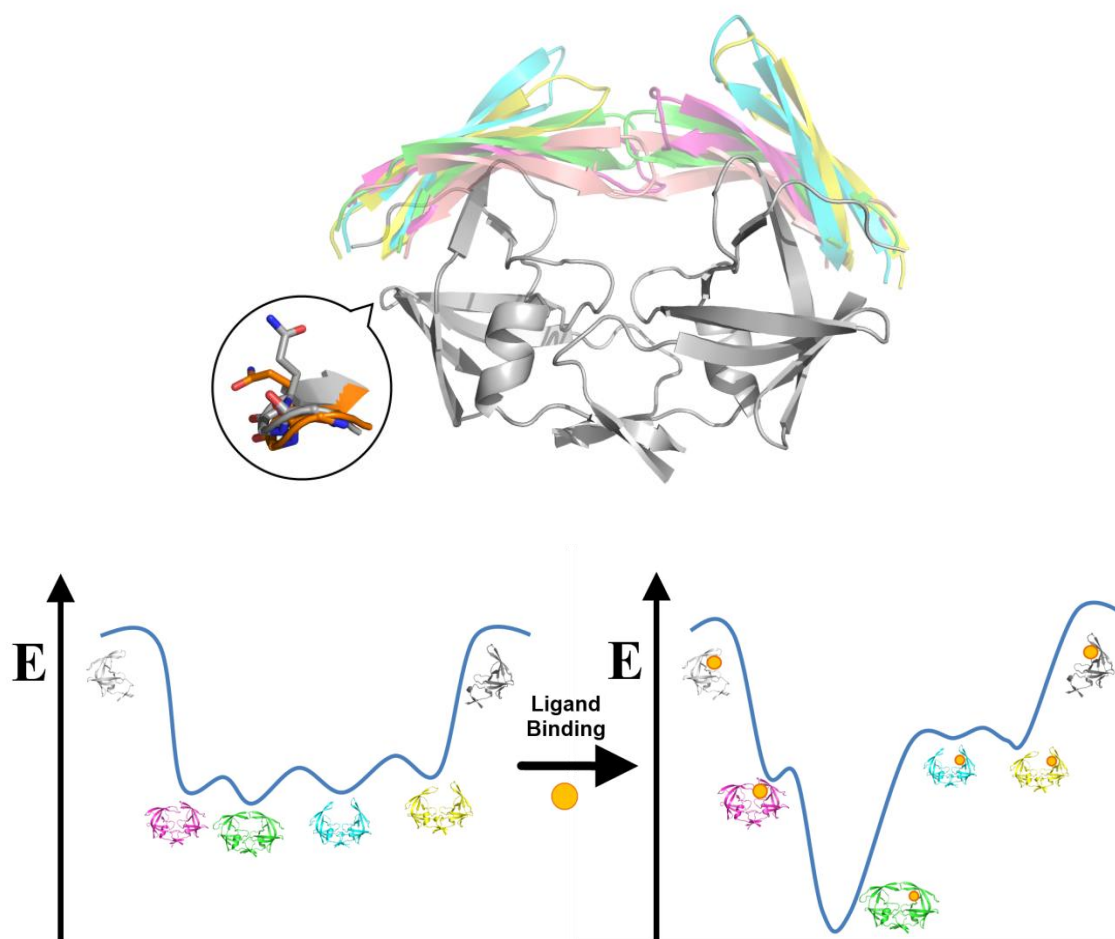
Figure 1.1. (Top) An ensemble of conformational states of HIV-1p. The flaps of HIV-1p, shown in various colors, encompass a wide range of conformations – closed, semi-open and open – and define the global conformational space of HIV-1p. The small insert illustrates the local fluctuations in the protein backbone and side chain (orange) commonly seen in proteins. **(Bottom)** Schematic diagram of the effect of ligand binding on a protein with multiple conformational states. The stable HIV-1 protease dimer is formed by two monomers and exists in an ensemble of conformational states. These conformation states are separated by energy barriers and their ratios depend on their positions in the free-energy landscape, as shown with protein in different sizes. When a ligand (orange orb) binds to the protein, it selectively binds to certain conformer and shifts the equilibrium of population, favoring a particular conformer and depopulates the other conformers.

As protein exists in equilibrium of differently populated conformational states, certain conformers may play crucial roles in protein functions, such as enzymatic activity and molecular recognition.[17,18] Ligands may bind to any of the conformers in the

conformational ensemble, regardless of the population of that conformer. This selective binding to a particular conformer in the ensemble perturbs the thermodynamics of the ensemble, shifts the equilibrium and alters the population distribution among the conformational states.[19] This shift of equilibrium may favor the catalytically active conformer and drive the enzymatic reaction of the bound ligand, or populates the conformational state with the right geometry for molecular recognition.[20]

Allostery and Protein Flexibility

Allostery is a phenomenon in which the binding of a ligand at a distal site affects the conformation and functional properties of the orthosteric site. Such distal site is called the allosteric site and the effector ligand is called an allosteric ligand. Since proteins are known to be intrinsically flexible and exist in an ensemble of pre-existing conformational states, the binding of a ligand to the allosteric site can induce a shift in the population of the conformational states, in which the conformer selectively populated by the allosteric ligand may have an altered local conformation at the orthosteric site. This allosteric regulation is known to play important roles in the regulation of protein activity, such as turning "on" or "off" a signaling protein in a signal transduction pathway, inducing the motion in molecular motors, and enhancing or diminishing the catalytic activity of enzymes.

Nussinov *et al.*[11] proposed that through the binding of allosteric ligand to a specific conformational state and altering the distribution of the conformer population in an ensemble, allostery can be used as a mean to modulate the activity of the targeted protein. Such modulation has been demonstrated in the inhibition of HIV reverse

transcriptase, in which the non-nucleoside reverse transcriptase inhibitor, nevirapine, binds to an allosteric site and affects the catalytic activity of the orthosteric site.[21] Another example is the regulation of $GABA_A$ receptor activity by diazepam, a positive allosteric modulator that enhances the effect of GABA by binding to an allosteric site far from the endogenous GABA binding site.[22]

Allosteric regulation is an attractive way to modulate a protein as it does not compete with the native substrate and can be highly selective. However, designing allosteric ligands for a particular protein using SBDD is considered a challenging "high-hanging fruit"; manipulating the allostery requires not only an extensive understanding of the dynamics of the protein target but also methods that can produce useful information from the few existing "snapshots" of most proteins. Since an allosteric site may exist in only some of the conformational states, X-ray crystallography may not be able to capture the particular conformation with the allosteric site for use in the development of allosteric modulator. Hence, alternative methods are sought after to produce additional structural information.

1.2 Dynamic Structure-Based Drug Discovery

Sampling Ensemble of Conformations

SBDD requires high-definition atomic details of a receptor structure. Experimentally, this structural information can be obtained through X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. Advances in

structural biology and technologies allow faster and more accurate X-ray crystal and NMR structure determination, as demonstrated by the increase in number of three-dimensional protein structures deposited into Protein Data Bank each year.[23]

To exemplify the dynamic features of a protein, a collection of crystal structures could be used to represent the conformational ensemble of the protein. On the other hand, NMR method intrinsically provides ensemble information of the determined protein.[24] However, both methods have shortcomings. Damm *et al.*[25] suggests that many crystal structures of the same protein are strikingly similar, likely due to the use of similar crystallization conditions and data solving processes. Conversely, NMR is applicable only to relatively small proteins, usually under 300 residues, as the complexity of the NMR data grows exponentially with the number of residues. To overcome the limitations and to supplement X-ray crystallography and NMR methods, computational techniques such as molecular dynamics (MD) simulation and normal mode analysis have been developed and gained popularity in estimating the flexibility of a protein and generating the ensemble information. Damm *et al.*[25] has demonstrated that the accuracy of MD-generated conformational ensemble is comparable to NMR technique and better than X-ray crystallography, which can be affected significantly by crystal packing effect and crystallization conditions.

Using the snapshots of the protein taken from MD simulation, the Carlson group has developed schemes that incorporate multiple protein structure (**MPS**) information into the standard SBDD.[25-33] These methodologies are robust and accurate[27,31,33,34] (Ung, in preparation), and have gained popularity among researchers in the SBDD field.[35-38]

Accounting for Protein Flexibility in Structure-Based Drug Discovery

Molecular docking technique is routinely used in SBDD. Its main application is to elucidate the binding modes of a small molecule in a receptor, in which the static receptor structure is usually obtained from the X-ray crystal, NMR or homolog model. Several approaches have been developed to incorporate protein flexibility into molecular docking[39]: (*1*) "soft docking": the van der Waals radii of atoms are scaled down to ease the steric clashes and to allow slight overlap of the receptor and ligand volumes during docking to mimic induced-fit effect,[40,41] (*2*) representative structure: docking of ligand to an average, representative receptor structure,[42] (*3*) "on the fly" receptor flexibility: conformation of protein backbone and side chains are sampled during ligand docking,[43-46] and (*4*) MPS: an ensemble of static receptor conformations generated from MD simulation, CONCOORD simulation, or experimental sources (X-ray crystal or NMR structure) are used in ligand docking.[47-51]

In addition to the application in molecular docking, MPS scheme also finds applications in the dynamics pharmacophore model and binding hot spot identification. Classic pharmacophore model and binding hot spot screening utilize only one static protein receptor in the experiment. However, such limited approach would not recognize binding hot spots that are not represented in that particular static structure. In addition, this approach usually suffers from the appearance of numerous distracting local minima, or spurious binding hot spots. The application of MPS in the binding hot spot screening provides a better representation of the conformational ensemble and suppresses the transient local minima by combining consistently appearing binding hot spots.

Carlson *et al.*[26,27] developed a robust MPS-based dynamic pharmacophore method that accounts for protein flexibility. To represent the conformational ensemble of the receptor in MPS, structures from various sources, such as a collection of crystal structures, NMR ensemble, or snapshots taken from MD simulations are used. To probe the binding hot spots, Carlson *et al.* utilized the multi-unit search for interacting conformers (MUSIC) technique available in the molecular modeling program BOSS.[52] The receptor surface of each snapshot of the MPS is flooded with non-interacting molecular probes and these probes are subjected to Monte Carlo energy minimization simultaneously to locate the energetically favorable sites, i.e. binding hot spots. The favorable sites from each snapshot are overlaid to identify the consensus sites that consistently appear across all snapshots, as shown in Figure 1.2. The consensus sites are then converted into a simple dynamic pharmacophore model usable by conventional virtual screening software to identify potential new ligands and scaffolds. This MPS-based dynamic pharmacophore method has been refined and applied to several systems, including HIV-1 protease,[25,28,29,34] dihydrofolate reductase[30,32] and MDM2.[31]

Multiple solvent crystal structure (MSCS)[53-57] and NMR[58] methods have been developed to experimentally probe the binding hot spots with organic solvents. To simplify and make dynamic binding hot spot identification more applicable, Lexa *et al.*[33] has developed a computational analog of these experimental approaches, the mixed-solvent molecular dynamics (MixMD) method. MixMD incorporates probe mapping into molecular dynamics simulations and uses water-miscible organic solvents (e.g. acetonitrile) as hydrophobic probe to sample the receptor for binding hot spots. MixMD has several advantages:
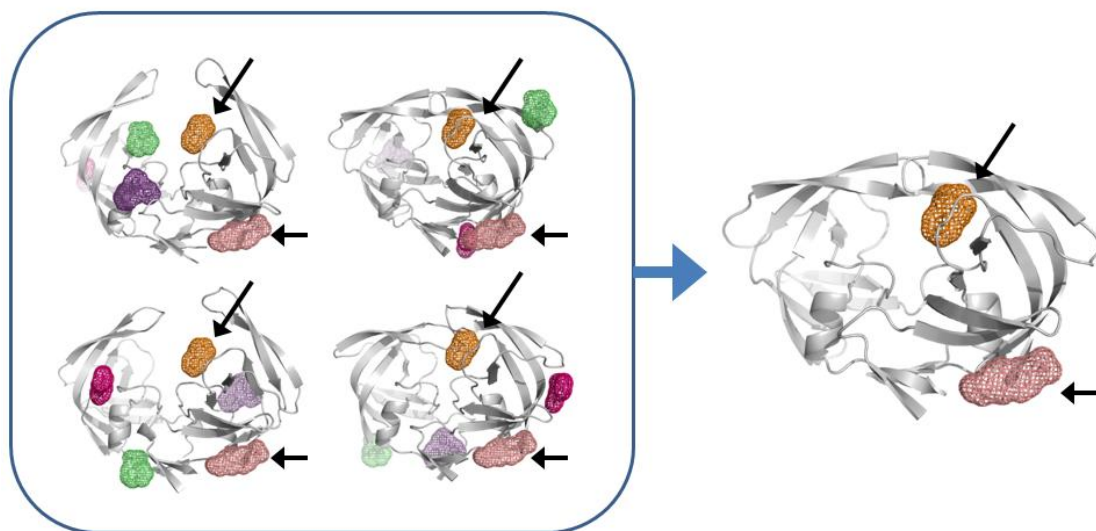
Figure 1.2. Schematic diagram of consensus analysis of data generated from MPS-based binding hot spot probing method. All analyzed frames of the ensemble are superimposed onto a common frame and the binding hot spots found on each frame are compared. Hot spots that consistently appear, shown by the arrows, are considered as the consensus binding hot spots.

1. comprehensive sampling of the receptor conformational space during solvent probing,

2. direct competition of probe and water molecules, and

3. readily available parameters for the organic solvents and thus no unphysical "tweaking" parameter is needed.

To find the consensus position of the probes, the positions of the probes are integrated over the simulation trajectory and are then converted into a three-dimensional occupancy map, which correlates to the probe occurrence on the protein surface. By comparing the MixMD and MSCS results on a model protein, the hen egg-white lysozyme,[55] Lexa *et al*. successful identified the substrate binding site as the binding hot spot while eliminated the spurious binding hot spots when MPS strategy is not applied. A similar approach, described by Bakan *et al*.,[59] uses isopropanol as the probe and converted the probe occupancy map into a binding free energy map to identify binding

hot spots. Another similar method, site identification by ligand competitive saturation (SILCS), has also been developed.[60-63] In contrast to MixMD, SILCS uses highly hydrophobic organic molecules (benzene and propane) as probes that require additional repulsive constraints to avoid aggregation of the probes and the method has a tendency to induce protein denaturation.[63]

1.3 HIV-1 Protease: A Principle Target of AIDS Treatment

The clinically important human immunodeficiency virus type-1 protease (**HIV-1p**) is used as the test case for the development of MPS-based methodologies. It has ample amount of NMR and X-ray crystallography data and its biochemical and physical properties are very well characterized, making it a suitable choice for our studies.

Clinical Importance and Genome Plasticity of HIV

Acquired immunodeficiency syndrome (AIDS), a relatively new disease first confirmed in the 1980s and declared a worldwide health issue by the World Health Organization, has since become a pandemic with an estimated 34 million infected cases globally.[64,65] This disease is caused by the infection of human immunodeficiency virus (HIV), a group of viruses from the family *Retroviridae* and the genus *Lentivirus*.[66,67] Two types of the virus have been characterized clinically: HIV-1 and HIV-2. HIV-1 is more infective and virulent[68] and is responsible for the majority of the infected cases globally, whereas HIV-2 is more closely related to the simian immunodeficiency virus

11

and is mostly confined in West Africa.[66,69] The viruses are transmitted through sexual contact, bodily fluids or tissues exposure, or vertical transmission from mother to child during pregnancy, delivery, or breast-feeding.[70] The virus primarily targets and kills the $CD4^+$ T lymphocytes of the T helper cells, compromising this vital component in the human adaptive immune system and thereby deteriorating the immune system as the infection progresses.[71-73] After an initial phase of acute flu-like illness following the exposure, the disease progresses into the clinical latency phase for $2 - 10$ years, in which the patient has a chronic, asymptotic HIV infection.[74-76] As the disease progresses, the immune system gradually loses the ability to confront common opportunistic infections and certain cancers, and eventually leads to the death of the patient.[77]

Although retroviruses are a diverse group of viruses that are found to infect a wide range of species, a prominent character of retroviruses is the genetic plasticity of the retroviral genome within a population.[78,79] The virus exists as a snowflake-like quasi-species and nearly every virion in a population differs from each other. This character has significant implication in the rapid rise of multi-drug resistant (MDR) strains of the virus under the selective pressure of anti-retroviral therapy. For HIV, the genetic diversity and plasticity may be the result of a combination of several factors:

1. High rate of replication.

2. Low fidelity of the HIV reverse transcriptase (RT) that lacks proof-reading and repair mechanisms.[80] Mutations propagate during syntheses of single-stranded viral DNA (vDNA) from viral RNA (vRNA) and the subsequent double-stranded DNA[81,82] for host genome insertion by HIV integrase.[83]

3. Frequent recombination of multiple viral genomes[84,85] as RT can complete DNA synthesis by polymerizing part of one vRNA and switch over to a homologous region on the copackaged vRNA template.[86,87] Clinically significant MDR strains can arise rapidly by the recombination of several existing single-drug resistant strains within a few replication cycles.[88,89]

4. Actions of host antiviral mechanisms.[78-90] Virally-encoded accessory proteins, such as viral infectivity factor (Vif) and viral protein R (Vpr), are found to target the vDNA-modifying cytidine deaminase APOBEC3G[90-92] and the DNA repairing machineries,[93] respectively. They modulate the antiviral effects of these host proteins while allowing the host-mutated vDNA to propagate and contribute to the genome diversity of the viral population.[94]

Properties, Structure and Dynamics of HIV-1p

HIV has been extensively studied in the past several decades and crucial viral-encoded products have been targeted for anti-retroviral therapeutic development. One of the viral products, the HIV-1p, has been the focal point of anti-retroviral research as it is critical in the maturation and propagation of the infective HIV viral particle. It promotes controlled autoproteolysis and release of mature protease from *gag-pol* polyprotein,[95,96] and cleaves the *gag* and *gag-pol* polyprotein precursors to release the structural proteins MA, CA, NC and p6, as well as the functional enzymes reverse transcriptase, integrase and protease.[97-99] As such, it has been characterized and extensively studied, and FDA-approved protease inhibitors have been developed.

The first crystal structure of HIV-1p, as shown in Figure 1.3, was solved in 1989 by two laboratories independently.[100,101] It was determined to be a C_2-symmetric homodimeric aspartyl protease that is comprised of two identical monomers of 99 amino acids each, in which the N- and C-terminal regions of each subunit interact to form the dimer interface. The substrate binding site is shaped at the dimer interface and each subunit contributes the catalytically active residues (Leu24, Asp25, Thr26 and Gly27) to form the "fireman's grip",[102,103] in which the Asp25 residues is the catalytic component of the active site. The formation and catalytic activity of HIV-1p homodimer is pH-dependent; K_d of dimerization is significantly lower at pH ~ 5.5 and has a longer dimerization half-life ($t_{1/2}$ ~ 0.5 hour) than at pH ~ 7.5, while catalytic activity is most optimal at pH ~ 5.0.[104,105]
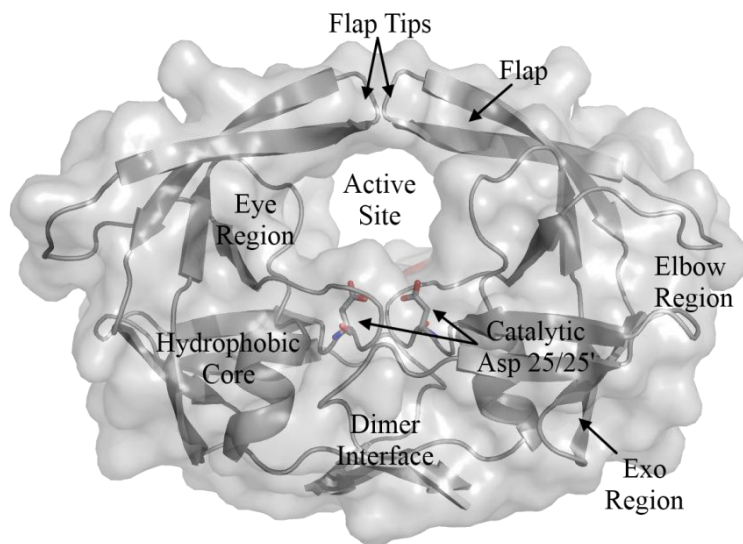
Figure 1.3. A cartoon representation with protein surface of HIV-1p in the semi-open conformation (PDB ID: 1HHP). Catalytic Asp 25/25' are shown in stick representation. Key regions of the protease are indicated.

The initial phase of the catalytic reaction involves the association of the substrate and the HIV-1p. The substrates can only access the catalytic site through the open

conformation of the protease.[106,107] The access and placement of the substrate in the catalytic site are controlled by the movement of a pair of flexible loops that serve as the gate to the orthosteric site of HIV-1p. This pair of glycine-rich, anti-parallel β-hairpins, one from each monomeric HIV-1p, is commonly referred to as the "**flaps**" (K45-M-I-G-G-I-G-G-F-I54). The mobility of the flaps is essential to HIV-1p activity as they control the access and positioning of substrate in the catalytic site during hydrolysis. Several studies based on X-ray crystallography,[101,108] electron paramegatic resonance,[109,110] NMR,[111-113] and various molecular dynamics simulations [28,107,114-117] suggest that the flaps of HIV-1p exist in an ensemble of conformational states and can adopt a range of conformations such as the closed, semi-open, and open forms.[118-120] Furthermore, simulations have shown that the flap movement is facilitated by the rearrangement of the hydrophobic core of the protease (residues 5, 11, 13, 15, 22, 24, 33, 36, 38, 62, 64, 75, 85, 89, 90, 93, 97).[121]

The conformational changes of the flaps can be substantial as the flap tips separation can be measured up to several tens of Å.[115,122,123] Due to the intrinsic flexibility of the flaps, apo protease can also exist in closed form[124] while ligand-bound protease can adopt the semi-open conformation.[125] NMR experiments of the apo protease showed that the dynamic equilibrium of the closed, semi-open, and fully open conformations of the flap occurs on a micro- to milli-second time scale with a faster flap tip fluctuations on a nanosecond time scale, while ligand-bound protease has flaps in the closed conformation with very limited flexibility except for the flap tip residues Ile50 and Gly51.[111,126,127]

In X-ray crystal structures, the closed conformation of the flaps is usually associated with the ligand-bound HIV-1p where the flap residues are in tight contacts with the bound ligand, while the apo structures adopt the semi-open conformation, in which the flap tips shift away from the binding site.[128,129] Although no experimental structure of HIV-1p with flaps in a fully open conformation has been observed, a crystal structure of a MDR mutant, PDB: 1TW7, has been observed with the flaps in partially open conformation.[130] However, the relevance of this partial open state in the flap equilibrium remains to be clarified, as subsequent molecular dynamics simulations have suggested that the 1TW7 structure is likely stabilized by crystal packing contacts.[131,132]

In addition to the various flap opening states of HIV-1p, the relative orientation of the flaps (Figure 1.4), i.e. the flap handedness, is observed to reverse upon the change from open/semi-open conformations to closed conformation.[115,128,133-136] This reorientation of flap handedness has been indicated as essential; it correlates to the formation of critical hydrogen bonds and the coordination of a structure water molecule between the flaps and the substrate,[137] and the correct placement of the substrate over the catalytic Asp 25/25' for optimal catalytic activity.[138]

Resistance and New Modes of Inhibition

Extensive researches have lead to several FDA-approved protease inhibitors (PIs) that specifically target the HIV-1p. Since the introduction of the first HIV-1 PI, saquinavir, in December 1995 and followed by other related inhibitors, these novel PIs have successfully slowed down the progress of the AIDS and have since been an integral component of the highly active anti-retroviral therapy, or HAART, in the management of

AIDS.[99,139] These PIs are divided into two major classes, of which 7 are peptidomimetic and 2 are non-peptidomimetic (Figure 1.5). Peptidomimetic PIs inhibit the protease competitively by mimicking the natural substrates or the transition state of substrate cleavage,[140] while non-peptidomimetic PIs displace a conserved structural water molecule that forms crucial hydrogen bonds and coordinates the substrates or peptidic inhibitors to the protease flap tips.[141]

(A)   **(B)**
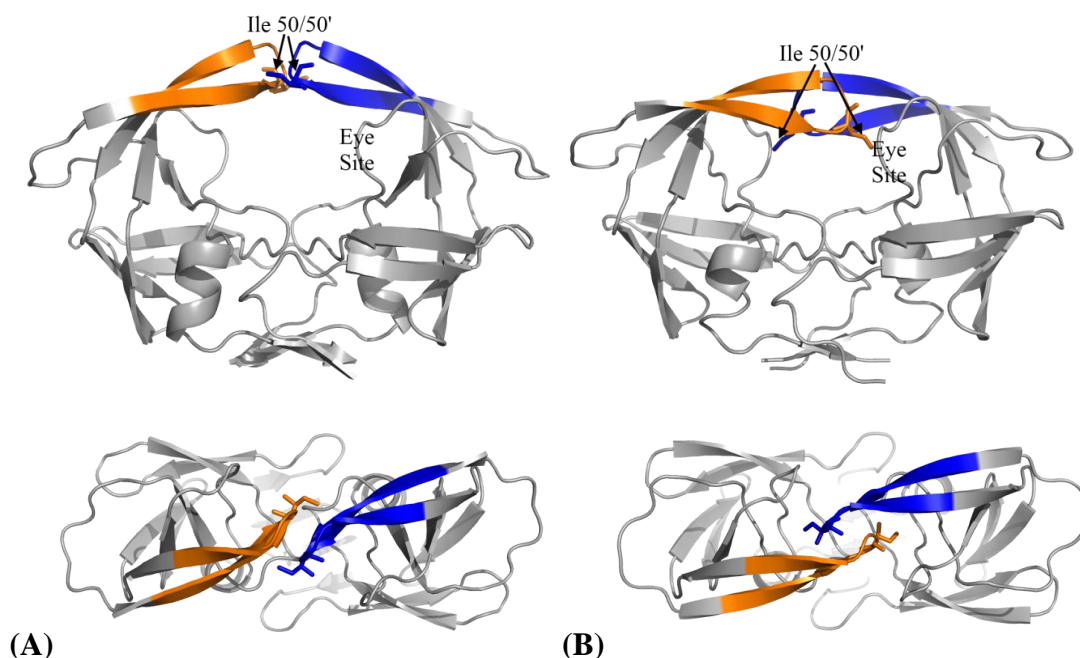
Figure 1.4. Configurations of HIV-1p flaps in semi-open and closed conformations. Flap regions (residues 45 − 54) are colored orange and blue for monomer 1 and 2, respectively. **(A)** HIV-1p in semi-open conformation. The flap tip residue Ile50 (displayed as stick) does not occupy the flap tip recognition site (Eye site) of the opposite monomer and is closer to the Eye site of the same monomer. **(B)** HIV-1p in closed conformation. The flap tip residue Ile50 occupies the Eye site of the opposite monomer.
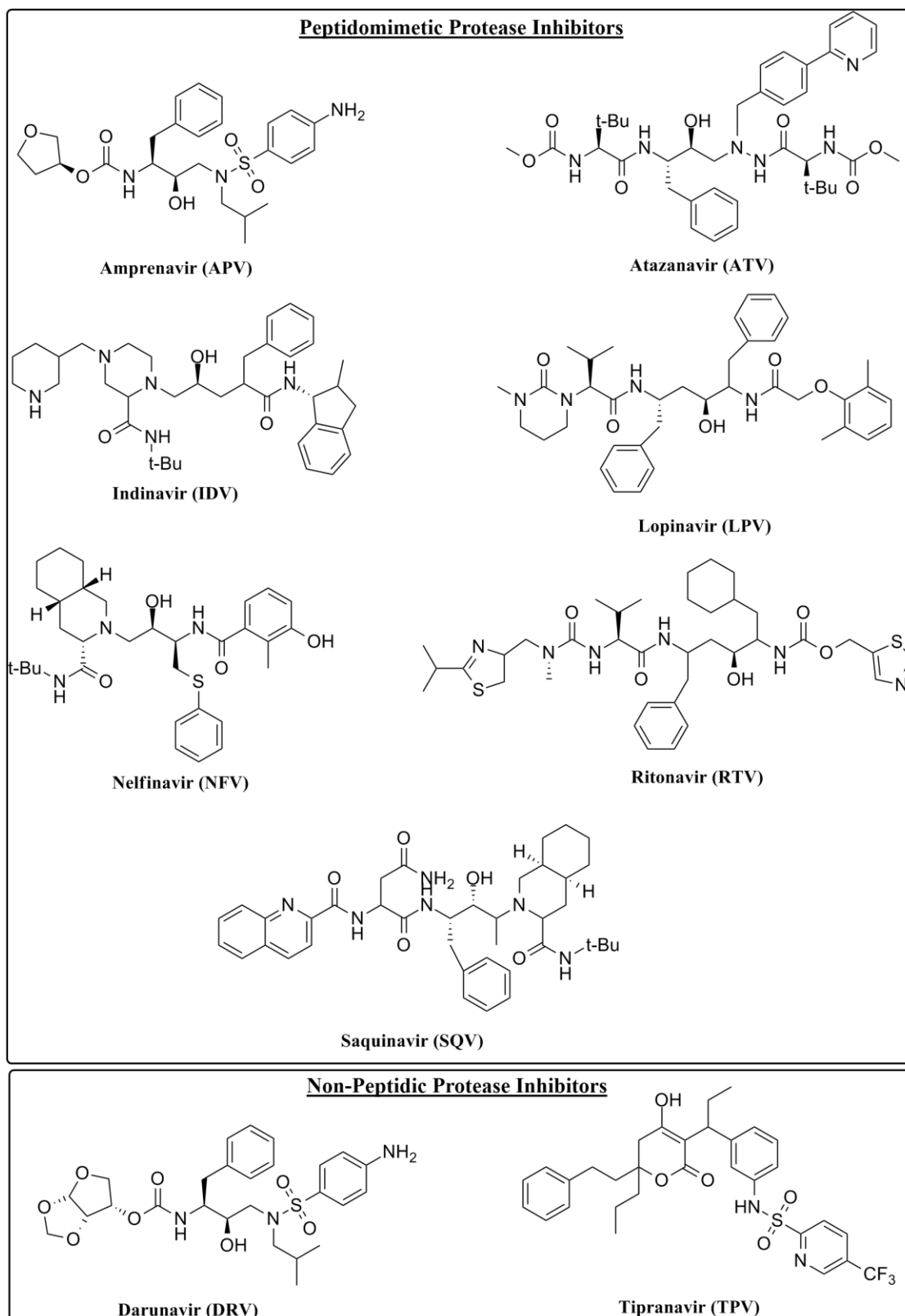
Figure 1.5. Structures of the HIV-1p inhibitors on the market.

Currently, all marketed PIs, both peptidomimetic and non-peptidic, act by targeting the catalytic site of HIV-1p.[142] However, under therapeutic pressure, HIV with the wild-type (WT) protease has rapidly evolved to strains that are resistant to multiple drugs through mutations.[143-147] Moreover, coevolution has been observed on the *gag* and *gag-pol* substrates, in which mutations of the substrates allow optimal binding with the mutant protease.[148-150] These mutations have rendered the currently available PIs ineffective and put patients at great risks of AIDS relapse. Hence, there is great urgency to develop new PIs that act through novel mechanisms that are least susceptible to the effects of mutations.

Extensive research and clinical results have shown that mutations in HIV-1p usually appear in the catalytic site. These mutations may favor the natural substrates over PIs by influencing the delicate balance of enzyme recognition events. This subtle change in PI binding can be a result of mutation of catalytic site or flag residues that are in direct contacts with PI,[151,152] or indirectly through altering the local environment of the catalytic site, such as polarity, sharp or size.[108] Mutations that are distal to the catalytic sites are also frequent; these mutations may distort the geometry of the catalytic site[153] or alter the flexibility and dynamics of the flaps.[138,154-158] Through understanding the drug resistant mechanisms that HIV evolves to evade the action of existing PIs, new PIs that act through novel mechanisms are being developed to combat the fast-evolving HIV. These new modes of inhibitions are summarized below.

Substrate Envelope Volume

Substrate recognition by HIV-1p is relatively unspecific as it does not strongly recognize specific primary sequence for substrate hydrolysis.[159] However, unlike the

promiscuous cellular aspartic protease pepsin, HIV-1p has a remarkably limited substrate range and cleaves only a very selective set of amino acid sequences.[160,161] By examining the crystal structures of protease-PI complexes, it is shown that substrates bind to the HIV-1p in an extended form and adopt a set of conserved three-dimensional shapes that are universal across all natural substrates. The consensus volume of these conserved shapes is termed as the "substrate envelope" and it is hypothesized to be the key element in substrate recognition and binding specificity.[108]

Further studies have validated the "substrate envelop" hypothesis and show that key resistant mutations of HIV-1p appear to cluster near locations where PIs protrude outside the substrate envelope volume.[162,163] This hypothesis explains the failures of the current PIs, in which most of them are shorter than the natural substrates and contain hydrophobic moieties that extend beyond the substrate envelope. It also explains the different mutant profile of the second-generation PI darunavir[146] and its effectiveness against clinically relevant protease mutants.[164,165] Further studies have applied this substrate envelope volume as a guide for next-generation PI development.[166-168] An updated substrate envelope, in which the dynamic features of the binding site in the form of probability distribution function, has also been developed.[150,169]

Dimer Interface

Protein-protein interfaces are popular therapeutic targets as they usually carry out critical functions.[170] At physiological pH, HIV-1p exists in an equilibrium of catalytically inactive monomer and catalytically active homodimer. During the formation of a dimer, two monomers interlink through non-covalent interactions of the four N- and C-terminal segments of the monomers (residues 1-4 and 96-99, respectively) that form a four-

stranded anti-parallel β-sheet at the dimer interface (Figure 1.3). This interlinked β-sheet stabilizes the dimer and is estimated to provide 75% of the dimerization stabilization energy.[171] In addition, studies have shown that the residues at the dimer interface are highly conserved in HIV-1 and in some HIV-2 isolates.[172-175] Because of the structural importance of this dimer interface, dimer formation has been studied extensively[176] and a novel method of inhibiting protease activity by disrupting the dimer formation has been developed.

Currently, development of dimerization inhibitors has mostly focused on peptide-based molecules, either with single-peptide molecules,[177-179] or double-peptide molecules that tether two peptides mimicking N- and C-termini.[180-185] These peptidyl inhibitors are generally very potent, and in some cases show "cross-inhibition" against other proteases, such as the human T cell leukaemia virus type-1 (HTLV-1) protease.[186] Recently, computational studies that examine the dynamics of the dimer interface have lead to the development of non-peptidyl small molecule inhibitors,[187] shown in Figure 1.6 A, which may potentially be more drug-like and allows better pharmacodynamics and pharmacokinetics properties that peptidyl inhibitors lack.

<u>Allosteric Sites</u>

The dynamics of the flaps have major influence on the catalytic activity of HIV-1p.[111-113,119,188] Various studies have suggested that drug resistant mutants usually have altered flap movement when compared to the WT protease.[110,114,158,189-191] Computational studies of the protease suggest that the movement of the elbow region of the flap (17-39
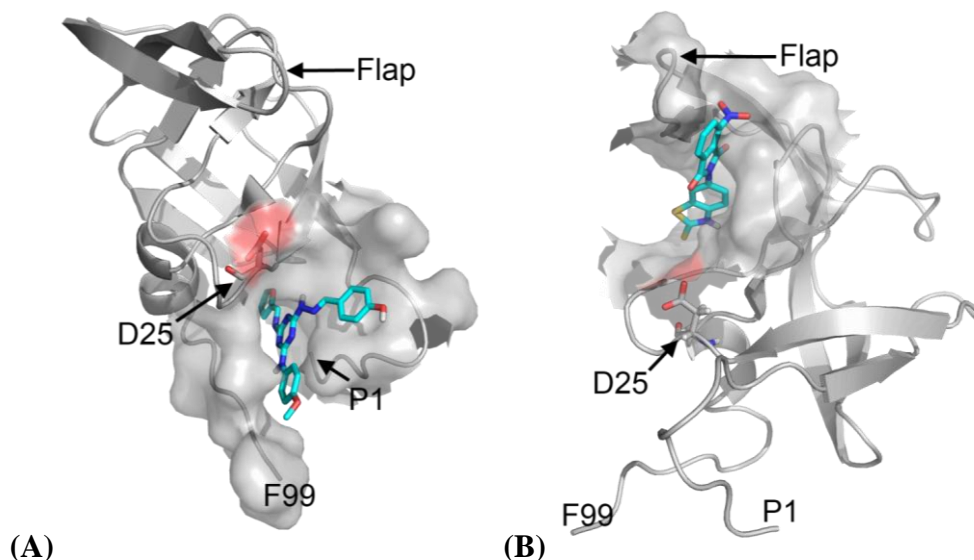
Figure 1.6. Inhibiting HIV-1p through non-substrate binding sites. **(A)** Docking pose of dimerization inhibitor Compound 8 in the β-sheet interface of a monomer (communication with Dr. Jerome J. Quintero). **(B)** Docking pose of allosteric inhibitor NIT in the Eye allosteric site under the flap region.

gap) is anti-correlated to the flap (25-50 gap), suggesting that modulating the movement of the elbow region may alter protease activity.[114,118,122,192] Experimental studies show that physically restraining the flap elbow movement through disulfide bridge diminishes proteolytic activity of HIV-1p, confirming the importance of flap movement in catalytic activity.[193]

In light of the importance of flap movement in HIV-1p, allosteric sites were sought that could modulate the flap movement. Using MPS-based dynamic pharmacophore method.[25,28-30] the Carlson group has proposed the existence of a previously undescribed allosteric site between the flap loop (residues 53-56) and the Face loop (residues 78-82), called the Eye site, in the protease.[34] This site is later confirmed by X-ray crystallography[194] and it is only observed in semi-open conformation, in which the flap handedness reverses upon closing of the flaps when substrate or inhibitor binds

(Figure 1.4 B). Damm *et al.* hypothesized that by selectively targeting the semi-open state of HIV-1p and interfering the proper movement of the flap, the protease catalytic activity will be adversely affected. Based on the dynamic pharmacophore method, Damm *et al.* constructed an Eye site pharmacophore model and applied it to virtually screen a chemical library, in which a small molecule was identified that inhibits HIV-1p proteolytic activity (Figure 1.6 B). A follow-up study identified another small molecule that modulates HIV-1p catalytic activity through a noncompetitive manner (Ung, in preparation).

In addition, Perryman *et al.*[114] proposed an allosteric site, originally called the "allosteric groove" and later renamed to "Exo site",[194] in a cleft formed between a loop (residues 15-19) and a β-strand (residues 59-63). A family of molecules was identified through high-throughput screening.[189] Kinetic studies show they do not act through the catalytic site as they do not compete with pepstatin A, a known active-site binder.[195,196] Through a docking experiment with protease in closed conformation, the authors hypothesized these molecules act by binding to the Exo site. However, the actual mode of inhibition of these molecules is currently unclear as computational studies with protease-ligand complex in semi-open conformation show the molecules can also bind to the Eye site (Ung, in preparation). In addition, although the unusual flap conformation of the protease in crystal structure 1TW7[131] is thought to be due to allosteric control through the Exo site, Lexa *et al.*[132] suggested that such observation was due to crystal packing and not a result of allosteric modulation through the Exo site.

1.4 Maid in the Cell: Hsp70 Chaperone

Biological Significance of Hsp70Chaperone

In a typical cell, the total protein concentration in the cytosol is usually in the range of 300 – 400 mg per mL.[197] In this environment, the volume available for biochemical reactions in solutions is limited by the high concentration of proteins. This macromolecular crowding promotes intermolecular interactions, in which exposed hydrophobic regions of unfolded proteins are prone to misfolding or aggregation,[198,199] resulting in a range of cellular problems and affect cell viability.[200] To maintain the protein homeostasis, or proteostasis, within a cell, all organisms have dedicated protein assemblies that regulate proteostasis by binding to the exposed hydrophobic regions of the unfolded proteins, or client proteins, to avoid protein aggregation and assist the clearing of the abnormal misfolded, aggregated proteins.[201-203]

One of the most abundant proteostasis regulators is heat shock protein 70 (**Hsp70**), a family of highly conserved 70-kDa molecular chaperones found in both eukaryotic and prokaryotic organisms.[204-207] Expression of this already abundant protein can be further induced under conditions of stress, such as temperature shift.[208,209] Hsp70 is involved in a wide range of cellular "house keeping" tasks: (**1**) it is responsible for the folding of nascent polypeptides[201] and refolding of denatured proteins, (**2**) facilitates disaggregation of amyloidogenic aggregates and deoligomerization of protein complexes,[210] (**3**) guides translocating proteins across organellar membranes,[211-213] and (**4**) assists proteasome-mediated degradation.[214,215] Furthermore, Hsp70 collaborates with

24

a range of other chaperones to regulate proteostasis,[215-218] such as the highly conserved Hsp90 family that is also found in both eukaryotic and prokaryotic organisms.[203]

As Hsp70 plays a role in many cellular processes, its modulations may have implications in various disease controls.[200,207,219] In human, overexpression of Hsp70 is associated with cancer survival[220,221] as it provides a survival advantage through its interaction with caspase-dependent and –independent apoptotic pathways.[222,223] Inhibition of Hsp70 removes this protection for cancer cells and is predicted to selectively induce apoptosis in cancer cells.[224] On the other hand, Hsp70 is linked to the clearance of toxic neurogenic plaques found in neurodegenerative disorders such as Alzheimer's disease, where Hsp70 and other chaperones prevent the formation and promote the degradation of β-amyloid fibrils[225,226] and neurofibrillary tangles of tau protein.[227-229] Hence, the modulation of Hsp70 activity has been suggested as a potential way to control neurodegenerative disorders.[230,231]

DnaK, a prokaryotic homolog that is ~ 50% identical (~ 78% similar) to eukaryotic Hsp70,[206] is a promising target for therapeutics. In addition to its function as a molecular chaperone in the cytosol, DnaK is also found on cell surface and has secondary function as virulence factor by "moonlighting" as adhesion.[232-235] Although it is not essential to cell growth under normal condition, cells with null mutations lose viability when challenged by stress, such as temperature,[236,237] immune-response,[235] and antibiotics.[238] It has been shown that inhibition of DnaK would sensitize clinically relevant pathogens such as *Staphylococcus aureus*[239] and *Klebsiella pneumonia.*[240] Hence, research and development of inhibitor targeting DnaK is being actively pursued.[241]

Recently, studies have suggested that Hsp70 and other molecular chaperones, such as Hsp60 and Hsp90, have extracellular roles where they function as intercellular signaling agents in immune and inflammatory responses.[242] In addition to cell lysis, release of these chaperones through endolysosome-mediated secretion has also been reported.[243,244] These chaperones have been observed to interact with cell surface receptors, such as Toll-like receptors, linked to innate immune responses[245,246] and induce immune responses.[247,248] Interestingly, cell surface receptors CD40[249] and CCR5,[250] both cytokine receptors, double as Hsp70 receptors. Binding of Hsp70 has shown to lower infectivity of HIV-1 virus through the stimulation of anti-viral, APOBEC3G-mediated response in $CD4^+$ and $CD8^+$ T cells.[251,252]

Structure of *Escherichia coli* DnaK

All members of the Hsp70 family are composed of 2 globular domains connected by a conserved flexible 5-residue interdomain linker[253,254]: an N-terminal nucleotide binding domain (**NBD**) and a C-terminal substrate binding domain (**SBD**), as shown in Figure 1.7. For *Escherichia coli* DnaK, a bacterial model of Hsp70, the NBD is a 44-kDa ATPase and has a fold similar to actin and hexokinase.[255,256] NBD consists of 2 globular subdomains, I and II, each can be further divided into A and B regions. NMR study has suggested movements of the subdomains take part in the catalytic activity and allow allosteric communication between the NBD and SBD.[257] A striking difference between Gram-(+) and Gram-(-) DnaK (and eukaryotic Hsp70) is the deletion of a small lobe in the I-B subdomain (residues 79-102 in *E. coli* DnaK) of Gram-(+) NBD.[258] The hydrophobic nucleotide binding pocket can be found at the intersection of all four
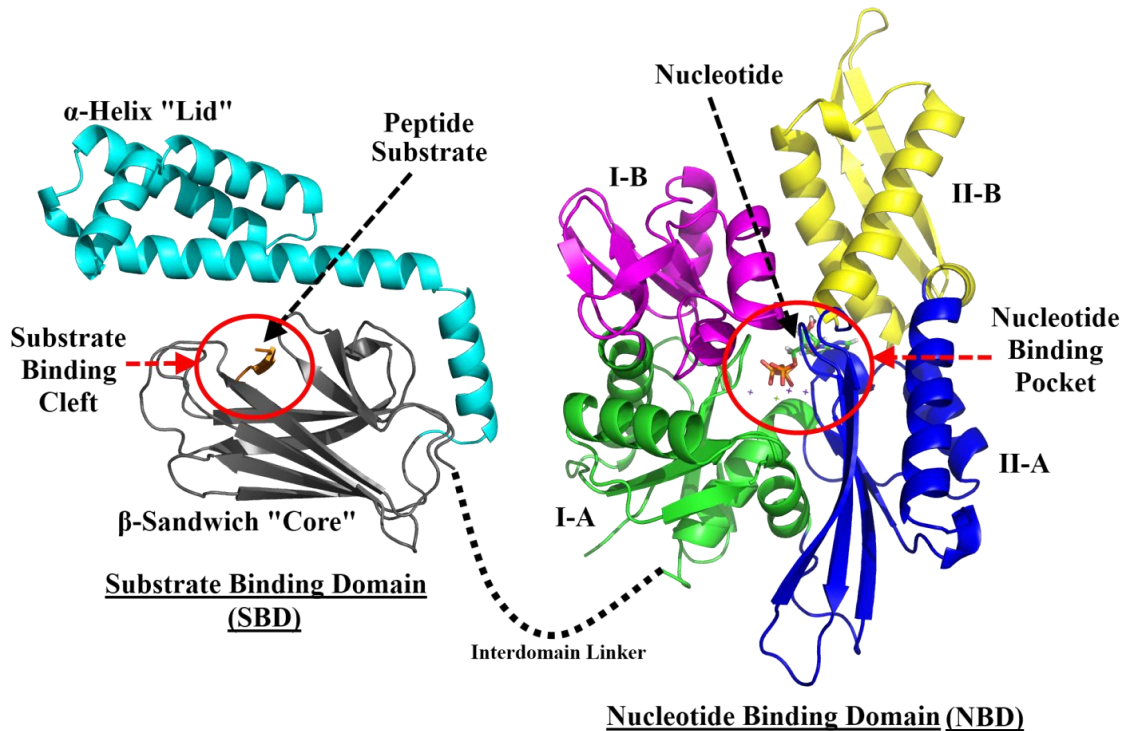
Figure 1.7. Architecture of *Escherichia coli* heat shock protein 70 (DnaK). The chaperone consists of two globular domains, substrate binding domain (SBD) and nucleotide binding domain (NBD) (PDB: 1DKG), linked by an interdomain linker that is usually unstructured. SBD consists of two subdomains, an α-helix "lid" (cyan) and a β-sandwich "core" (gray) that contains a substrate binding cleft. NBD consists of two globular subdomains I and II, for which each subdomain can be divided into two regions, hence subdomains I-A (green), I-B (purple), II-A (blue) and II-B (yellow). NBD has a nucleotide binding pocket located between the subdomains IA and IIA.

subdomains, in which 2 phosphate-binding loops, each from subdomains I-A and II-A, interact with the phosphate tail of the nucleotides; a Mg^{2+} cation is required for nucleotide binding and ATP hydrolysis, while two K^+ cations are essential for Hsp70 activation and client protein dissociation.[259,260]

The SBD is a 25-kDa domain consists of a 10-kDa α-helix subdomain that acts as a "lid", and a 15-kDa β-sandwich subdomain that acts as the "base". The client protein is bound in the cavity between the two subdomains.[261] SBD binds to the hydrophobic region of a linear nascent polypeptide, or a fully or partially unfolded client protein; SBD has

low sequence selectivity and thus can recognize a wide range of client proteins.[262] SBD preferentially binds to an extended hydrophobic core of 4-5 residues (> 10 Å from end to end) flanked by basic residues, where leucine is the most commonly recognized core residue and acidic residues are disfavored.[216,262,263]

Peptide binding of SBD is regulated by the NBD through an ATP-dependent mechanism. Studies suggest that the binding of ATP triggers conformational change in the NBD, which promotes the docking of the SBD to the NBD through a conserved interdomain interface.[216] This association may involve the interdomain linker, whose function as the communicator between the NBD and SBD is supported by crystallography, NMR, and mutagenesis studies.[264,265] When NBD binds ATP, conformational change induces the linker to pack to both NBD and SBD, which promotes the docking and interactions of the two domains. This docking alters the conformation of SBD and induces the opening of the "lid" subdomain. SBD in ATP-bound state has low affinity and fast exchange rate for the client proteins;[265] hence, it is associated with the release of bound substrate.[266] Conversely, ATP hydrolysis changes the NBD to ADP-bound state, in which the alternation triggers the dissociation of the interdomain linker from the domains, exposing the interlinker region and making it susceptible to proteolysis.[254,261,265,267] The undocking of the two domains promotes lid-closing of the SBD, strengthening substrate binding affinity and lowering the exchange rate for client proteins; hence, ADP-bound DnaK is associated with the binding of the client peptide. Mutations in the conserved region of the interdomain linker decouple the communication between NBD and SBD, in which the substrate binding activity of SBD does not respond to the change in the nucleotide-bound state of NBD.[268]

Regulations of DnaK

The spontaneous transition between the ATP-bound and ADP-bound states of DnaK is extremely slow. Without stimulation, *E. coli* DnaK has very low basal ATPase activity (0.002 – 0.2 ATP/min),[269] and is minimally stimulated by the client protein by ~ 1 – 5 folds.[268,270] However, its activity can be enhanced greatly by various partners: the cochaperone J proteins (e.g. **DnaJ**) facilitate ATP hydrolysis and client protein capture;[268] and nucleotide exchange factors (NEFs) (e.g. **GrpE**) stimulate the release of ADP and client.[271] Together, these partners regulate the heat shock response mechanism, as demonstrated in Figure 1.8.[218,272,273]

DnaJ and related co-chaperones have a conserved J-domain of ~ 80aa near the N-terminus[274,275] that interacts with DnaK.[276] The conserved C-terminus region may bind reversibly to unfolded client proteins and deliver them to DnaK that it interacts with.[218, 273,277-279] DnaJ binds to an area near subdomain II-A of a DnaK-ATP complex while some regions of DnaJ may bind at or near the DnaK substrate binding site.[280-282] Individually, DnaJ stimulates the rate of ATP hydrolysis by ~ 50-fold;[280,283] DnaJ and substrate synergistically stimulate the rate of ATP hydrolysis by > 1000-fold.[266,268,277] Hence, DnaJ promotes the tight binding of the protein substrate by DnaK.

GrpE binds to the DnaK-ADP-substrate complex and stimulates the exchange rate of ADP and ATP by 2500 – 5000 folds[284] but does not affect ATP hydrolysis rate.[266] This exchange induces the release of substrate held in ADP-bound DnaK by reversing the SBD into the low-affinity/fast-exchange ATP-bound state. The released substrate will either be folded or unfolded, or is transferred to chaperones such as Hsp90[285,286] or chaperonin.[216] Together, DnaJ and GrpE act synergistically and enhance ATPase activity

29
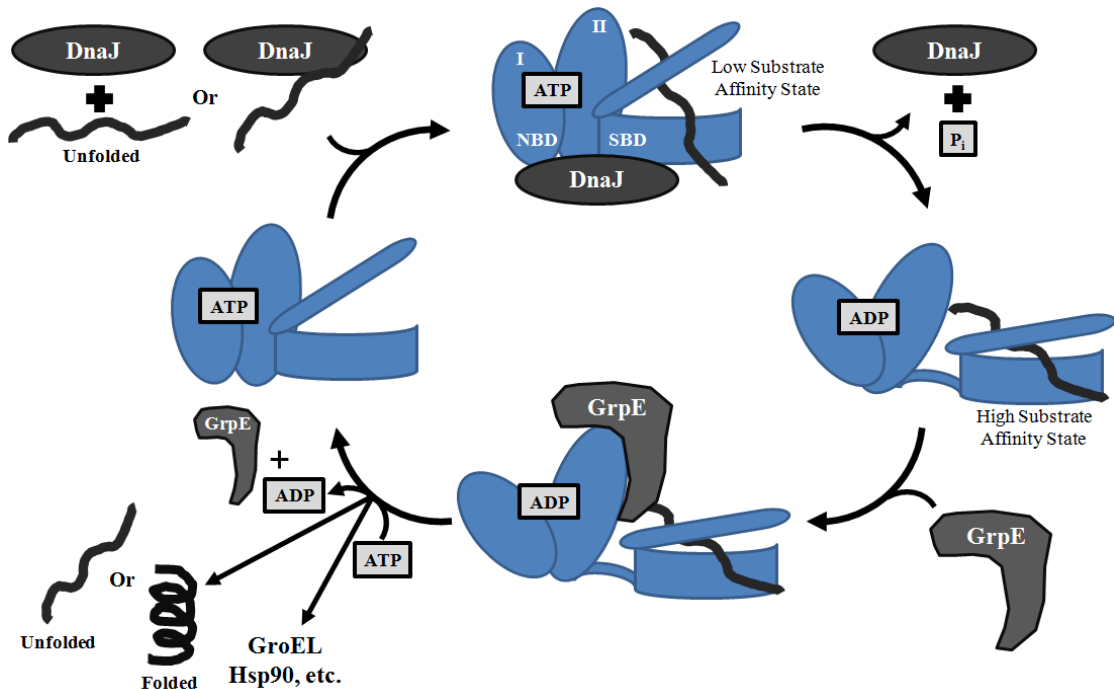
Figure 1.8. ATP hydrolysis and substrate binding/release are coupled in the DnaK-DnaJ-GrpE complex. Binding of the cochaperone DnaJ, which may carry and deliver the client protein to DnaK, works in concert with the binding of client in the SBD to promote ATP hydrolysis. In ADP-bound state, the lid of SBD closes and SBD has an enhanced binding affinity for the client. Association of nucleotide exchange factor, GrpE, to the NBD promotes ADP dissociation and ATP binding, which induces the opening of the SBD lid and the release of the bound substrate. DnaK may transfer the client to another chaperone system, e.g. GroEL or Hsp90 for refolding, or release the client in unfolded or folded form.

by > 50-fold.[266,268,273] The DnaK-DnaJ-GrpE chaperone system functions in an ATP-regulated cycle of substrate binding and release, captures the short hydrophobic segments in unfolded client proteins and prevents further unfolding or aggregation by shielding these segments. This system collaborates with other chaperones that refold, e.g. GroEL-ES, or degrade, ClpB, the client proteins and thus regulates the protein quality within the cell and maintains the cell viability during conditions of stress.

In vivo, DnaJ and GrpE are present in sub-stoichiometric amounts relative to DnaK. It has been demonstrated *in vitro* and *in vivo* that the refolding capacity of DnaK

diminishes when DnaJ or GrpE rises above or below the optimal ratios to DnaK; hence, the relative amounts of DnaJ and GrpE and their interactions with DnaK can be exploited to positively or negatively control the Hsp70 machinery[284,287-289] for therapeutic use.[290-292]

Modulation of Hsp70

The Hsp70 chaperone machinery provides a safe haven for proteins during times of stress through stabilizing partially or fully unfolded proteins and prevents aggregation that may threaten the survival of the cell. At times of stress, Hsp70 is overexpressed to compensate the adverse conditions, whereas down-regulation of Hsp70 results in destabilizing the delicate proteostasis. It is known that tumor cells have elevated level of Hsp70;[292] down-regulation of Hsp70 promotes apoptosis of tumor cells.[293] Similarly, down-regulation of Hsp70 sensitizes bacteria to antibiotics.[240,294] Hence, Hsp70 may have therapeutic implications and its modulation has been explored. Extensive crystallography, NMR, and mutagenesis studies of Hsp70 have been carried out to better understand its structure and allostery. There is interest in what conformational change takes place in both SBD and NBD when the ATP is hydrolyzed. It is of great interest to understand how the cross-talk between the two domains happens and which residues are involved.

Served as the surrogate of the mammalian Hsp70 chaperone system, the *E. coli* DnaK chaperone system has been studied extensively as its DnaK-DnaJ-NEF system is simpler. Six DnaJ homologs have been identified in *E. coli*, as compare to 22 in *Saccharomyces cerevisiae* and over 50 in human.[218,295,296] Conversely, only one GrpE-like NEF is found in bacteria and mitochondria; the eukaryotes have many vastly diverse NEFs, e.g. Bag family of NEFs, and Hsp70-like NEFs (Hsp110).[218,297,298] Through "gray

box" high-throughput screening campaigns that constitute DnaK and its DnaJ and NEF partners, the Gestwicki group has discovered several classes of compounds that modulate Hsp70 activity through allosteric regulations of the NBD.[241,299-301] These compounds do not have strong activity when added to DnaK or DnaJ alone; the modulation manifests only in the presence of both DnaK and DnaJ. NMR chemical shift data of these compounds indicates they act on different regions of the NBD. Hence, novel modes of modulation through these new allosteric sites distal to the nucleotide binding site, as shown in Figure 1.9, have been identified

One of the allosteric modulators, 115-7C, havs behaviors similar to an artificial co-chaperone, which stimulates ATPase activity and promotes protein folding activity of DnaK.[301] Interestingly, it does not compete against DnaJ but acts in concert. NMR chemical shift data, shown in the right panel of Figure 1.9 A, confirms that compound 115-7C binds in a region adjacent to the proposed docking site of DnaJ.[280-282] By applying molecular simulation and computational docking methods, a model was generated that places 115-7C into a transiently appeared pocket formed inside the IIA subdomain. In this model, 115-7C makes contact with several residues that have significant chemical shifts in the NMR experiment. Through this model, 115-7C was modified and the new compound reverses the 115-7C activity on DnaK and becomes an inhibitor.

On the other hand, myricetin and several chemically similar compounds that modulate DnaK activity belong to the family of flavonoids.[241] These compounds inhibit DnaJ-mediated ATPase activity of DnaK and its effect is enhanced by GrpE. As shown in the left panel of Figure 1.9 B.
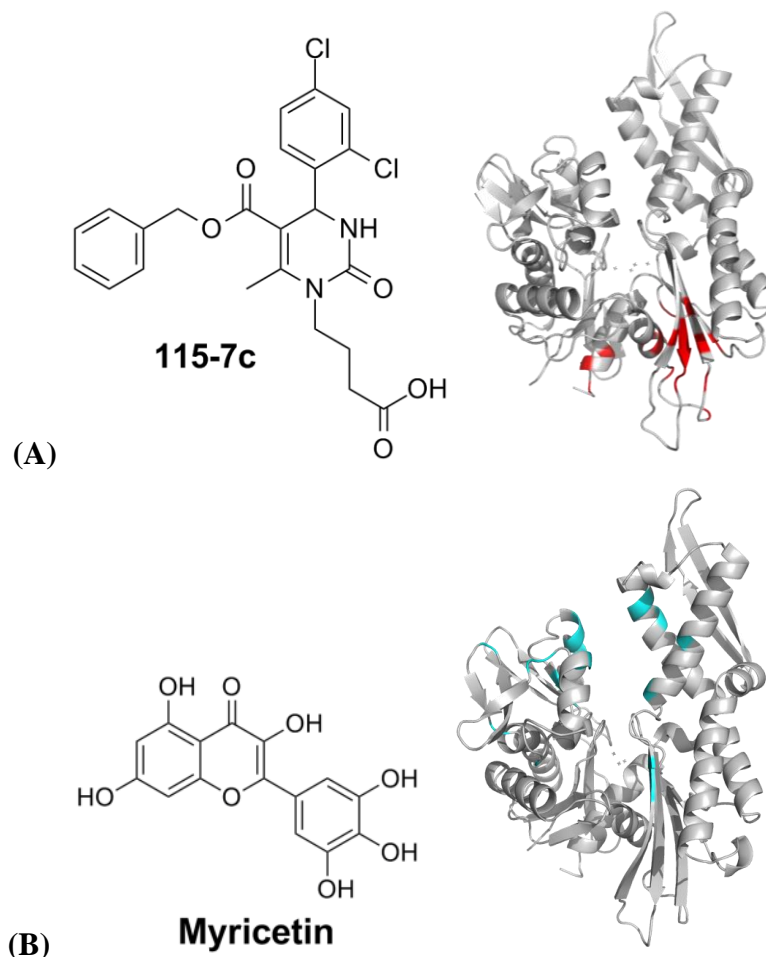
Figure 1.9. DnaK modulators induce NMR chemical shifts at sites distal from the nucleotide-binding site of DnaK. **(A)** Compound 115-7C induces chemical shifts (red) in the IA and IIA subdomains. **(B)** Myricetin induces chemical shifts (cyan) in residues mostly in the IB and IIB subdomains.

NMR result reveals that myricetin induces chemical shifts that cluster mostly on subdomains I-B and II-B, a site situated ~ 10 Å away from the nucleotide-binding site, together with some chemical shifts in the distal II-A. The clustering of the chemical shifts indicates the ligand binds to a pocket formed between subdomains I-B and II-B when NBD is in closed conformation; however, it does not explain the distal chemical shifts on subdomain II-A. It is known that NEFs interact with the subdomain II-B and control the rate of nucleotide exchange through altering the relative position of subdomains I-B, II-A

and II-B.[202] Hence, these distal residues on subdomain II-A may involve in the ATP turnover by acting as structural elements in the conformational change essential to the catalysis process. Furthermore, the change in conformation may propagate along the structure and affect the proline switch (Pro143),[302] a highly conserved residue that is central to a network of hydrogen-bonded residues (K70, F146, R151, and E171) that senses the nucleotide in the binding site and relays the information to SBD, thereby regulating SBD chaperone activity through conformation control.

Studying small molecule modulators may provide insights im how allostery in Hsp70 can be achieved. Furthermore, these modulators offer new opportunities to modulate various disease states through Hsp70. Hence, understanding the mechanisms of allosteric regulation within the Hsp70-cochaperone system is essential. Several computational studies have applied MD to understand the global conformational changes in both the NBD and SBD of Hsp70 in various nucleotide-bound states.[303,304]

1.5 Theory

Molecular Mechanics

Originally developed in the 1950s for use in theoretical physics experiments,[305] computational methods have been further developed and adapted to model complex biomolecular systems such as proteins, nucleic acids, and lipid membranes. To model a chemical system with high accuracy, quantum mechanism (**QM**) that considers explicit electrons in a system would be used; however, this method is only applicable to small

systems because the complexity of the calculation grows exponentially as the system size increases. Since most modeling of bimolecular systems do not involve bond breakage and formation, and interests are mostly focused on the dynamics of the molecules, modeling can be achieved by applying molecular mechanics (**MM**), a simple parametric method based on Newtonian mechanics and does not involve treatment of explicit electrons. With a set of properly optimized parameters, which can be derived from QM calculations and from experimental results such as X-ray crystallographic data and NMR measurements, MM can simulate the behavior of molecular systems with high accuracy.[306,307]

MM describes a set of atoms with parameters and relates atomic coordinates, energies and forces within a simulated system with empirical potential functions called force field (**FF**). Several major FFs currently in use are: AMBER[308] (Assisted Model Building with Energy Refinement), CHARMM[309] (Chemistry of HARvard Macromolecular Mechanics), GROMOS[310] (Groningen Molecular Simulation), and OPLS-AA[311] (Optimized Potentials for Liquid Simulations). Although there are many varieties of FFs, they all share similar modeling ideas: for non-polarizable FFs, each atom is considered as a hard sphere with fixed van der Waals radius and constant atomic partial charge. For polarizable FFs (e.g. AMBER's *ff03*[312]), the distribution of the atomic partial charge over an atom, modeled by additional charge dipoles, is dependent on the chemical environment. Similarly, the energy functions used in most FFs share similar terms that describe the intramolecular (bond stretching, bending, and dihedral angle) and intermolecular (electrostatic/dipole and van der Waals) interactions and tend to be simple pairwise additive functions. Although these approximations are not flawless,[313] they allow fast computations with reasonable accuracy.

In AMBER force fields, the potential energy function used in the dynamic simulations described in the later chapters, has the following form:

$$V_{AMBER}(\vec{R}) = \sum_{bonds} K_d(d - d_0)^2$$

$$+ \sum_{angle} K_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} K_\varphi[1 + \cos(n\varphi - \delta)]$$

$$+ \sum_{non-bond} \left[\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6}\right) + \frac{q_i q_j}{\epsilon_e r_{ij}}\right]$$

(Equation 1.1)

for the bonded interactions, d, θ, φ, and δ are the bond length, bond angle, dihedral angle, and phase angle (either 0° or 180°), respectively; the zero subscript represents the equilibrium values for each of these terms. K_d, K_θ, and K_φ are the bond length, bond angle, and dihedral angle force constants, respectively. The non-bonded interactions are described by the combination of Lennard-Jones potentials and Coulombic function: A_{ij} and B_{ij} represent the van der Waals attraction and London dispersion terms, respectively, ε_e and q_i represent the effective dielectric constant and partial atomic charge of the atom i, respectively, and r_{ij} is the distance between atoms i and j.

Like most other FFs, the potential energy function of AMBER FF (Eq 1.1) describes a linear addition of the bonded and non-bonded interactions (Figure 1.10). The intramolecular bond stretching and bending are described by harmonic functions while the dihedral angle term is modeled as a Fourier series of a periodic function. For non-bonded interactions, a Lennard-Jones potential and a Coulombic function, both pairwise atomic interactions terms, to describe the van der Waals interactions and electrostatic/ dipole interactions, respectively.
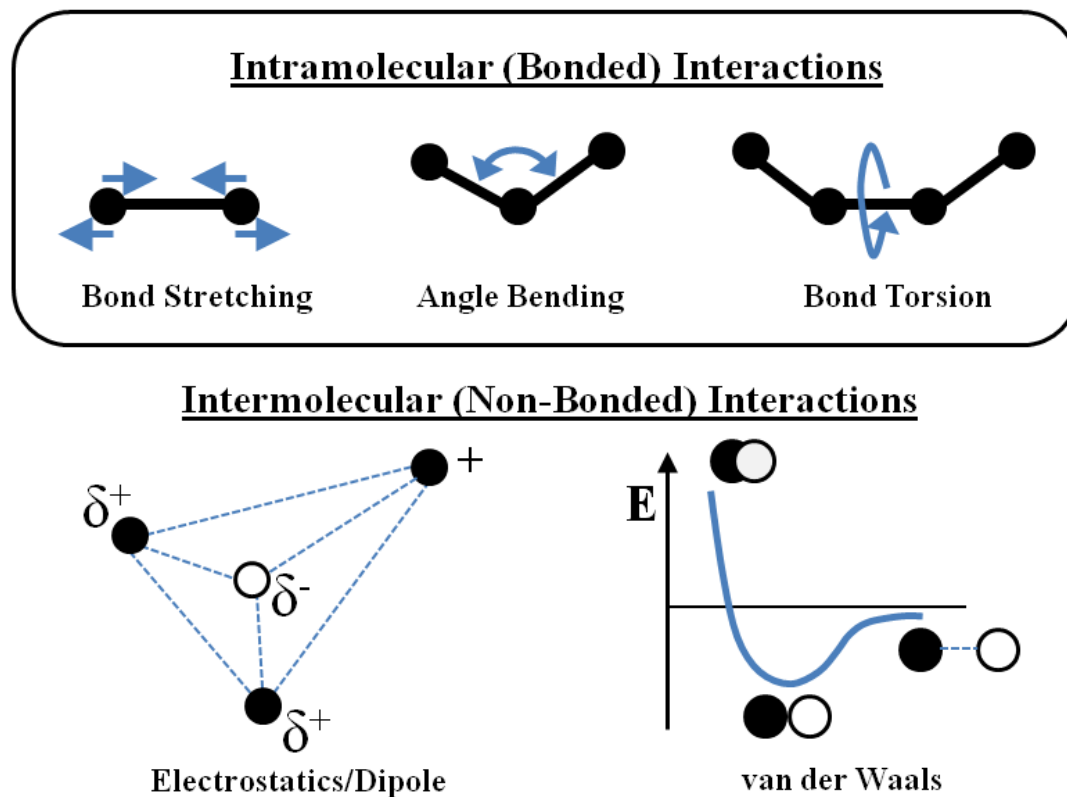
Figure 1.10. Intramolecular and intermolecular components of a Molecular Mechanics used in the AMBER force fields. **(Top)** The three intramolecular interactions: 1-2 bond interaction, 1-3 bond angle, and 1-4 dihedral angle. **(Below)** Intermolecular interactions: charge-charge, charge-dipole, and dipole-dipole interactions; and van der Waals interactions.

Although only a few types of chemical elements are present in biomolecule simulations, these elements usually have striking chemical and physical differences when incorporated into large biomolecules such as nucleic acids, polypeptides, and carbohydrates than into small organic molecules. Hence, in order to accurately describe these different types of molecules in MM, FFs specially parameterized for each group should be used. Some FFs are specially tuned for proteins (ff99SB,[314] CHARMM22,[315] OPLS-AA[311]), nucleic acids (ff99SB, CHARMM27[316]) or carbohydrates (e.g.

GLYCAM06[317]), while other are optimized for use in organic small molecules (e.g. mmff94,[318] AM1-BCC,[319,320] and general AMBER FF[321]).

Molecular Dynamics

Since the first molecular dynamics (**MD**) simulation of a biomolecule was performed in 1976 by McCammon *et al.*,[322,323] MD has been routinely used to study the dynamics of biomolecules. It conducts iterative numerical calculation of instantaneous forces present in a MM system and the consequential movements in that system. MD simulates the trajectory of a collection of atoms based on the atomic coordinates and an initial set of velocity distributions according to the classical Newtonian equations of motion. Newton's second law of motion (Eq. 1.2) describes the relationship among the forces, moments, velocities, and accelerations acting on the atoms in the MM system:

$$F_i(r) = -\frac{dV}{dr_i} = \frac{dp_i}{dt} = \frac{d}{dt}\left(m_i \frac{dr_i}{dt}\right) = m_i \frac{d^2 r_i}{dt^2}$$

(Equation 1.2)

where $F_i(r)$ and p_i are the force and moment exerted on an atom i, V is the potential energy of the MM system, m_i and r_i are the mass and coordinates of the atom i, respectively, at a specific time t.

Equation 1.2 is a crude approximation of the motion of the MM system as this equation is truncated at the second-order term (acceleration) of a Taylor expansion, and assumes the higher-order terms sum to zero. However, such assumption would lead to poor conservation of system energy and induce significant fluctuations in the system energy and a drift of the coordinates over time. Several integration algorithms were developed to improve the integration of the equations of motions: Verlet, velocity Verlet,

leapfrog, and Beeman algorithms.[324] The SHAKE algorithm[325] is usually applied to constrain bonds with hydrogen atoms to allow the use of longer simulation time step without compromising the stability of the integration algorithm.

Condensed phase MD simulation is routinely carried out using static coordinates of a biomolecule, obtained from X-ray crystallography, NMR, or homology model, is immersed in a volume of explicit solvent molecules. In addition to the popular water models including SPC/E, TIP3P, TIP4P, TIP4P-Ew, and TIP5P,[326] organic solvents are also available for use,[311,327,328] as seen in the mix-solvent MD performed by Lexa *et al*.[33] To simulate an infinitely tiled system, a geometric box such as central cubic or truncated octahedron is duplicated periodically in all directions. These periodic boundary conditions ensure all simulated atoms are surrounded by neighboring atoms.

Long-range electrostatic interactions are typically the most computationally expensive calculations in a MD simulation. Particle mesh Ewald method[329-331] was introduced for rigorous evaluation of the electrostatic interactions in an infinite periodic system. It avoids the use of abrupt cutoff distance seen in older MD methods while providing good efficiency with the use of fast Fourier transforms.

Langevin Dynamics

In Langevin dynamics a continuum solvation model replaces the explicit solvent molecules, greatly reduces the number of atoms in the MM system and enhances conformational sampling over standard MD. Langevin equation, Eq 1.3, is used as an alternative to Newton's second law, where two additional terms, a frictional function and a random force function, are incorporated.

$$m_i \frac{d^2r}{dt^2} = F_i(r) - \omega_i \frac{dr}{dt} + R_i(t)$$

(Equation 1.3)

$F_i(r)$ is the force term used in conventional MD, ω_i is the friction coefficient, and $R_i(t)$ is the random force experienced by the atom i. The frictional function is controlled by the collision frequency[332] and introduces a fictional drag experienced by solute molecules in an implicit solvent environment, while the random force term mimics the random impulses associated with the thermal motions of the explicit solvent molecules.

Currently two approaches, Poisson-Boltzmann[333,334] and Generalized Born,[335] are available in AMBER force field to evaluate the electrostatic solvation free energy of the solute in continuum solvation model. Although Poisson-Boltzmann approach is computational expensive, it has been demonstrated to reliably reproduce the energetics and conformations when compared to explicit solvent simulations.[336] Generalized Born approach is an efficient and relatively simple approximation of the Poisson-Boltzmann model that has been widely adopted by MD developers, although it may not be as accurate as Poisson-Boltzmann approach.[337]

Essential Dynamics

Principal component analysis is a method used for reducing the dimensionality of a dataset by performing orthogonal transformation and extracting the strongly correlated variables within the dataset into a set of values of linearly uncorrelated variables. The first principal component, or mode, is thus the linear combination of the variables that produce the largest variances in the data provided. The subsequent principal components

are the linear combinations of the remaining variables that provide the next largest variance.

The major challenge of MD trajectory analysis is the extraction of significant, large-scale, global correlated motions from the local movements and thermal fluctuations occurring in the analyzed simulations. By performing an eigen-decomposition of the covariance matrix of the trajectory, principal component analysis can isolate the collective modes of motions of the trajectory. Since this approach decomposes the trajectory into its most fundamental components, this method is also called the essential dynamics analysis.[338,339] Each mode of motion, or eigenvector, contains the collective motion of the residues of the examined system while the associated eigenvalue quantifies the contribution of the eigenvector to the overall dynamics of the trajectories. Each of these eigenvectors can be individually visualized and inspected.[340] The first few eigenvectors usually describe the global domain movements and changes in protein conformation, whereas the later eigenvectors usually describe the irrelevant local fluctuations, such as side chain rotations (Figure 1.1 Top).

Molecular Docking

Molecular docking is a structure-based computational tool that places a small molecule, generally called the ligand, in the binding site of a macromolecule, called receptor. The strength of the interactions between the ligand and receptor is estimated by the scoring function (**SF**). Since the first application of the molecular docking method in the 1980s by Kuntz *et al.*,[341] molecular docking technique has become a standard tool in SBDD research and is the workhorse in many drug discovery programs.[342]

Compared to quantum mechanics and molecular mechanics, molecular docking may not be as accurate; however, molecular docking offers unparalleled processing speed at an astonishingly low computational cost. These advantages allow molecular docking technique to be used in virtual screening, in which the receptor docks against a large chemical library, and accelerates the drug discovery process and reduces the cost of purchasing and synthesizing compounds that are unlikely to be hits of the receptor. Over the years, many in-house molecular docking programs were created;[343] a few programs developed by academics and industries are popular among researchers and are available in the public domain for use. The latest docking programs have the capability to perform VS: AutoDock,[46] AutoDock Vina,[344] DOCK,[341] FlexX,[345] FRED,[346] Glide,[347,348] GOLD,[349] and ICM.[350]

Docking of a ligand to a binding site of the receptor *in silico* involves two components: searching and scoring. The searching component has two tasks: generates ligand conformations, and performs orientational sampling of the ligand within the receptor binding site. Various ligand conformation sampling strategies are implemented: exhaustive approaches such as systematic sampling (FRED and Glide) and fragment-based incremental construction (FlexX); stochastic techniques such as Monte Carlo (AutoDock Vina, ICM) and genetic algorithm (AutoDock 3 and 4, GOLD); and MM-based methods such as simulated annealing (AutoDock 2.4[351]) and energy minimization methods[14] (DOCK, ICM). Orientational sampling strategies fall into a few categories: shape matching (FRED, DOCK); grid-based potential map (AutoDock, AutoDock Vina, Glide, ICM); and pharmacophore model (FlexX, GOLD).

Another critical component of molecular docking is scoring function. SF is an active area of research because it directly determines the accuracy of the docking results.[352-357] It is used to direct the docking of the ligand in the orientational sampling stage and to evaluate the binding affinity of the final docked pose. Many SFs, standalone (ChemScore,[358] DrugScore,[359] M-Score,[360] and X-Score[361]) or native to the docking programs, have been developed with the goal to accurately predict the binding affinity of the ligand-receptor complex to distinguish true binders from the non-binders, and the distinction between true binding modes from all other alternative modes explored. However, since efficiency is a major requirement of SF in molecular docking, assumptions and simplifications have to be made to reduce the complexity of the algorithm and speed up the calculations. Hence, SF is a compromise of speed and accuracy and is not used as a rigorous method for binding affinity evaluation. Currently, there are three major groups of SFs: empirical functions (FlexX, FRED, Glide, GOLD (ChemScore), X-Score), force-field-based functions (AutoDock, DOCK, GOLD (GoldScore), ICM) and knowledge-based functions (AutoDock Vina, DrugScore, M-Score).

A major limitation of docking is the use of rigid receptor and thus flexibility of the biomolecule is not included. This can lead to failed docking results, such as the well-known inability of most docking programs to place a ligand into the nucleotide binding site of protein kinases in closed conformation, and the failure to identify allosteric sites by docking, such as the Eye allosteric pocket of HIV-1p.[34] Plasticity is an intrinsic character of most biomolecules;[20] hence, incorporation of flexibility has been suggested to improve docking results.[15,19] An adoption of MPS strategy is the relaxed complex

scheme,[47-49,51] where protein conformations generated from MD simulation were sampled for docking. Although this scheme proved to be very powerful and accurate, the need to perform MD makes the method computational expensive. To reduce the computational cost, methods have been developed that allow only the selected residue side-chains to be flexible during docking, as seen in AutoDock 4.2,[46] Induced Fit Docking of Glide,[44] and GOLD. In addition, methods that account for protein backbone movement have also been developed: FlipDock,[45] Induced Fit Docking of Glide, ROSETTALIGAND,[362] SLIDE,[43] and a CONCOORD-based docking scheme developed by Seeliger *et al.*[55]

In addition to improving docking result by optimizing the sampling algorithm, new approaches to improve scoring accuracy are also developed. A straightforward approach is to improve the SF itself; newer scoring functions are being developed and fitted to larger and more diverse dataset for better data coverage.[46,344] Another strategy is to combine multiple scoring results into a consensus score. This approach arises from the assumption that a SF may not perform well in certain receptors but a collection of diverse scoring functions should eliminate such outliers and enhance the success rate of identifying the true binders.[363-365]

A general interest is which docking program and scoring function perform the best, as most people want to use only the best available. Several studies have been done to compare the SFs[352] but no conclusive results have been drawn.[355,357,366-369] This ambiguity may result from the small set of test case for evaluation and the lack of a standardized testing platform for the comparison. As such, a comprehensive, high-quality and diverse dataset by the Community Structure-Activity Resource (CSAR) project has

been developed to provide a leveled testing ground for critical assessment of the SFs.[357,369]

1.6 Specific Aims

The goals of this project are to apply the MPS method, and other computational methods, to study HIV-1p and DnaK-NBD. Specifically,

1. Identify and characterize allosteric HIV-1p inhibitors found using MPS-based dynamic pharmacophore model

2. Develop MixMD method to identify binding hot spots

3. Study the dynamics of DnaK-NBD with MPS strategy

4. Develop CSAR through comparing various scoring functions

1.7 References

1. Lindsay, M. A. Target Discovery. *Nat. Rev. Drug Dis.* **2003**, *2*, 831-838.

2. Ruffner, H.; Bauer, A.; Bouwmeester, T. Human Protein-Protein Interaction Networks and the Value for Drug Discovery. *Drug Dis. Today* **2007**, *12*, 709-716.

3. Sleno, L.; Emili, A. Proteomic Methods for Drug Target Discovery. *Curr. Opin. Chem. Biol.* **2008**, *12*, 46-54.

4. Congreve, M.; Murray, C. W.; Blundell, T. L. Keynote Review: Structural Biology and Drug Discovery. *Drug Dis. Today* **2005**, *10*, 895-907.

5. Scapin, G. Structural Biology and Drug Discovery. *Curr. Pharm. Des.* **2006**, *12*, 2087-2097.

6. Macarron, R. Critical Review of the Role of HTS in Drug Discovery. *Drug Dis. Today* **2006**, *11*, 277-279.

7. Hardy, J. A.; Wells, J. A. Searching for New Allosteric Sites in Enzymes. *Curr. Opin. Struct. Biol.* **2004**, *14*, 1-10.

8. Guido, R. V. C.; Oliva. G. Structure-Based Drug Discovery for Tropical Diseases. *Curr. Top. Med. Chem.* **2009**, *9*, 824-843.

9. Dailey, M. M.; Hait, C.; Holt, P. A.; Maguire, J. M.; Meier, J. B.; Miller, M. C.; Petraccone, L.; Trent, J. O. Structure-Based Drug Design: From Nucleic Acid to Membrane Protein Targets. *Exp. Mol. Pathol.* **2009**, *86*, 141-150.

10. Lu, Y.; Nikolovska-Coleska, Z.; Fang, X.; Gao, W.; Shangary, S.; Qiu, S.; Qin, D.; Wang, S. Discovery of a Nanomolar Inhibitor of the Murine Double Minute 2 (MDM2)-p53 Interaction through an Integrated, Virtual Database Screening Strategy. *J. Med. Chem.* **2006**, *49*, 3759-3762.

11. Gunasekaran, K.; Ma, B.; Nussinov, R. Is Allostery an Intrinsic Property of All Dynamic Proteins? *Proteins* **2004**, *57*, 433-443.

12. Henzler-Wildman, K.; Kern, D. Dynamic Personality of Proteins. *Nature* **2007**, *450*, 964-972.

13. Gutfreund, H.; Knowles, J. R. The Foundations of Enzyme Action. *Essays Biochem.* **1967**, *3*, 25-72.

14. Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein-Ligand Docking: Current Status and Future Challenges. *Proteins* **2006**, *65*, 15-26.

15. Carlson, H. A. Protein Flexibility and Drug Design: How to Hit a Moving Target. *Curr. Opin. Chem. Biol.* **2002**, *6*, 447-452.

16. Freire, E. The Propagation of Binding Interactions to Remote Sites in Proteins: Analysis of the Binding of the Monoclonal Antibody D1.3 to Lysozyme. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 10118-10122.

17. Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. The Energy Landscape and Motions of Proteins. *Science* **1991**, *254*, 1598-1603.

18. Boehr, D. D.; Nussinov, R.; Wright, P. E. The Role of Dynamic Conformational Ensembles in Biomolecular Recognition. *Nat. Chem. Bio.*, **2009**, *5*, 789-796.

19. Carlson, H. A.; McCammon, J. A. Accommodating Protein Flexibility in Computational Drug Design. *Mol. Pharmaco.* **2000**, *57*, 213-218.

20. Henzler-Wildman, K. A.; Thai, V.; Lei, M.; Ott, M.; Wolf-Watz, M.; Fenn, R.; Pozharski, E.; Wilson, M. A.; Petsko, G. A.; Karplus, M.; Hubner, C. G.; Kern, D. Intrinsic Motions along an Enzymatic Reaction Trajectory. *Nature* **2007**, *450*, 838-844.

21. Kroeger Smith, M. B.; Rouzer, C. A.; Taneyhill, L. A.; Smith, N. A.; Hughes, S. H.; Boyer, P. L.; Janssen, P. A. J.; Moereels, H.; Koymans, L.; Arnold, E.; Ding, J.; Das, K.; Zhang, W.; Michejda, C. J.; Smith, Jr. R. H. Molecular Modeling Studies of HIV-1 Reverse Transcriptase Nonnucleoside Inhibitors: Total Energy of Complexation as a Predictor of Drug Placement and Activity. *Protein Sci.* **1995**, *4*, 2203-2222.

22. Campo-Soria, C.; Chang, Y.; Weiss, D. S. Mechanism of Action of Benzodiazepines on $GABA_A$ Receptors. *Br. J. Pharmacol.* **2006**, *148*, 984-990.

23. Berman, H. M. The Protein Data Bank: A Historical Perspective. *Acta. Cryst.* **2008**, *A64*, 88-95.

24. Roberts, G. C. NMR Spectroscopy in Structure-Based Drug Design. *Curr. Opin. Biotechnol.* **1999**, *10*, 42-47.

25. Damm, K. L.; Carlson, H. A. Exploring Experimental Sources of Multiple Protein Conformations in Structure-Based Drug Design. *J. Am. Chem. Soc.* **2007**, *129*, 8225-8235.

26. Carlson, H. A.; Masukawa, K. M.; McCammon, J. A. Method for Including the Dynamic Fluctuations of a Protein in Computer-Aided Drug Design. *J. Phys. Chem. A* **1999,** *103*, 10213-10219.

27. Carlson, H. A.; Masukawa, K. M.; Jorgensen, W. L.; Lins, R. D.; Briggs, J. M.; McCammon, J. A. Developing a Dynamic Pharmacophore Model for HIV-1 Integrase. *J. Med. Chem.* **2000**, *43*, 2100-2114.

28. Meagher, K. L.; Carlson, H. A. Incorporating Protein Flexibility in Structure-Based Drug Discovery: Using HIV-1 Protease as a Test Case. *J. Am. Chem. Soc.* **2004**, *126*, 13276-13281.

29. Meagher, K. L.; Lerner, M. G.; Carlson, H. A. Refining the Multiple Protein Structure Pharmacophore Method: Consistency across Three Independent HIV-1 Protease Models. *J. Med. Chem.* **2006**, *49*, 3478-3484.

30. Bowman, A. L.; Lerner, M. G.; Carlson, H. A. Protein Flexibility and Species Specificity in Structure-Based Drug Discovery: Dihydrofolate Reductase as a Test System. *J. Am. Chem. Soc.* **2007**, *129*, 3634-3640.

31. Bowman, A. L.; Nikolovska-Coleska, Z.; Zhong, H.; Wang, S.; Carlson, H. A. Small Molecule Inhibitors of the MDM2-p53 Interaction Discovered by Ensemble-Based Receptor Models. *J. Am. Chem. Soc.* **2007**, *129*, 12809-12814.

32. Lerner, M. G.; Bowman, A. L.; Carlson, H. A. Incorporating Dynamics in *E. coli* Dihydrofolate Reductase Enhances Structure-Based Drug Discovery. *J. Chem. Inf. Model.* **2007**, *47*, 2358-2365.

33. Lexa, K. W.; Carlson, H. A. Full Protein Flexibility is Essential for Proper Hot-Spot Mapping. *J. Am. Chem. Soc.* **2011**, *133*, 200-202.

34. Damm, K. L.; Ung, P. M.-U.; Quintero, J. J.; Gestwicki, J. E.; Carlson, H. A. A Poke in the Eye: Inhibiting HIV-1 Protease through its Flap-Recognition Pocket. *Biopolymers* **2008**, *89*, 643-652.

35. Mason, J. S.; Good, A. C.; Martin, E. J. 3-D Pharmacophores in Drug Discovery. *Curr. Pharm. Des.* **2001**, *7*, 567-597.

36. Mustata, G. I.; Briggs, J. M. A Structure-Based Design Approach for the Identification of Novel Inhibitors: Application to an Alanine Racemase. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 935-953.

37. Schechner, M.; Sirockin, F.; Stote, R. H.; Dejaegere, A. P. Functionality Maps of the ATP Binding Site of DNA Gyrase B: Generation of a Consensus Model of Ligand Binding. *J. Med. Chem.* **2004**, *47*, 4373-4390.

38. Deng, J.; Sanchez, T.; Neamati, N.; Briggs, J. M. Dynamic Pharmacophore Model Optimization: Identification of Novel HIV-1 Integrase Inhibitors. *J. Med. Chem.* **2006**, *49*, 1684-1692.

39. Teague, S. J. Implications of Protein Flexibility for Drug Discovery. *Nat. Rev. Drug Discov.* **2003**, *2*, 527-541.

40. Jiang, F.; Kim, S. H. "Soft Docking": Matching of Molecular Surface Cubes. *J. Mol. Biol.* **1991**, *219*, 79-102.

41. Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft Docking and Multiple Receptor Conformations in Virtual Screening. *J. Med. Chem.* **2004**, *47*, 5076-5084.

42. Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in Molecular Recognition: The Effects of Ligand and Protein Flexibility on Molecular Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 45-55.

43. Zavodszky, M. I.; Lei, M.; Thorpe, M. F.; Day, A. R.; Kuhn, L. A. Modeling Correlated Main-Chain Motions in Proteins for Flexible Molecular Recognition. *Proteins* **2004**, *57*, 243-261.

44. Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel Procedure for Modeling Ligand/Receptor Induced Fit Effects. *J. Med. Chem.* **2006**, *49*, 534–553.

45. Zhao, Y.; Sanner, M. F. Flipdock: Docking Flexible Ligands into Flexible Receptors. *Proteins* **2007**, *68*, 726-737.

46.   Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olsen, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30*, 2785-2791.

47.   Lin, J.-H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A. Computational Drug Design Accommodating Receptor Flexibility: The Relaxed Complex Scheme. *J. Am. Chem. Soc.* **2002**, *124*, 5632-5633.

48.   Ung, M.-U.; Lu, B.; McCammon, J. A. E230Q Mutation of the Catalytic Subunit of cAMP-dependent Protein Kinase Affects Local Structure and the Binding of Peptide Inhibitor. *Biopolymers* **2006**, *81*, 428-439.

49.   Amaro, R. E.; Baron, R.; McCammon, J. A. An Improved Relaxed Complex Scheme for Receptor Flexibility in Computer-Aided Drug Design. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 693-705.

50.   Seeliger, D.; de Groot, B. L. Conformational Transitions upon Ligand Binding: Holo-Structure Prediction from Apo Conformations. *PLoS Comput. Biol*. **2010**, *6*, e1000634.

51.   Machado, K. S.; Schroeder, E. K.; Ruiz, D. D.; Cohen, E. M. L.; de Souza, O. N. FReDoWS: A Method to Automate Molecular Docking Simulations with Explicit Receptor Flexibility and Snapshot Selection. *BMC Genomic* **2011**, *12*, S6.

52.   Jorgensen, W. L. *BOSS Version 3.8*l Yale University: New Haven, CT, **1997**.

53.   Allen, K. N.; Bellamacina, C. R.; Ding, X.; Jeffery, C. J.; Mattos, C.; Petsko, G. A.; Ringe, D. An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins. *J. Phys. Chem.* **1996**, *100*, 2605-2611.

54.   Mattos, C.; Ringe, D. Locating and Characterizing Binding Sites on Proteins. *Nat. Biotech.* **1996**, *14*, 595-599.

55.   Wang, Z.; Zhu, G.; Huang, Q.; Qian, M.; Shao, M.; Jia, Y.; Tang, Y. X-Ray studies on Cross-Linked Lysozyme Crystals in Acetonitrile-Water Mixture. *Biochim. Biophy. Acta* **1998**, *1384*, 335-344.

56.   English, A. C.; Groom, C. R.; Hubbard, R. E. Experimental and Computational Mapping of the Binding Surface of a Crystalline Protein. *Prot. Eng.* **2001**, *14*, 47-59.

57.   Mattos, C.; Bellamacina, C. R.; Peisach, E.; Pereira, A.; Vitkup, D.; Petsko, G. A.; Ringe, D. Multiple Solvent Crystal Structures: Probing Binding Sites, Plasticity and Hydration. *J. Mol. Biol.* **2006**, *357*, 1471-1482.

58.   Liepinsh, E.; Otting, G. Organic Solvents Identify Specific Ligand Binding Sites on Protein Surfaces. *Nat. Biotech.* **1997**, *15*, 264-268.

59.   Bakan, A.; Nevins, N.; Lakdawala, A. S.; Bahar, I. Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules. *J. Chem. Theory Comp.* **2012**, *8*, 2435-2447.

60. Guvench, O.; MacKerell, A. D. Computational Fragment-Based Binding Site Identification by Ligand Competitive Saturation. *PLoS Comput. Biol.* **2009**, *5*, e1000435.

61. Raman, E. P.; Yu, W.; Guvench, O.; Mackerell, A. D. Reproducing Crystal Binding Modes of Ligand Functional Groups Using Site-Identification by Ligand Competitive Saturation (SILCS) Simulations. *J. Chem. Inf. Model.* **2011**, *51*, 877-896.

62. Raman, E. P.; Vanommeslaeghe, K.; MacKerell, A. D. Site-Specific Fragment Identification Guided by Single-Step Free Energy Perturbation. *J. Chem. Theory Comput.* **2012**, *8*, 3513-3525.

63. Foster, T. J.; Mackerell, A. D.; Guvench, O. Balancing Target Flexibility and Target Denaturation in Computational Fragment-Based Inhibitor Discovery. *J. Comp. Chem.* **2012**, *33*, 1880-1891.

64. Kallings, L. O. The First Postmodern Pandemic: 25 Years of HIV/AIDS. *J. Intern. Med.* **2008**, *263*, 218-243.

65. UNAIDS. 2010 Report on the Global AIDS Epidemic, Joint United Nations Programme on HIV/AIDS. UN Headquarters New York, Geneva, Johannesburg, Mexico City and Port of Spain, **2010**.

66. Sharp, P. M.; Hahn, B. H. Origins of HIV and the AIDS Pandemic. *Cold Spring Harb. Perspect. Med.* **2011**, *1*, a006841.

67. ICTV. International Committee on Taxonomy of Viruses, **2011**. Retrieved 2012-09-04.
http://ictvonline.org/taxonomyHistory.asp?taxnode_id=20114900&taxa_name=Human%20immunodeficiency%20virus%201

68. Gilbert, P. B.; McKeague, I. W.; Eisen, G.; Mullins, C.; Guéye-NDiaye, A.; Mboup, S.; Janki, P. J. Comparison of HIV-1 and HIV-2 Infectivity from a Prospective Cohort Study in Senegal. *Stat. in Med.* **2003**, *22*, 573-593.

69. Santiago, M. L.; Range, F.; Keele, B. F.; Li, Y.; Bailes, E.; Bibollet-Ruche, F.; Fruteau, C.; Noe, R.; Peeters, M.; Brookfield, J. F. Y.; Shaw, G. M.; Sharp, P. M.; Hahn, B. H. Simian Immunodeficiency Virus Infection in Free-Ranging Sooty Mangabeys (*Cercocebus atys atys*) from the Taï Forest, Côte d'Ivoire: Implications for the Origin of Epidemic Human Immunodeficiency Virus Type 2. *J. Virol.* **2005**, *79*, 12515–12527.

70. Rom, W. N.; Markowitz, S. Environmental and Occupational Medicine (4th ed.). Philadelphia: Wolters Kluwer/Lippincott Williams & Wilkins, **2007**, p.745. ISBN: 978-0-7817-6299-1.

71. Copeland, K. F.; Heeney, J. L. T Helper Cell Activation and Human Retroviral Pathogenesis. *Microbiol. Mol. Biol. Rev.* **1996**, *60*, 722-742.

72. Rosenberg, E. S.; Walker, B. D. HIV Type-1 Specific Helper T Cells: A Critical Host Defense. *AIDS Res. Hum. Retroviruses* **1998**, *14*, S143-147.

73. Alimonti, J. B.; Ball, T. B.; Fowke, K. R. Mechanisms of CD4$^+$ T lymphocyte Cell Death in Human Immunodeficiency Virus Infection and AIDS. *J. Gen. Virol.* **2003**, *84*, 1649-1661.

74. Clarke, S. J.; Saag, M. S.; Decker, W. D.; Campbell-Hill, S.; Roberson, J. L.; Veldkamp, P. J.; Kappes, J. C.; Hahn, B. H.; Shaw, G. M. High Titers of Cytopathic Virus in Plasma of Patients with Symptomatic Primary HIV-1 Infection. *New Engl. J. Med.* **1991**, *324*, 954-960.

75. Daar, E. S.; Moudgil, T.; Meyer, R. D.; Ho, D. D. Transient High Levels of Viremia in Patients with Primary Human Immunodeficiency Virus Type 1 Infection. *New Engl. J. Med.* **1991**, *324*, 961-964.

76. Levy, J. A. HIV Pathogenesis: Knowledge Gained after Two Decades of Research. *Adv. Dent. Res.* **2006**, *19*, 10-16.

77. Chu, C.; Selwyn, P. A. Complications of HIV Infection: A Systems-Based Approach. *Am. Fam. Physician* **2011**, *83*, 395-406.

78. Coffin, J. M. HIV Population Dynamics In Vivo: Implication for Genetic Variation, Pathogenesis, and Therapy. *Science* **1995**, *267*, 483-489.

79. Mansky, L. M. Retrovirus Mutation Rates and Their Role in Genetic Variation. *J. Gen. Virol.* **1998**, *79*, 1337-1345.

80. Ji, J.; Loeb, L. A. Fidelity of HIV-1 Reverse Transcriptase Copying RNA in Vitro. *Biochemistry* **1992**, *31*, 954-958.

81. Kati, W. M.; Johnson, K.A.; Jerva, L. F.; Anderson, K. S. Mechanism and Fidelity of HIV Reverse Transcriptase. *J. Biol. Chem.* **1992**, *267*, 25988-25997.

82. Preston, B. D.; Dougherty, J. P. Mechanisms of Retroviral Mutation. *Trends Microbiol.* **1996**, *4*, 16–21.

83. Asante-Appiah, E.; Skalka, A. M. HIV-1 Integrase: Structural Organization, Conformational Changes, and Catalysis. *Adv. Virus Res.* **1999**, *52*, 351-369.

84. Kuwata, T.; Miyazaki, Y.; Igarashi, T.;Takehisa, J.; Hayami, M. The Rapid Spread of Recombinants During a Natural In Vitro Infection with Two Human Immunodeficiency Virus Type 1 Strains. *J. Virol.* **1997**, *71*, 7088–7091.

85. Onafuwa-Nuga, A.; Telesnitsky, A. The Remarkable Frequency of Human Immunodeficiency Virus Type 1 Genetic Recombination. *Microbiol. Mol. Biol. Rev.* **2009**, *73*, 451-480.

86. Coffin, J. Retroviridae: The Viruses and Their Replication. In B. N. Fields, D. M. Knipe, P. M. Howley, *et al*. (ed.), *Fundamental Virology, 3rd ed.* **1996**, Lippincott, Philadelphia, PA.

87. An, W.; Telesnitsky, A. HIV-1 Genetic Recombination: Experimental Approaches and Observations. *AIDS Rev*. **2002**, *4*, 195–212.

88. Moutouh, L.; Corbeil, J.; Richman, D. D. Recombination Leads to the Rapid Emergence of HIV-1 Dually Resistant Mutants under Selective Drug Pressure. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 6106–6111.

89. Garcia-Lerma, J. G.; Heneine, W. Resistance of Human Immunodeficiency Virus Type 1 to Reverse Transcriptase and Protease Inhibitors: Genotypic and Phenotypic Testing. *J. Clin. Virol.* **2001**, *21*, 197–212.

90. Chiu, Y. L.; Greene, W. C. Multifaceted Antiviral Actions of APOBEC3 Cytidine Deaminases. *Trends Immunol.* **2006**, *27*, 291–297.

91. Lecossier, D.; Bouchonnet, F.; Clavel, F.; Hance, A. J. Hypermutation of HIV-1 DNA in the Absence of the Vif Protein. *Science* **2003**, *300*, 1112.

92. Zhang, H.; Yang, B.; Pomerantz, R. J.; Zhang, C.; Arunachalam, S. C.; Gao, L. The Cytidine Deaminase CEM15 Induces Hypermutation in Newly Synthesized HIV-1 DNA. *Nature* **2003**, *424*, 94-98.

93. Schröfelbauer, B.; Yu, Q.; Zeitlin, S. G.; Landau, N. R. Human Immunodeficiency Virus Type 1 Vpr Induces the degradation of the UNG and SMUG Uracil-Dna Glycosylases. *J. Virol.* **2005**, *79*, 10978-10987.

94. Monajemi, M.; Woodworth, C. F.; Benkaroun, J.; Grant, M.; Larijani, M. Emerging Complexities of APOBEC3G Action on Immunity and Viral Fitness during HIV Infection and Treatment. *Retrovirology* **2012**, *9*, 35.

95. Louis, J. M.; Clore, G. M.; Gronenborn, A. M. Autoprocessing of HIV-1 Protease Is Tightly Coupled to Protein Folding. *Nature Struct. Biol.* **1999**, *6*, 868-875.

96. Huang, L.; Li, Y.; Chen, C. Flexible Catalytic Site Conformations Implicated in Modulation of HIV-1 Protease Autoprocessing Reactions. *Retrovirology* **2011**, *8*, 79.

97. Kohl, N. E.; Emini, E. A.; Schleif, W. A.; Davis, L. J.; Heimbach, J. C.; Dixon, R. A.; Scolnick, E. M.; Sigal, I. S. Active Human Immunodeficiency Virus Protease is Required for Viral Infectivity. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 4686-4690.

98. Kräusslich, H.-G. Human Immunodeficiency Virus Proteinase Dimer as Component of the Viral Polyprotein Prevents Particle Assembly and Viral Infectivity. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 3213-3217.

99. Flexner, C. HIV Drug Development: the Next 25 Years. *Nat. Rev. Drug Dis.* **2007**, *6*, 959-966.

100. Navia, M. A.; Fitzgerald, P. M.; McKeever, B. M.; Leu, C. T.; Heimbach, J. C.; Herber, W. K.; Sigal, I. S.; Darke, P. L.; Springer, J. P. Three-Dimensional Structure of Aspartyl Protease from Human Immunodeficiency Virus HIV-1. *Nature* **1989**, *337*, 615-620.

101. Wlodawer, A.; Miller, M.; Jaskólski, M.; Sathyanarayana, B. K.; Baldwin, E.; Weber, I. T.; Selk, L. M.; Clawson, L.; Schneider, J.; Kent, S. B. H. Conserved Folding in Retroviral Proteases: Crystal Structure of a Synthetic HIV-1 Protease. *Science* **1989**, *245*, 616-621.

102. Strisovsky, K.; Tessmer, U.; Langner, J.; Konvalinka, Kräusslich, H.-G. Systematic Mutational Analysis of the Active-Site Threonine of HIV-1 Protease: Rethinking the "Fireman-s Grip" Hypothesis. *Prot. Sci.* **2000**, *9*, 1631-1641.

103. Ingr, M.; Uhliková, T.; Strisovsky, K.; Majerová, E.; Konvalinka, J. Kinetics of the Dimerization of Retroviral Protease: The "Fireman's Grip" and Dimerization. *Prot. Sci.* **2003**, *12*, 2173-2182.

104. Darke, P. L.; Jordan, S. P.; Hall, D. L.; Zugay, J. A.; Shafer, J. A.; Kuo, L. C. Dissociation and Association of the HIV-1 Protease Dimer Subunits: Equilibria and Rates. *Biochemistry* **1994**, *33*, 98-105.

105. Jordan, S. P.; Zugay, J.; Darke, P. L.; Kuo, L. C. Activity and Dimerization of Human Immunodeficiency Virus Protease as a Function of Solvent Composition and Enzyme Concentration. *J. Biol. Chem.* **1992**, *267*, 20028-20032.

106. Rick, S. W.; Erickson, J. W.; Burt, S. K. Reaction Path and Free Energy Calculations of the Transition between Alternate Conformations of HIV-1 Protease. *Proteins* **1998**, *32*, 7-16.

107. Chang, C.-N. A.; Trylska, J.; Tozzini, V.; McCammon, J. A. Binding Pathways of Ligands to HIV-1 Protease: Coarse-Grained and Atomistic Simulations. *Chem. Biol. Drug Des.* **2007**, *69*, 5-13.

108. Prabu-Jeyabalan, M.; Nalivaika, E.; Schiffer, C. A. Substrate Shape Determines Specificity of Recognition for HIV-1 Protease Analysis of Crystal Structures of Six Substrate Complexes. *Structure* **2002**, *10*, 369-381.

109. Ding, F.; Layten, M.; Simmerling, C. Solution Structure of HIV-1 Protease Flaps Probed by Comparison of Molecular Dynamics Simulation Ensembles and EPR Experiments. *J. Am. Chem. Soc.* **2008**, *130*, 7184-7185.

110. Galiano, L.; Ding, F.; Veloro, A. M.; Blackburn, M. E.; Simmerling, C.; Fanucci, G. E. Drug Pressure Selected Mutations in HIV-1 Protease Alter Flap Conformations. *J. Am. Chem. Soc.* **2009**, *131*, 430-431.

111. Ishima, R.; Freeberg, D. I.; Wang, Y. X.; Louis, J. M.; Torchia, D. A. Flap Opening and Dimer-Interface Flexibility in the Free and Inhibitor-Bound HIV Protease and Their Implications for Function. *Structure* **1999**, *7*, 1047-1055.

112. Katoh, E.; Louis, J. M.; Yamazaki, T.; Gronenborn, A. M.; Torchia, D. A.; Ishima, R. A Solution NMR Study of the Binding Kinetics and the Internal Dynamics of an HIV-1 Protease-Substrate Complex. *Protein Sci.* **2003**, *12*, 1376-1385.

113. Ishima, R.; Louis, J. M. A Diverse View of Protein Dynamics from NMR Studies of HIV-1 Protease Flaps. *Proteins* **2008**, *70*, 1408-1415.

114. Perryman, A. L.; Lin, J.-H.; McCammon, J. A. Restrained Molecular Dynamics Simulations of HIV-1 Protease: the First Step in Validating a New Target for Drug Design. *Biopolymers* **2006**, *82*, 272-284.

115. Hornak, V.; Okur, A.; Rizzo, R. C.; Simmerling, C. HIV-1 Protease Flaps Spontaneously Open and Reclose in Molecular Dynamics Simulations. *Proc. Nat. Am. Sci. U.S.A.* **2006**, *103*, 915-920.

116. Sadiq, S. K.; De Fabritiis, G. Explicit Solvent Dynamics and Energetics of HIV-1 Protease Flap Opening and Closing. *Proteins* **2010**, *78*, 2873-2885.

117. Karthik, S.; Senapati, S. Dynamic Flaps in HIV-1 Protease Adopt Unique Ordering at Different Stages in the Catalytic Cycle. *Proteins* **2011**, *79*, 1830-1840.

118. Tozzini, V.; McCammon, J. A. A Coarse Grained Model of the Dynamics of Flap Opening in HIV-1 Protease. *Chem. Phys. Lett.* **2005**, *413*, 123-128.

119. Hornak, V.; Simmerling, C. Targeting Structural Flexibility in HIV-1 Protease Inhibitor Binding. *Drug Discov. Today* **2007**, *12*, 132-138.

120. Deng, N.-J.; Zheng, W.; Gallicchio, E.; Levy, R. M. Insights into the Dynamics of HIV-1 Protease: A Kinetic Network Model Constructed from Atomistic Simulations. *J. Am. Chem. Soc.* **2011**, *133*, 9387-9394.

121. Foulkes-Murzycki, J. E.; Scott, W. R.; Schiffer, C. A. Hydrophobic Sliding: A Possible Mechanism for Drug Resistance in Human Immunodeficiency Virus Type 1 Protease. *Structure* **2007**, *15*, 225-233.

122. Tozzini, V.; Trylska, J.; Chang, C,-e; McCammon, J. A. Flap Opening Dynamics in HIV-1 Protease Explored with a Coarse-Grained Model. *J. Struct. Biol.* **2007**, *157*, 606-615.

123. Hamelberg, D.; McCammon, J. A. Fast Peptidyl Cis-Trans Isomerization within the Flexible Gly-Rich Flaps of HIV-1 Protease. *J. Am. Chem. Soc.* **2005**, *127*, 13778-13779.

124. Robbins, A. H.; Coman, R. M.; Bracho-Sanchez, E.; Fernandez, M. A.; Gilliland, C. T.; Li, M.; Agbandje-McKenna, M.; Wlodawer, A.; Dunn, B. M.; McKenna, R. Structure of the Unbound Form of HIV-1 Subtype A Protease: Comparison with Unbound Forms of Proteases from Other HIV Subtypes. *Acta. Crystallogr. D: Biol. Crystallogr.* **2010**, *66*, 233-242.

125. Böttcher, J.; Blum, A.; Dörr, S.; Heine, A.; Diederich, W. E.; Klebe, G. Targeting the Open-Flap Conformation of HIV-1 Protease with Pyrrolidine-Based Inhibitors. *ChemMedChem* **2008**, *3*, 1337-1344.

126. Nicholson, L. K.; Yamazaki, T.; Torchia, D. A.; Grzesiek, S.; Bax, A.; Stahl, S. J.; Kaufman, J. D.; Wingfield, P. T.; Lam, P. Y. S.; Jadhav, P. K.; Hodge, C. N.; Domaille, P. J.; Chang, C. H. Flexibility and Function in HIV-1 Protease. *Nat. Struct. Biol.* **1995**, *2*, 274-280.

127. Freeberg, D. I.; Ishima, R.; Jacob, J.; Wang, Y. X.; Kustanovich, I.; Louis, J. M.; Torchia, D. A. Rapid Structural Fluctuations of the Free HIV Protease Flaps in Solution: Relationship to Crystal Structures and Comparison with Predictions of Dynamics Calculations. *Prot. Sci.* **2002**, *11*, 221-232.

128. Spinelli, S.; Liu, Q. Z.; Alzari, P. M.; Hirel, P. H.; Poljak, R. J. The Three-Dimensional Structure of the Aspartyl Protease from the HIV-1 Isolate BRU. *Biochimie* **1991**, *73*, 1391-1396.

129. Heaslet, H.; Rosenfeld, R.; Giffin, M.; Lin, Y.-C.; Tam, K.; Torbett, B. E.; Elder, J. H.; McRee, D. E.; Stout, C. D. Conformational Flexibility in the Flap Domains of Ligand-Free HIV Protease. *Acta Cryst.* **2007**, *D63*, 866-875.

130. Martin, P.; Vickrey, J. F.; Proteasa, G.; Jimenez, Y. L.; Wawrzak, Z.; Winters, M. A.; Merigan, T. C.; Kovari, L. C. "Wide-Open" 1.3 Å Structure of a Multidrug-Resistant HIV-1 Protease as a Drug Target. *Structure* **2005**, *13*, 1887-1895.

131. Layten, M.; Nornak, V.; Simmerling, C. The Open Structure of a Multi Drug Resistant HIV-1 Protease is Stabilized by Crystal Packing Contacts. *J. Am. Chem. Soc*. **2005**, *128*, 13360-13361.

132. Lexa, K. W.; Damm, K. L.; Quintero, J. J.; Gestwicki, J. E.; Carlson, H. A. Clarifying Allosteric Control of the Flap Conformations in the 1TW7 Crystal Structure of HIV-1 Protease. *Proteins* **2009**, *74*, 872-880.

133. Sham, H. L.; Zhao, C.; Stewart, K. D.; Betebenner, D. A.; Lin, S.; Park, C. H.; Kong, X.-P.; Rosenbrook, W.; Herrin, T.; Madigan, D.; Vasavanonda, S.; Lyons, N.; Molla, A.; Saldivar, A.; Marsh, K. C.; McDonald, E.; Wideburg, N. E.; Denissen, J. F.; Robins, T.; Kempf, D. J.; Plattner, J. J.; Norbeck, D. W. A Novel, Picomolar Inhibitor of Human Immunodeficiency Virus Type 1 Protease. *J. Med. Chem.* **1996**, *39*, 392-397.

134. Hornak, V.; Okur, A.; Rizzo, R. C.; Simmerling, C. HIV-1 Protease Flaps Spontaneously Close to the Correct Structure in Simulations Following Manual Placement of an Inhibitor into the Open State. *J. Am. Chem. Soc.* **2006**, *128*, 2812-2813.

135. Toth, G.; Borics, A. Flap Opening Mechanism of HIV-1 Protease. *J. Mol. Graph. Model.* **2006**, *24*, 465-474.

136. Toth, G.; Borics, A. Closing of the Flaps of HIV-1 Protease Induced by Substrate Binding: A Model of a Flap Closing Mechanism in Retroviral Aspartic Proteases. *Biochemistry* **2006**, *45*, 6606-6614.

137. Baca, M.; Kent, S. B. Catalytic Contribution of Flap-Substrate Hydrogen Bonds in "HIV-1 Protease" Explored by Chemical Synthesis. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 11638-11642.

138. Piana, S.; Carloni, P.; Parrinello, M. Role of Conformational Fluctuations in the Enzymatic Reaction of HIV-1 Protease. *J. Mol. Biol.* **2002**, *319*, 567-583.

139. Ledergerber, B.; Telenti, A.; Egger, M. Risk of HIV Related Kaposi's Sarcoma and Non-Hodgkin's Lymphoma with Potent Antiretroviral Therapy: Prospective Cohort Study. *Biol. Med. J.* **1999**, *319*, 23-24.

140. Eder, J.; Hommel, U.; Cumin, F.; Martoglio, B.; Gerhartz, B. Aspartic Proteases in Drug Discovery. *Curr. Pharm. Des.* **2007**, *13*, 271-285.

141. Flexner, C.; Bate, G.; Kirkpatrick, P. Tipranavir. *Nat. Rev. Drug Discov.* **2005**, *4*, 955-956.

142. Molla, A.; Granneman, G. R.; Sun, E.; Kempf, D. J. Recent Developments in HIV Protease Inhibitor Therapy. *Antiviral Res*. **1998**, *39*, 1–23.

143. Boden, D.; Hurley, A.; Zhang, L.; Cao, Y.; Guo, Y.; Jones, E.; Tsay, J.; Ip, J.; Farthing, C.; Limoli, K.; Parkin, N.; Markowitz, M. HIV-1 Drug Resistance in Newly Infected Individuals. *J. Am. Med. Assoc.* **1999**, *282*, 1135-1141.

144. Kozisek, M.; Saskova, K. G.; Rezacova, P.; Brynda, J.; van Maarseveen, N. M.; de Jong, D.; Boucher, C. A.; Kagan, R. M.; Nijhuis, M.; Konvalinka, J. Ninety-Nine Is Not Enough: Molecular Characterization of Inhibitor-Resistant Human Immunodeficiency Virus Type 1 Protease Mutants with Insertions in the Flap Region. *J. Virol.* **2008**, *82*, 5869-5878.

145. Pereira-Vaz, J.; Duque, V.; Trindade, L.; Saraiva-da-Cunha, J.; Melico-Silvestre, A. Detection of the Protease Codon 35 Amino Acid Insertion in Sequences from Treatment-Naive HIV-1 Subtype C Infected Individuals in the Central Region of Portugal. *J. Clin. Virol.* **2009**, *46*, 169-172.

146. Weber, I. T.; Agniswamy, J. HIV-1 Protease: Structural Perspectives on Drug Resistance. *Viruses* **2009**, *1*, 1110-1136.

147. Johnson, V. A; Brun-Vézinet, F.; Clotet, B.; Günthard, H. F.; Juritzkes, D. R.; Pillay, D.; Schapiro, J. M.; Richman, D. D. Update of the Drug Resistance Mutations in HIV-1: December 2010. *Top HIV Med.* **2010**, *18*, 156-163.

148. Beck, Z. Q.; Hervio, L.; Dawson, P. E.; Elder, J. H.; Madison, E. L. Protease Using a Phage Display Library and Use in Inhibitor Development. *Virology* **2000**, *274*, 391-401.

149. Dam, E.; Quercia, R.; Glass, B.; Descamps, D.; Launay, O.; Duval, X.; Kräusslich, H.-G.; Jance, A. J.; Clavel, F. Gag Mutations Strongly Contribute to HIV-1 Resistance to Protease Inhibitors in Highly Drug-Experienced Patients besides Compensating for Fitness Loss. *PLoS Pathog.* **2009**, *5*, e1000345.

150. Özen, A.; Haliloglu, T.; Schiffer, C. A. HIV-1 Protease and Substrate Coevolution Validates the Substrate Envelope as the Substrate Recognition Pattern. *J. Chem. Theory Comput.* **2012**, *8*, 703-714.

151. Velazquez-Campoy, A.; Muzammil, S.; Ohtaka, H.; Schon, A.; Vega, S.; Freire, E. Structural and Thermodynamic Basis of Resistance to HIV-1 Protease Inhibition: Implications for Inhibitor Design. *Curr. Drug Targets. Infect. Disord.* **2003**, *3*, 311-328.

152. Tie, Y.; Boross, P. I.; Wang, Y. F.; Gaddis, L.; Hussain, A. K.; Leshchenko, S.; Ghosh, A. K.; Louis, J. M.; Harrison, R. W.; Weber, I. T. High Resolution Crystal Structures of HIV-1 Protease with a Potent Non-Peptide Inhibitor (UIC-94017) Active Against Multi-Drug-Resistant Clinical Strains. *J. Mol .Biol.* **2004**, *338*, 341-352.

153. Muzammil, S.; Ross, P.; Freire, E. A Major Role for a Set of Non-Active Site Mutations in the Development of HIV-1 Protease Drug Resistance. *Biochemistry* **2003**, *42*, 631-638.

154. Clemente, J. C.; Moose, R. E.; Hemrajani, R.; Whitford, L. R.; Govindasamy, L.; Reutzel, R.; McKenna, R.; Agbandje-McKenna, M.; Goodenow, M. M.; Dunn, B. M. Comparing the Accumulation of Active- and Nonactive-Site Mutations in the HIV-1 Protease. *Biochemistry* **2004**, *43*, 12141-12151.

155. Ode, H.; Neya, S.; Hata, M.; Sugiura, W.; Hoshino, T. Computational Simulations of HIV-1 Protease − Multi-Drug Resistance Due to Nonactive Site Mutation L90M. *J. Am. Chem. Soc.* **2006**, *128*, 7887-7895.

156. Ode, H.; Matsuyama, S.; Hata, M.; Neya, S.; Kakizawa, J.; Sugiura, W.; Hoshino, T. Computational Characterization of Structural Role of the Non-Active Site Mutation M36I of Human Immunodeficiency Virus Type 1 Protease. *J. Mol. Biol.* **2007**, *370*, 598-607.

157. Seibold, S. A.; Cukier, R. I. A Molecular Dynamics Study Comparing a Wild-Type with a Multiple Drug Resistant HIV Protease: Differences in Flap and Aspartate 25 Cavity. *Proteins* **2007**, *69*, 551-565.

158. Liu, F.; Kovalevsky, A. Y.; Tie, Y.; Ghosh, A. K.; Harrison, R. W.; Weber, I. T. Effect of Flap Mutations on Structure of HIV-1 Protease and Inhibition by Saquinavir and Darunavir. *J. Mol. Biol.* **2008**, *381*, 102-115.

159. Cameron, C. E.; Grinde, B.; Jacques, P.; Jentoft, J.; Leis, J.; Wlodawer, A.; Weber, I. T. Comparison of the Substrate-Binding Pockets of the Rous Sarcoma Virus and Human Immunodeficiency Virus Type 1 Proteases. *J. Biol. Chem.* **1993**, *268*, 11711-11720.

160. Darke, P. L.; Nutt, R. F.; Brady, S. F.; Garsky, V. M.; Ciccarone, T. M.; Leu, C.-T.; Lumma, P. K.; Freidinger, R. M.; Veber, D. F.; Sigal, I. S. HIV-1 Protease Specificity of Peptide Cleavage Is Sufficient for Processing of Gag and Pol Polyproteins. *Biochem. Biophys. Res. Commun.* **1988**, *156*, 297-303.

161. Ridky, T. W.; Cameron, C. E.; Cameron, J.; Leis, J.; Copeland, T.; Wlodawer, A.; Weber, I. T.; Harrison, R. W. Protease Substrate Specificity Is Limited by Interactions between Substrate Amino Acids Bound in Adjacent Enzyme Subsites. *J. Biol. Chem.* **1996**, *271*, 4709-4717.

162. King, N. M.; Prabu-Jeyabalan, M.; Nalivaika, E. A.; Schiffer, C. A. Combating Susceptibility to Drug Resistance: Lessons from HIV-1 Protease. *Chem. Biol.* **2004**, *11*, 1333-1338.

163. Prabu-Jeyabalan, M.; King, N. M.; Nalivaika, E. A.; Heilek-Snyder, G.; Cammack, N.; Schiffer, C. A. Substrate Envelope and Drug Resistance: Crystal Structure of Ro1 in Complex with Wild-Type Human Immunodeficiency Virus Type 1 Protease. *Antimicrob. Agents Chemother.* **2006**, *50*, 1518-1521.

164. King, N. M.; Prabu-Jeyabalan, M.; Nalivaika, E. A.; Wigerinck, P.; de Bethune, M. P.; Schiffer, C. A. Structural and Thermodynamic Basis for the Binding of TMC114, a Next-Generation Human Immunodeficiency Virus Type 1 Protease Inhibitor. *J. Virol.* **2004**, *78*, 12012–12021.

165. Surleraux, D. L.; Tahri, A.; Verschueren, W. G.; Pille, G. M.; de Kock, H. A.; Jonckers, T. H.; Peeters, A.; De Meyer, S.; Azijn, H.; Pauwels, R.; de Bethune, M. P.; King, N. M.; Prabu-Jeyabalan, M.; Schiffer, C. A.; Wigerinck, P. B. Discovery and Selection of TMC114, a Next Generation HIV-1 Protease Inhibitor. *J. Med. Chem.* **2005**, *48*, 1813-1822.

166. Chellappan, S.; Kairys, V.; Fernandes, M. X.; Schiffer, C.; Gilson, M. K. Evaluation of the Substrate Envelope Hypothesis for Inhibitors of HIV-1 Protease. *Proteins*, **2007**, *68*, 561-567.

167. Chellappan, S.; Reddy, G. S. K. K.; Ali, A.; Nalam, M. N. L.; Anjum, S. G.; Cao, H.; Kairys, V.; Fernandes, M. X.; Altman, M. D.; Tidor, B.; Rana, T. M.; Schiffer, C. A.; Gilson, M. K. Design of Mutation-Resistant HIV Protease Inhibitors with the Substrate Envelope Hypothesis. *Chem. Biol. Drug Des.* **2007**, *69*, 298-313.

168. Parai, M. K.; Huggins, D. J.; Cao, H.; Nalam, M. N. L.; Ali, A.; Schiffer, C. A.; Tidor, B.; Rana, T. M. Design, Synthesis, and Biological and Structural Evaluations of Novel HIV-1 Protease Inhibitors to Combat Drug Resistance. *J. Med. Chem.* **2012**, *55*, 6328-6341.

169. Özen, A.; Haliloglu, T.; Schiffer, C. A. Dynamics of Preferential Substrate Recognition in HIV-1 Protease: Redefining the Substrate Envelope. *J. Mol. Biol.* **2011**, *410*, 726-744.

170. Kozakov, D.; Hall, D. R.; Chuang, G.-Y.; Cencic, R.; Brenke, R.; Gove, L. E.; Beglov, D.; Pelletier, J.; Whitty, A.; Vajda, S. Structural Convervation of Druggable Hot Spots in Protein-Protein Interfaces. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 13528-13533.

171. Todd, M. J.; Semo, N.; Freire, E. The structural Stability of the HIV-1 Protease. *J. Mol. Biol.* **1998**, *283*, 475-488.

172. Gustchina, A.; Weber, I. T. Comparative Analysis of the Sequences and Structures of HIV-1 and HIV-2 Proteases. *Proteins* **1991**, *10*, 325-339.

173. Wang, W.; Kollman, P. A. Free Energy Calculations on Dimer Stability of the HIV Protease Using Molecular Dynamics and a Continuum Solvent Model. *J. Mol. Biol.* **2000**, *303*, 567-582.

174. Miller, V. International Perspectives on Antiretroviral Resistance. Resistance to Protease Inhibitors. *J. Acquir. Immune Defic. Syndr.* **2001**, *26*, S34-S50.

175. D'Aquila, R. T.; Schapiro, J. M.; Brun-Vezinet, F.; Clotet, B.; Conway, B.; Demeter, L. M.; Grant, R. M.; Johnson, V. A.; Kuritzkes, D. R.; Loveday, C.; Shafer, R. W.; Richman, D. D. Drug Resistance Mutations in HIV-1. *Top HIV Med.* **2002**, *10*, 21-25.

176. Frutos, S.; Rodriguez-Mias, R. A.; Madurga, S.; Collinet, B.; Reboud-Ravaux, M.; Ludevid, D.; Giralt, E. Disruption of the HIV-1 Protease Dimer and Interfacial Peptides: Structural Studies Using NMR Spectroscopy, Combined with [2-^{13}C]-Trp Selective Labeling. *Peptid. Sci.* **2007**, *88*, 164-173.

177. Zhang, Z. Y.; Poorman, R. A.; Maggiora, L. L.; Heinrikson, R. L.; Kezdy, F. J. Dissociative Inhibition of Dimeric Enzymes: Kinetic Characterization of the Inhibition of HIV-1 Protease by Its COOH-Terminal Tetrapeptide. *J. Biol. Chem.* **1991**, *266*, 15591-15594.

178. Schramm, H. J.; Boetzel, J.; Buttner, J.; Fritsche, E.; Gohring, W.; Jaeger, E.; Konig, S.; Thumfart, O.; Wenger, T.; Nagel, N. E.; Schramm, W. The Inhibition of

Human Immunodeficiency Virus Protease by 'Interface Peptides'. *Antiviral Res.* **1996**, *30*, 155-170.

179. Breccia, P.; Boggetto, N.; Perez-Fernandez, R.; Van Gool, M; Takahashi, M.; Rene, L.; Prados, P.; Badet, B.; Reboud-Ravaux, M.; de Mendoza, J. Dimerization Inhibitors of HIV-1 Protease Based on a Bicyclic Guanidinium Subunit. *J. Med. Chem.* **2003**, *46*, 5196-5207.

180. Zutshi, R.; Franciskovich, J.; Shultz, M.; Schweitzer, B.; Bishop, P.; Wilson, M.; Chmielewski, J. Targeting the Dimerization Interface of HIV-1 Protease: Inhibition with Cross-Linked Interfacial Peptides. *J. Am. Chem. Soc.* **1997**, *119*, 4841-4845.

181. Shultz, M. D.; Bowman, M. J.; Ham, Y.-W.; Zhao, X.; Tora, G.; Chmielewski, J. Small-Molecule Inhibitors of HIV-1 Protease Dimerization Derived from Cross-Linked Interfacial Peptides. *Angew, Chem.* **2000**, *112*, 2822-2825.

182. Bowman, M. J.; Byrne, S.; Chmielewski, J. Switching between Allosteric and Dimerization Inhibition of HIV-1 Protease. *Chem. Biol.* **2005**, *12*, 439-444.

183. Bannwarth, L.; Kessler, A.; Pethe, S.; Collinet, B.; Merabet, N.; Boggetto, N.; Sicsic, S.; Reboud-Ravaux, M.; Ongeri, S. Molecular Tongs, Containing Amino Acid Mimetic Fragments: New Inhibitors of Wild-Type and Mutated HIV-1 Protease Dimerization. *J. Med. Chem.* **2006**, *49*, 5647-5664.

184. Lee, S.-G.; Chmielewski, J. Rapid Synthesis and In Situ Screening of Potent HIV-1 Protease Dimerization Inhibitors. *Chem. Biol.* **2006**¸*13*, 421-426.

185. Bowman, M. J.; Chmielewshi, J. Sidechain-Linked Inhibitors of HIV-1 Protease Dimerization. *Bioorg. Med. Chem.* **2009**, *17*, 967-976.

186. Daenke, S.; Schramm, H. J.; Bangham, C. R. M. Analysis of Substrate Cleavage by Recombinant Protease of Human T Cell Leukaemia, Virus Type 1 Reveals Preferences and Specificity of Binding. *J. Gen. Virol.* **1994**, *75*, 2233-2239.

187. Quintero, J. J. Computational Studies of the HIV-1 Protease Dimer Interface. University of Michigan. Doctoral Thesis. **2011**.

188. Trylska, J.; Tozzini, V.; Chang, C.-e. A; McCammon, J. A. HIV-1 Protease Substrate Binding and Product Release Pathways Explored with Coarse-Grained Molecular Dynamics. *Biophys. J.* **2007**, *92*, 4179-4187.

189. Chang, M. W.; Giffin, M. J.; Muller, R.; Savage, J.; Lin, Y. C.; Hong, S.; Jin, W.; Whitby, L. R.; Elder, J. H.; Boger, D. L.; Torbett, B. E. Identification of Broad-Based HIV-1 Protease Inhibitors from Combinatorial Libraries. *Biochem. J.* **2010**, *429*, 527-532.

190. Agniswamy, J.; Shen, C.-H.; Anian, A.; Sayer, J. M.; Louis, J. M.; Weber, I. T. HIV-1 Protease with 20 Mutations Exhibits Extreme Resistance to Clinical Inhibitors through Coordinated Structural Rearrangements. *Biochemistry* **2012**, *51*, 2819-2828.

191. Cai, Y.; Yilmaz, N. K.; Myint, W.; Ishima, R.; Schiffer, C. A. Differential Flap Dynamics in Wild-Type and a Drug Resistant Variant of HIV-1 Protease Revealed

Human Immunodeficiency Virus Protease by 'Interface Peptides'. *Antiviral Res.* **1996**, *30*, 155-170.

179. Breccia, P.; Boggetto, N.; Perez-Fernandez, R.; Van Gool, M; Takahashi, M.; Rene, L.; Prados, P.; Badet, B.; Reboud-Ravaux, M.; de Mendoza, J. Dimerization Inhibitors of HIV-1 Protease Based on a Bicyclic Guanidinium Subunit. *J. Med. Chem.* **2003**, *46*, 5196-5207.

180. Zutshi, R.; Franciskovich, J.; Shultz, M.; Schweitzer, B.; Bishop, P.; Wilson, M.; Chmielewski, J. Targeting the Dimerization Interface of HIV-1 Protease: Inhibition with Cross-Linked Interfacial Peptides. *J. Am. Chem. Soc.* **1997**, *119*, 4841-4845.

181. Shultz, M. D.; Bowman, M. J.; Ham, Y.-W.; Zhao, X.; Tora, G.; Chmielewski, J. Small-Molecule Inhibitors of HIV-1 Protease Dimerization Derived from Cross-Linked Interfacial Peptides. *Angew, Chem.* **2000**, *112*, 2822-2825.

182. Bowman, M. J.; Byrne, S.; Chmielewski, J. Switching between Allosteric and Dimerization Inhibition of HIV-1 Protease. *Chem. Biol.* **2005**, *12*, 439-444.

183. Bannwarth, L.; Kessler, A.; Pethe, S.; Collinet, B.; Merabet, N.; Boggetto, N.; Sicsic, S.; Reboud-Ravaux, M.; Ongeri, S. Molecular Tongs, Containing Amino Acid Mimetic Fragments: New Inhibitors of Wild-Type and Mutated HIV-1 Protease Dimerization. *J. Med. Chem.* **2006**, *49*, 5647-5664.

184. Lee, S.-G.; Chmielewski, J. Rapid Synthesis and In Situ Screening of Potent HIV-1 Protease Dimerization Inhibitors. *Chem. Biol.* **2006**¸*13*, 421-426.

185. Bowman, M. J.; Chmielewshi, J. Sidechain-Linked Inhibitors of HIV-1 Protease Dimerization. *Bioorg. Med. Chem.* **2009**, *17*, 967-976.

186. Daenke, S.; Schramm, H. J.; Bangham, C. R. M. Analysis of Substrate Cleavage by Recombinant Protease of Human T Cell Leukaemia, Virus Type 1 Reveals Preferences and Specificity of Binding. *J. Gen. Virol.* **1994**, *75*, 2233-2239.

187. Quintero, J. J. Computational Studies of the HIV-1 Protease Dimer Interface. University of Michigan. Doctoral Thesis. **2011**.

188. Trylska, J.; Tozzini, V.; Chang, C.-e. A; McCammon, J. A. HIV-1 Protease Substrate Binding and Product Release Pathways Explored with Coarse-Grained Molecular Dynamics. *Biophys. J.* **2007**, *92*, 4179-4187.

189. Chang, M. W.; Giffin, M. J.; Muller, R.; Savage, J.; Lin, Y. C.; Hong, S.; Jin, W.; Whitby, L. R.; Elder, J. H.; Boger, D. L.; Torbett, B. E. Identification of Broad-Based HIV-1 Protease Inhibitors from Combinatorial Libraries. *Biochem. J.* **2010**, *429*, 527-532.

190. Agniswamy, J.; Shen, C.-H.; Anian, A.; Sayer, J. M.; Louis, J. M.; Weber, I. T. HIV-1 Protease with 20 Mutations Exhibits Extreme Resistance to Clinical Inhibitors through Coordinated Structural Rearrangements. *Biochemistry* **2012**, *51*, 2819-2828.

191. Cai, Y.; Yilmaz, N. K.; Myint, W.; Ishima, R.; Schiffer, C. A. Differential Flap Dynamics in Wild-Type and a Drug Resistant Variant of HIV-1 Protease Revealed

by Molecular Dynamics and NMR Relaxation. *J. Chem. Theory Comput*. **2012**, *8*, 3452-3462.

192. Perryman, A. L.; Lin, J.-H.; McCammon, J. A. HIV-1 Protease Molecular Dyanamics of a Wild-Type and of the V82F/I84V Mutant: Possible Contributions to Drug Resistance and a Potential New Target site for Drugs. *Protein Sci.* **2004**, *3*, 1108-1123.

193. Mittal, S.; Cai, Y.; Nalam, M. N. L.; Bolon, D. N. A.; Schiffer, C. A. Hydrophobic Core Flexibility Modulates Enzyme Activity in HIV-1 Protease. *J. Am. Chem. Soc.* **2012**, *134*, 4163-4168.

194. Perryman, A. L.; Zhang, Q.; Soutter, H. H.; Rosenfeld, R.; McRee, D. E.; Olsen, A. J.; Elder, J. E.; Stout, C. D. Fragment-Based Screen Against HIV Protease. *Chem. Biol. Drug Des.* **2010**, *75*, 257-268.

195. Seelmeier, S.; Schmidt, H.; Turk, V.; von der Helm, K. Human Immunodeficiency Virus Has an Aspartic-Type Protease that Can Be Inhibited by Pepstatin A. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 6612-6616.

196. Dunn, B. M.; Gustchina, A.; Wlodawer, A.; Kay, J. Subsite Preferences of Retroviral Proteinases. *Methods Enzymol.* **1994**, *241*, 254-278.

197. Zimmerman, S. B.; Minton, A. P. Estimation of Macromolecule Concentrations and Excluded Volume Effects for the Cytoplasm of *Escherichia coli*. *J. Mol. Biol.* **1991**, *222*, 599-620.

198. Ellis, R. J.; Minton, A. P. Protein Aggregation in Crowded Environments. *Biol. Chem.* **2006**, *387*, 485-497.

199. Zou, Z.; Cao, L.; Zhou, P.; Su, Y.; Sun, Y.; Li, W. Hyper-acidic Protein Fusion Partners Improve Solubility and Assist Correct Folding of Recombinant Proteins Expressed in *Escherichia coli*. *J. Biotechnol.* **2008**, *135*, 333-339.

200. Balch, W. E.; Morimoto, R. I.; Dillin, A.; Kelly, J. W. Adapting Proteostasis for Disease Intervention. *Science* **2008**, *319*, 916-919.

201. Frydman, J. Folding of Newly Translated Proteins *in vivo*: The Role of Molecular Chaperones. *Annu. Rev. Biochem.* **2001**, *70*, 603-647.

202. Bukau, B.; Weissman, J.; Horwich, A. Molecular Chaperones and Protein Quality Control. *Cell* **2006**, *125*, 443-451.

203. Taipale, M.; Jarosz, D. F.; Lindquist, S. Hsp90 at the Hub of Protein Homeostasis: Emerging Mechanistic Insights. *Nature Rev. Mol. Cell Biol.* **2010**, *11*, 515-528.

204. Boorstein, W. R.; Ziegelhoffer, T.; Craig, E. A. Molecular Evolution of the HSP70 Multigene Family. *J. Mol. Evol.* **1994**, *38*, 1-17.

205. Mayer, M. P.; Brehmer, D.; Gassler, C. S.; Bukau, B. Hsp70 Chaperone Machines. *Adv. Protein Chem.* **2001**, *59*, 1-44.

206. Daugaard, M.; Rohde, M.; Jaattela, M. The Heat Shock Protein 70 Family: Highly Homologous Proteins with Overlapping and Distinct Functions. *FEBS Lett.* **2007**, *581*, 3702-3710.

207. Patury, S.; Miyata, Y.; Gestwicki, J. E. Pharmacological Targeting of the Hsp70 Chaperone. *Curr. Top. Med. Chem.* **2009**, *9*, 1337-1351.

208. Craig, E. A. The Heat Shock Response. *CRC Crit. Rev. Biochem.* **1985**, *18*, 239-280.

209. Lindquist, S. The Heat-shock Response. *Ann. Rev. Biochem.* **1986**, *55*, 1151-1191.

210. Liberek, K.; Lewandowska, A.; Zietkiewicz, S. Chaperones in Control of Protein Disaggregation. *EMBO J.* **2008**, *27*, 328-335.

211. Pratt, W. B.; Toft, D. O. Regulation of Signaling Protein Function and Trafficking by the Hsp90/Hsp70-based Chaperone Machinery. *Exp. Biol. Med. (Maywood, NJ, U.S.)* **2003**, *228*, 111-133.

212. Young, J. C.; Hoogenraad, N. J.; Hartl, F. U. Molecular Chaperones Hsp90 and Hsp70 Deliver Preproteins to the Mitochondrial Import Receptor Tom70. *Cell* **2003**, *112*, 41-50.

213. De Los Rios, P.; Ben-Zvi, A.; Slutsky, O.; Azem, A.; Goloubinoff, P. Hsp70 Chaperones Accelerate Protein Translocation and the Unfolding of Stable Protein Aggregates by Entropic Pulling. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 6166-6171.

214. Hohfeld, J.; Cyr, D. M.; Patterson, C. From the Cradle to the Grave: Molecular Chaperones that May Choose Between Folding and Degradation. *EMBO Rep.* **2001**, *2*, 885-890.

215. McClellan, A. J.; Tam, S.; Kaganovich, D.; Frydman, J. Protein Quality Control: Chaperones Culling Corrupt Conformations. *Nat. Cell Biol.* **2005**, *7*, 736-741.

216. Bukau, B.; Horwich, A. L. The Hsp70 and Hsp60 Chaperone Machines. *Cell* **1998**, *92*, 351-366.

217. Albanèse, V.; Yam, A. Y.-W.; Baughman, J.; Parnot, C.; Frydman, J. Systems Analyses Reveal Two Chaperone Networks with Distinct Functions in Eukaryotic Cells. *Cell* **2006**, *124*, 75-88.

218. Kampinga, H. H.; Craig, E. A. The HSP70 Chaperone Machinery: J Proteins as Drivers of Functional Specificity. *Mol. Cell Biol.* **2010**, *11*, 579-592.

219. Evans, C. G.; Chang, L.; Gestwicki, J. E. Heat Shock Protein 70 (Hsp70) as an Emerging Drug Target. *J. Med. Chem.* **2010**, *53*, 4585-4602.

220. Ciocca, D. R.; Calderwood, S. K. Heat Shock Proteins in Cancer: Diagnostic, Prognostic, Predictive, and Treatment Implications. *Cell Stress Chaperones* **2005**, *10*, 86-103.

221. Rohde, M.; Daugaard, M.; Jensen, M. H.; Helin, K.; Nylandsted, J.; Jaattela, M. Members of the Heat-Shock Protein 70 Family Promote Cancer Cell Growth by Distinct Mechanisms. *Genes & Dev.* **2005**, *19*, 570-582.

222. Buzzard, K. A.; Giaccia, A. J.; Killender, M.; Anderson, R. L. Heat Shock Protein 72 Modulates Pathways of Stress-induced Apoptosis. *J. Biol. Chem.* **1998**, *273*, 17147-17153.

223. Garrido, C.; Gurbuxani, S.; Ravagnan, L.; Kroemer, G. Heat Shock Proteins: Endogenous Modulators of Apoptotic Cell Death. *Biochem. Biophys. Res. Commun.* **2001**, *286*, 433-442.

224. Powers, M. V.; Clarke, P. A.; Workman, P. Dual Targeting of HSC70 and HSP72 Inhibits HSP90 Function and Induces Tumor-Specific Apoptosis. *Cancer Cell* **2008**, *14*, 250-262.

225. Evans, C. G.; Wisen, S.; Gestwicki, J. E. Heat Shock Proteins 70 and 90 Inhibit Early Stages of Amyloid Beta-(1-42) aggregation in vitro. *J. Biol. Chem.* **2006**, *281*, 33182-33191.

226. Kumar, P.; Ambasta, R. K.; Veereshwarayya, V.; Rosen, K. M.; Kosik, K. S.; Band, H.; Mestril, R.; Patterson, C.; Querfurth, H. W. CHIP and HSPs Interact with Beta-APP in a Proteasome-Dependent Manner and Influence Abeta Metabolism. *Hum. Mol. Genet.* **2007**, *16*, 848-864.

227. Petrucelli, L.; Dickson, D.; Kehoe, K.; Taylor, J.; Snyder, H.; Grover, A.; De Lucia, M.; McGowan, E.; Lewis, J.; Prihar, G.; Kim, K.; Dillmann, W. H.; Browne, S. E.; Hall, A.; Voellmy, R.; Tsuboi, Y.; Dawson, T. M.; Wolozin, B.; Hardy, J.; Hutton, M. CHIP and Hsp70 Regulate Tau Ubiquitination, Degradation and Aggregation. *Hum. Mol. Genet.* **2004**, *13*, 703-714.

228. Evans, C. G.; Jinwal, U. K.; Makley, L. N.; Dickey, C. A.; Gestwicki, J. E. Identification of Dihydropyridines That Reduce Cellular Tau Levels. *Chem. Commun.* **2011**, *47*, 529-531.

229. Miyata, Y.; Koren, J. III; Kiray, J.; Dickey, C. A.; Gestwicki, J. E. Molecular Chaperones and Regulation of Tau Quality Control: Strategies for Drug Discovery. *Future Med. Chem.* **2011**, *3*, 1523-1537.

230. Jinwal, U. K.; Koren, J.; O'Leary, J. C.; Jones, J. R.; Abisambra, J. F.; Dickey, C. A. Hsp70 ATPase Modulators as Therapeutics for Alzheimer's and Other Neurodegenerative Diseases. *Mol. Cell Pharmacol.* **2010**, *2*, 43-46.

231. Hoshino, T.; Murao, N.; Namba, T.; Takehara, M.; Adachi, H.; Katsuno, M.; Sobue, G.; Matsushima, T.; Suzuki, T.; Mizushima, T. Suppression of Alzheimer's Disease-Related Phenotypes by Expression of Heat Shock Protein 70 in Mice. *J. Neurosci.* **2011**, *31*, 5225-5234.

232. Huesca, M.; Goodwin, A.; Bhagwansingh, A.; Hoffman, P.; Lingwood, C. A. Characterization of an Acidic-pH-Inducible Stress Protein (Hsp70), a Putative Sulfatide Binding Adhesin, from *Helicobacter pylori*. *Infect. Immun.* **1998**, *66*, 4061-4067.

233. Macellaro, A.; Tujulin, E.; Hjalmarsson, K.; Norlander, L. Identification of a 71-Kilodalton Surface-Associated Hsp70 Holologue in *Coxiella burnetii*. *Infect. Immun.* **1998**, *66*, 5882-5888.

234. de Jesus, M. C.; Urban, A. A.; Marasigan, M. E.; Barnett Foster, D. E. Acid and Bile-Salt Stress of Enteropathogenic *Escherichia coli* Enhances Adhesion to Epithelial Cells and Alters Glycolipid Receptor Binding Specificity. *J. Infect. Dis.* **2005**, *192*, 1430-1440.

235. Henderson, B.; Allan, E.; Coates, A. R. M. Stress Wars: The Direct Role of Host and Bacterial Molecular Chaperones in Bacterial Infection. *Infect. Immun.* **2006**, *74*, 3693-3706.

236. Paek, K. H.; Walker, G. C. *Escherichia coli dnaK* Null Mutant Are Inviable at High Temperature. *J. Bacteriol.* **1987**, *169*, 283-290.

237. Bukau, B.; Walker, G. C. Cellular Defects Caused by Deletion of the *Escherichia coli dnaK* Gene Indicate Roles for Heat Shock Protein in Normal Metabolism. *J. Bacteriol.* **1989**, *171*, 2337-2346.

238. Wolska, K. I.; Bugajska, E.; Jurkiewicz, D.; Kuc, M.; Jozwik, A. Antibiotic Susceptibility of *Escherichia coli dnaK* and *dnaJ* Mutants. *Microbial Drug Resis.* **2000**, *6*, 119-126.

239. Singh, V. K.; Utaida, S.; Jackson, L. S.; Jayaswal, R. K.; Wilkinson, B. J.; Chamberlain, N. R. Role for *dnaK* Locus in Tolerance of Multiple Stresses in *Staphylococcus aureus*. *Microbiology* **2007**, *153*, 3162-3173.

240. Credito, K.; Lin, G.; Koeth, L.; Sturgess, M. A.; Appelbaum, P. C. Activity of Levofloxacin Alone and in Combination with a DnaK Inhibitor against Gram-Negative Rods, Including Levofloxacin-Resistant Strains. *Antimicrob. Agents Chemother.* **2009**, *53*, 814-817.

241. Chang, L.; Miyata, Y.; Ung, P. M.-U.; Bertelsen, E. B.; McQuade, T. J.; Carlson, H. A.; Zuiderweg, E. R. P.; Gestwicki, J. E. Chemical Screens Against a Reconstituted Multiproteincomplex: Myricetin Blocks DnaJ Regulation of DnaK through an Allosteric Mechanism. *Chem. Biol.* **2011**, *18*, 210-221.

242. Henderson, B.; Calderwood, S. K.; Coates, A. R. M.; Cohen, I.; van Eden, W.; Lehner, T.; Pockley, A. G. Caught with Their PAMPs Down? The Extracellular Signaling Actions of Molecular Chaperones Are Not Due to Microbial Contaminants. *Cell Stress Chap.* **2010**, *15*, 123-141.

243. Mambula, S. S.; Calderwood, S. K. Heat Shock Protein 70 Is Secreted from Tumor Cells by a Non-Classical Pathway Involving Lysosomal Endosomoes. *J. Immunol.* **2006**, *177*, 7849-7857.

244. Mambula, S. S.; Stevenson, M. A.; Ogawa, K.; Calderwood, S. K. Mechanisms for Hsp70 Secretion: Crossing Membranes without a Leader. *Methods* **2007**, *43*, 168-175.

245. Vabulas, R. M.; Ahmad-Nejad, P.; Ghose, S.; Kirshning, C. J.; Issels, R. D.; Wagner, H. HSP70 as Endogenous Stimulus of the Toll/Interleukin-1 Receptor Signal Pathway. *J. Biol. Chem.* **2002**, *277*, 15107-15112.

246. Caldwood, S. K.; Theriault, J.; Gray, P. J.; Gong, J. Cell Surface Receptors for Molecular Chaperones. *Methods* **2007**, *43*, 199-206.

247. Wang, Y.; Whittall, T.; McGowan, E.; Younson, J.; Kelly, C.; Bergmeier, L. A.; Singh, M.; Lehner, T. Identification of Stimulating and Inhibitory Epitopes within the Heat Shock Protein 70 Molecule that Modulate Cytokine Production and Maturation of Dendritic Cells. *J. Immunol.* **2005**, *174*, 3306-3316.

248. Wang, X. Y.; Facciponte, J.; Chen, X.; Subjeck, J. R.; Repasky, E. A. Scavenger Receptor-A Negatively Regulates Antitumor Immunity. *Cancer Res.* **2007**, *67*, 4996-5002.

249. Futagami, S.; Hiratsuka, T.; Shindo, T.; Hamamoto, T.; Horie, A.; Ueki, N.; Kusunoki, M.; Gudis, K.; Miyake, K.; Tsukui, T.; Sakamoto, C. Extracellular HSP70 Blocks CD40L-Induced Apoptosis and Tubular Formation in Endothelial Cells. *J. Gastroenterol. Hepatol.* **2008**, *2*, S222-228.

250. Mackay, C. R.; Sallusto, F. A New Role for CCR5 in Innate Immunity – Binding to Bacterial Heat Shock Protein 70. *Eur. J. Immunol.* **2006**, *36*, 2293-2295.

251. Pido-Lopez, J.; Whittall, T.; Wang, Y.; Bergmeier, L. A.; Babaahmady, K.; Singh, M.; Lehner, T. Stimulation of Cell Surface CCR5 and CD40 Molecules by Their Ligands or by HSP70 Up-regulate APOBEC3G Expression in CD4(+) T Cells and Dendritic Cells. *J. Immunol.* **2007**, *178*, 1671-1679.

252. Sugiyama, R.; Nishitsuji, H.; Furukawa, A.; Katahira, M.; Habu, Y.; Takeuchi, H.; Ryo, A.; Takaku, H. Heat Shock Protein 70 Inhibits HIV-1 Vif-Mediated Ubiquitination and Degradation of APOBEC3G. *J. Biol. Chem.* **2011**, *286*, 10051-10057.

253. Harrison, C. J.; Hayer-Hartl, M.; Di Liberto, M.; Hartl, F.; Kuriyan, J. Crystal Structure of the Nucleotide Exchange Factor GrpE Bound to the ATPase Domain of the Molecular Chaperone DnaK. *Science* **1997**, *276*, 431-435.

254. Bertelsen, E. B.; Chang, L.; Gestwicki, J. E.; Zuiderwig, E. R. P. Solution Conformation of wild-Type E. coli Hsp70 (DnaK) Chaperone Complexed with ADP and Substrate. *Proc. Nat. Am. Sci. U.S.A.* **2009**, *106*, 8471-8476.

255. Bork, P.; Sander, C.; Valencia, A. An ATPase Domain Common to Prokaryotic Cell Cycle Proteins, Sugar Kinases, Actin and Hsp70 Heat Shock Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 7290-7294.

256. Schüler, H. ATPase Activity and Conformational Changes in the Regulation of Actin. *Biochim. Biophys. Acta.* **2001**, *1549*, 137-147.

257. Bhattacharya, A.; Kurochkin, A. V.; Yip, G. N. B.; Zhang, Y.; Bertelsen, E. B.; Zuiderweg, E. R. P. Allostery in Hsp70 Chaperones Is Transduced by Subdomain Rotations. *J. Mol. Biol.* **2009**, *388*, 475-490.

258. Chang, Y.-W.; Sun, Y.-J.; Wang, C.; Hsiao, C.-D. Crystal Structures of the 70-kDa Heat Shock Proteins in Domain Disjoining Conformation. *J. Biol. Chem.* **2008**, *283*, 15502-15511.

259. Flaherty, K. M.; Deluca-Flaherty, C.; McKay, D. B. Three-Dimensional Structure of the ATPase Fragment of a 70K Heat-Shock Cognate Protein. *Nature* **1990**, *346*, 623-628.

260. Palleros, D. R.; Reid, K. L.; Shi, L.; Welch, W. J.; Fink, A. L. ATP-Induced Protein Hsp70 Complex Dissociation Requires K+ But Not ATP Hydrolysis. *Nature* **1993**, *365*, 664-666.

261. Zhu, X.; Zhao, X.; Burkholder, W. F.; Gragerov, A.; Ogata, C. M.; Gottesman, M. E.; Hendrickson, W. A. Structural Analysis of Substrate Binding by the Molecular Chaperone DnaK. *Science* **1996**, *272*, 1606-1614.

262. Rüdiger, S.; Germeroth, L.; Schneider-Mergener, J.; Bukau, B. Substrate Specificity of the DnaK Chaperone Determined by Screening Cellulose-Bound Peptide Libraries. *EMBO J.* **1997**, *16*, 1501-1507.

263. Mayer, M. P.; Bukau, B. Hsp70 Chaperones: Cellular Functions and Molecular Mechanism. *Cell. Mol. Life Sci.* **2005**, *62*, 670-684.

264. Jiang J.; Prasad, K.; Lafer, E. M.; Sousa, R. Structural Basis of Interdomain Communication in the Hsc70 Chaperone. *Mol. Cell* **2005**, *20*, 513-524.

265. Swain, J. F.; Dinler, G.; Sivendran, R.; Montgomery, D. L.; Stotz, M.; Gierasch, L. M. Hsp70 Chaperone Ligands Control Domain Association via an Allosteric Mechanism Mediated by the Interdomain Linker. *Mol. Cell* **2007**, *26*, 27-39.

266. Liberek, K.; Skowyra, D.; Zylicz, M.; Johnson, C.; Georgopoulos, C. The Escherichia coli DnaK Chaperone, the 70-kDa Heat Shock Protein Eukaryotic Equivalent, Changes Conformation Upon ATP Hydrolysis, Thus Triggering Its Dissociation from a Bound Target Protein. *J. Biol. Chem.* **1991**, *266*, 14491-14496.

267. Wang, H.; Kurochkin, A. V.; Pang, Y.; Hu, W.; Flynn, G. C.; Zuiderweg, E. R. NMR Solution Structure of the 21 kDa Chaperone Protein DnaK Substrate Binding Domain: A Preview of Chaperone Protein Interaction. *Biochemistry* **1998** *37*, 7929-7940.

268. Laufen, T.; Mayer, M. P.; Beisel, C.; Klostermeier, D.; Mogk, A.; Reinstein, J.; Bukau, B. Mechanism of Regulation of Hsp70 Chaperones by DnaJ Cochaperones. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 5452-5457.

269. Zylicz, M.; LeBowitz, J. H.; McMacken, R.; Georgopoulos, C. The dnaK Protein of *Escherichia coli* Possesses an ATPase and Autophosphorylation Activity and Is Essential in an *in Vitro* DNA Replication System. *Proc. Natl. Acad. Sci. U.S.A.* **1983**, *80*, 6431-6435.

270. Sadis, S.; Hightower, L .E. Unfolded Proteins Stimulate Molecular Chaperone Hsc70 ATPase by Accelerating ADP/ATP Exchange. *Biochemistry* **1992**, *31*, 9406-9412.

271. McCarty, J. S.; Buchberger, A.; Reinstein, J.; Bukau, B. The Role of ATP in the Functional Cycle of the DnaK Chaperone System. *J. Mol. Biol.* **1995**, *249*, 126-137.

272. Szabo, A.; Langer, T.; Schroder, H.; Flanagan, J.; Bukau, B.; Hartl, F. U. The ATP Hydrolysis-Dependent Reaction Cycle of the *Escherichia coli* Hsp70 System – DnaK, DnaJ, and GrpE. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 10345-10349.

273. Gamer, J.; Multhaup, G.; Tomoyasu, T.; McCarty, J. S.; Rudiger, S.; Schonfeld, H.-J.; Schirra, C.; Bujard, H.; Bukau, B. A Cycle of Binding and Release of the DnaK, DnaJ and GrpE Chaperones Regulates Activity of the *Escherichia coli* Heat Shock Transcription Factor Sigma32. *EMBO J.* **1996**, *15*, 607-617.

274. Huang, K.; Flanagan, J. M.; Prestegard, J. H. The Influence of C-terminal Extension on the Structure of the "J-Domain" in *E. coli* DnaJ. *Prot. Sci.* **1999**, *8*, 203-214.

275. Genevaux, P.; Schwager, F.; Georgopoulos, C.; Kelley, W. L. Scanning Mutagenesis Identifies Amino Acid Residues Essential for the *in Vivo* Activity of the *Escherichia coli* DnaJ (Hsp40) J-Domain. *Genetics* **2002**, *162*, 1045-1053.

276. Wall, D.; Zylicz, M.; Georgopoulos, C. The NH2-Terminal 108 Amino Acids of the *Escherichia coli* DnaJ Protein Stimulate the ATPase Activity of DnaK and Are Sufficient for Delta Replication. *J. Biol. Chem.* **1994**, *269*, 5446-5451.

277. Karzai, A. W.; McMacken, R. A Bipartite Signaling Mechanism Involved in DnaJ-Mediated Activation of the *Escherichia coli* DnaK Protein. *J. Biol. Chem.* **1996**, *271*, 11236-11246.

278. Sha, B.; Lee, S.; Cyr, D. M. The Crystal Structure of the Peptide-Binding Fragment from the Yeast Hsp40 Protein Sis1. *Structure* **2000**, *8*, 799-807.

279. Srinivasan, S. R.; Gillies, A. T.; Chang, L.; Thompson, A. D.; Gestwicki, J. E. Molecular Chaperones DnaK and DnaJ Share Predicted Binding Sites on Most Proteins in the *E. coli* Proteome. *Mol. BioSyst.* **2012**, *8*, 2323-2333.

280. Gässler, C. S.; Buchberger, A.; Laufen, T.; Mayer, M. P.; Schröder, H.; Valencia, A.; Bukau, B. Mutations in the DnaK Chaperone Affecting Interaction with the DnaJ Cochaperone. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 15229-15234.

281. Suh, W. C.; Burkholder, W. F.; Lu, C. Z.; Zhao, X.; Gottesman, M. E.; Gross, C. A. Interaction of the Hsp70 Molecular Chaperone, DnaK, with Its Co-chaperone DnaJ. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 15223-15228.

282. Ahmad, A.; Bhattacharya, A.; McDonald, R. A.; Cordes, M.; Ellington, B.; Bertelsen, E. B.; Zuiderweg, E. R. P. Heat Shock Protein 70 kDa Chaperone/DnaJ Cochaperone Complex Employs an Unusual Dynamic Interface. *Proc. Nat. Acad. Sci. U.S.A.* **2011**, *108*, 18966-18971.

283. Cyr, D. M. Swapping Nucleotides, Tuning Hsp70. *Cell* **2008**, *133*, 945-947.

284. Pachschies, L.; Theyssen, H.; Buchberger, A.; Bukau, B.; Goody, R. S.; Reinstein, J. GrpE Accelerates Nucleotide Exchange of the Molecular Chaperone DnaK with an Associative Displacement Mechanism. *Biochemistry* **1997**, *36*, 3417-3422.

285. Bohen, S. P.; Kralli, A.; Yamamoto, K. R. Hold'em and Fold'em: Chaperones and Signal Transduction. *Science* **1996**, *268*, 1303-1304.

286. Genest, O.; Hoskins, J. R.; Camberg, J. L.; Doyle, S. M.; Wickner, S. Heat Shock Protein 90 from *Escherichia coli* Collaborates with the DnaK Chaperone System in Client Protein Remodeling. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 8206-8211.

287. Diamant, S.; Goloubinoff, P. Temperature-Controlled Activity of DnaK-DnaK-GrpE Chaperones: Protein-Folding Arrest and Recovery During and After Heat Shock Depends on the Substrate Protein and the GrpE Concentration. *Biochemistry* **1998**, *37*, 9688-9694.

288. Nollen, E. A.; Brunsting, J. F.; Song, J.; Kampinga, H. H.; Morimoto, R. I. Bag 1 Functions In Vivo As a Negative Regulator of Hsp70 Chaperone Activity. *Mol. Cell Biol.* **2000**, *20*, 1083-1088.

289. Gässler, C. S.; Wiederkehr, T.; Brehmer, D.; Bukau, B.; Mayer, M. P. Bag-1M Accelerates Nucleotide Release for Human Hsc70 and Hsp70 and Can Act Concentration-Dependent as Positive and Negative Cofactor. *J. Biol. Chem.* **2001**, *276*, 32538-32544.

290. Redford, N. B.; Fina, M.; Benjamin, I. J.; Moreadith, R. W.; Graves, K. H.; Zhao, P.; Gawa, S.; Wiethoff, A.; Sherry, A. D.; Malloy, C. R.; Williams, R. S. Cardioprotective Effects of 70-kDa Heat Shock Protein in Transgenic Mice. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 2339-2342.

291. Cummings, C. J.; Sun, Y.; Opal, P.; Antalffy, B.; Mestril, R.; Orr, H. T.; Dillmann, W. H.; Zoghbi, H. Y. Over-Expression of Inducible HSP70 Chaperone Suppresses Neuropathology and Improves Motor Function in SCA1 Mice. *Hum. Mol. Genet.* **2001**, *10*, 1511-1518.

292. Jaattela, M. Over-Expression of Hsp70 Confers Tumorigenicity to Mouse Fibrosarcoma Cells. *Int. J. Cancer* **1995**, *60*, 689-693.

293. Nylandsted, J.; Wick, W.; Hirt, U. A.; Brand, K.; Rohde, M.; Leist, M.; Weller, M.; Jaattela, M. Eradication of Glioblastoma, and Breast and Colon Carcinoma Xenografts by Hsp70 Depletion. *Cancer Res.* **2002**, *62*, 7139-7142.

294. Liebscher, M.; Jahreis, G.; Lucke, C.; Grabley, S.; Raina, S.; Schiene-Fischer, C.; Fatty Acyl Benzamido Antibacterials Based on Inhibition of DnaK-Catalyzed Protein Folding. *J. Biol. Chem.* **2007**, *282*, 4437-4446.

295. Walsh, P.; Bursac, D.; Law, Y. C.; Cyr, D.; Lithgow, T. The J-Protein Family: Modulating Protein Assembly, Disassembly and Translocation. *EMBO Rep.* **2004**, *5*, 567-571.

296. Qiu, X.-B.; Shao, Y.-M.; Miao, S.; Wang, L. The Diversity of the DnaJ/Hsp40 Family, the Crucial Partner for Hsp70 Chaperones. *Cell. Mol. Life Sci.* **2006**, *63*, 2560-2570.

297. Takayama, S.; Reed, J. C. Molecular Chaperone Targeting and Regulation by BAG Family Proteins. *Nature Cell Biol.* **2001**, *3*, 237-241.

298. Dragovic, Z.; Broadley, S. A.; Shomura, Y.; Bracher, A.; Hartl, F. U. Molecular Chaperones of the Hsp110 Family Act as Nucleotide Exchange Factors of Hsp70s. *EMBO J.* **2006**, *25*, 2519-2528.

299. Chang, L.; Bertelsen, E. B.; Wisen, S.; Larsen, E. M.; Zuiderweg, E. R. P.; Gestwicki, J. E. High-Throughput Screen for Small Molecules That Modulate the ATPase Activity of the Molecular Chaperone DnaK. *Anal. Biochem.* **2008**, *372*, 167-176.

300. Miyata, Y.; Chang, L.; Bainor, A.; McQuade, T. J.; Walczak, C. P.; Zhang, Y.; Larsen, M. J.; Kirchhoff, P.; Gestwicki, J. E. High-Throughput Screen for

Escherichia coli Heat Shock Protein 70 (Hsp70/DnaK): ATPase Assay in Low Volume by Exploiting Energy Transfer. *J. Biomol. Screen.* **2010**, *15*, 1211-1219.

301. Wisen, S.; Bertelsen, E. B; Thompson, A. D.; Patury, S.; Ung, P.; Chang, L.; Evans, C. G.; Walter, G. M.; Wipf, P.; Carlson, H. A.; Brodsky, J. L.; Zuiderweg, E. R. P.; Gestwicki, J. E. Binding of a Small Molecule at a Protein-Protein Interface Regulates the Chaperone Activity of Hsp70-Hsp40. *Chem. Biol.* **2010**, *5*, 611-622.

302. Vogel, M.; Bukau, B.; Mayer, M. P. Allosteric Regulation of Hsp70 Chaperone by a Proline Switch. *Mol. Cell* **2006**, *21*, 359-367.

303. Woo, H.-J.; Jiang, J.; Lafer, E. M.; Sousa, R. ATP-Induced Conformational Changes in Hsp70: Molecular Dynamics and Experimental Validation of an in Silico Predicted Conformation. *Biochemistry* **2009**, *48*, 11470-11477.

304. Nicolaï, A.; Senet, P.; Delarue, P.; Ripoll, D. R. Human Inducible Hsp70: Structures, Dynamics, and Interdomain Communication from All-Atom Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2010**, *6*, 2501-2519.

305. Alder, B. J.; Wainwright, T. E. Phrase Transition for Hard Sphere System. *J. Chem. Phys.* **1957**, *27*, 1208-1209.

306. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc*. **1995**, *117*, 5179-5197.

307. Beauchamp, K. A.; Lin, Y.-S.; Das, R.; Pande, V. S. Are Protein Force Fields Getting Better? A Systemic Benchmark on 524 Diverse NMR Measurements. *J. Chem. Theory Comput.* **2012**, *8*, 1409-1414.

308. Case, D. A.; Darden, T. A.; Cheatham, III., T.E. ; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D.R.; Mathews, D.H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER 11*, University of California, San Francisco.: **2010**.

309. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. Charmm – A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comp. Chem.* **1983**, *4*, 187-217.

310. Hermans, J.; Berendsen, H. J. C.; van Gunsteren, W. F.; Postma, J. P. M. A Consistent Empirical Potential for Water-Protein Interactions. *Biopolymers* **1984**, *23*, 1513-1518.

311. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225-11236.

312. Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee. T. A Point-Charge Force Field for Molecular Mechanics

Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *J. Comput. Chem.***2003**, *24*, 1999-2012.

313. Cheatham, T. E. Simulations and Modeling of Nucleic Acid Structure, Dynamics and Interactions. *Curr. Opin. Struct. Biol*. **2004**, *14*, 360-367.

314. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins* **2006**, *65*, 712-725.

315. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586-3616.

316. MacKerell, A. D.; Banavali, N.; Foloppe, N. Development and Current Status of the CHARMM Force Field for Nucleic Acids. *Biopolymers* **2001**, *56*, 257-265.

317. Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; Gonzalez-Outeirino, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. GLYCAM06: A Generalizable Biomolecular Force Field. Carbohydrates. *J. Comput. Chem.* **2008**, *29*, 622-655.

318. Halgren, T. A. Merck Molecular Force Field. I. Basis Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490-519.

319. Jakalian, A; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comp. Chem.* **2000**, *21*, 132-146.

320. Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comp. Chem.* **2002**, *23*, 1623-1641.

321. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A. Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157-1174.

322. McCammon, J. A. Molecular Dynamics Study of the Bovine Pancreatic Trypsin Inhibitor. *Models for Protein Dynamics*; CECAM: Orsay, France, **1976**; p137.

323. McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of Folded Proteins. *Nature* **1977**, *267*, 585-590.

324. Adcock, S. A.; McCammon, J. A. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.* **2006**, *106*, 1589-1615.

325. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of *n*-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327-341.

326. Scheraga, H. A.; Khalili, M.; Liwo, A. Protein-Folding Dynamics: Overview of Molecular Simulation Techniques. *Annu. Rev. Phys. Chem.* **2007**, *58*, 57-83.

327. Jorgensen, W. L. Optimized Intermolecular Potential Functions for Liquid Alcohols. *J. Phys. Chem.* **1986**, *90*, 1276-1284.

328. Jorgensen, W. L.; McDonald, N. A. Development of an All-Atom Force Field for Hetercycles. Properties of Liquid Pyridine and Diazenes. *J. Mol. Struct. (Theochem)* **1998**, *424*, 145-155.

329. Darden, T. A.; York, D. M.; Pedersen, L. G. Particle Mesh Ewald. An N.log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089-10092.

330. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577-8593.

331. Crowley, M. F.; Darden, T. A.; Cheatham, T. E. III; Deerfield, D. W. II. Adventures in Improving the Scaling and Accuracy of a Parallel Molecular Dynamics Program. *J. Supercomput.* **1997**, *11*, 255-278.

332. Loncharich, R. J.; Brooks, B. R.; Paster, R. W. Langevin Dynamics of Peptides: The Frictional Dependence of Isomerization Rates of N-actylananyl-N'-methylamide. *Biopolymers* **1992**, *32*, 523-535.

333. Edinger, S. R.; Cortis, C.; Shenkin, P. S.; Friesner, R. A. Solvation Free Energies of Peptides: Comparison of Approximate Continuum Solvation Models with Accurate Solution of the Poisson-Boltzmann Equation. *J. Phys. Chem. B* **1997**, *101*, 1190-1197.

334. Cramer, C. J.; Truhlar, D. G. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chem. Rev.* **1999**, *99*, 2161-2200.

335. Bashford, D.; Case, D. A. Generalized Born Models of Macromolecular Solvation Effects. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129-152.

336. Honig, B.; Nicholls, A. Classical Electrostatics in Biology and Chemistry. *Science* **1995**, *268*, 1144-1149.

337. Onufriev, A.; Bashford, D.; Case, C. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins* **2004**, *55*, 383-394.

338. García, A. E. Large-Amplitude Nonlinear Motions in Proteins. *Phys. Rev. Lett.* **1992**, *68*, 2696-2699.

339. Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential Dynamics of Proteins. *Proteins* **1993**, *17*, 412-425.

340. Mongan, J. Interactive Essential Dynamics. *J. Comp.-aided Mol. Des.* **2004**, *18*, 433-436.

341. Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269-288.

342. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for drug Discovery: Methods and Application. *Nat. Rev. Drug Discov.* **2004**, *3*, 935-948.

343. Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the Development of Universial, Fast and Highly Accurate Docking/Scoring Methods: A Long Way to Go. *Brit. J. Pharmaco.* **2008**, *153*, S7-S26.

344. Trott, O.; Olson, A. J. Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461.

345. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470-489.

346. FRED version 2.2.5, OpenEye Scientific Software, Inc., Santa Fe, NM, **2009**.

347. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.

348. Glide, version 5.5, Schrödinger, LLC, New York, **2009**.

349. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727-748.

350. Abagyan, R. A.; Mazur, A. K. New Methodology for Computer-Aided Modeling of Biomolecular Structure and Dynamics. 2. Local Deformations and Cycles. *J. Biomol. Struct. Dyn.* **1989**, *6*, 833-845.

351. Goodsell, D. S.; Olson, A. J. Automated Docking of Substrates to Proteins by Simulated Annealing. *Proteins* **1990**, *8*, 195-202.

352. Wang, R.; Lu, Y.; Wang, S. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem.* **2003**, *46*, 2287-2303.

353. Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors. *J. Chem. Inf. Model.* **2006**, *46*, 401–415.

354. Rajamani, R.; Good, A. C. Ranking Poses in Structure-Based Lead Discovery and Optimization: Current Trends in Scoring Function Development. *Curr. Opin. Drug Discov. Devel.* **2007**, *10*, 308-315.

355. Cheng, T.; Li, X.; Li, Y.; Liu, Z. C.; Wang, R. Comparative Assessment of Scoring Functions on a Diverse Test Set. *J. Chem. Inf. Model.* **2009**, *49*, 1079–1093.

356. Englebienne, P.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 4. Are Popular Scoring Functions Accurate for This Class of Proteins? *J. Chem. Inf. Model.* **2009**, *49*, 1568–1580.

357. Smith, R. D.; Dunbar, J. B.; Ung, P. M.-U.; Esposito, E. X.; Yang, C.-Y.; Wang, S.; Carlson, H. A. CSAR Benchmark Exercise of 2010: Combined Evaluation Across All Submitted Scoring Functions. *J. Chem. Inf. Model.* **2011**, *51*, 2115-2131.

358. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical Scoring Functions: I. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. *J. Comput. Aided Mol. Des.* **1997**, *11*, 425–445.

359. Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions. *J. Mol. Biol.* **2000**, *295*, 337–356.

360. Yang, C.-Y.; Wang, R.; Wang, S. M-Score: A Knowledge-Based Potential Scoring Function Accounting for Protein Atom Mobility. *J. Med. Chem.* **2006**, *49*, 5903–5911.

361. Wang, R.; Lai, L.; Wang, S. Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. *J. Comput. Aided Mol. Des.* **2002**, *16*, 11–26.

362. Davis, I. W.; Baker, D. ROSETTALIGAND Docking with Full Ligand and Receptor Flexibility. *J. Mol. Biol.* **2009**, *385*, 381-392.

363. Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42*, 5100-5109.

364. Terp, G. E.; Johansen, B. N.; Christensien, I. T.; Jorgensen, F. A New Concept for Multidimensional Selection of Ligand Conformations (Multiselect) and Multidimensional Scoring (Multiscore) of Protein-Ligand Binding Affinities. *J. Med. Chem.* **2001**, *44*, 2333-2343.

365. Feher, M. Consensus Scoring for Protein-Ligand Interactions. *Drug Dis. Today* **2006**, *11*, 421-428.

366. Cole, J. C.; Murray, C. W.; Nissink, W. M.; Taylor, R. D.; Taylor, R. Comparing Protein-Ligand Docking Programs Is Difficult. *Proteins* **2005**, *60*, 325-332.

367. Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912-5931.

368. Yuriev, E.; Agostino, M.; Ramsland, P. A. Challenges and Advances in Computational Docking: 2009 in Review. *J. Mol. Recognit.* **2010**, *24*, 149-164.

369. Dunbar, J. B.; Smith, R. D.; Yang, C. Y.; Ung, P. M.-U.; Lexa, K. L.; Khazanov, N. A.; Stuckey, J. A.; Wang, S.; Carlson, H. A. CSAR Benchmark Exercise of 2010: Selection of the Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2011**, *51*, 2036-2046.

CHAPTER 2

Allosteric Modulation of HIV-1 Protease via
Targeting the Flap-Tip Recognition Site

2.1 Abstract

NMR and molecular dynamics simulations have demonstrated that the β-hairpins, or flaps, of HIV-1 protease can adopt a range of conformations and that these conformational states are strongly associated with the enzymatic activity of the protease. A multiple protein structure pharmacophore model for high-throughput virtual screening was generated for allosteric site located between the flap and the main body of the protease, first described and named the Eye site by Damm *et al.*[1] The result from the screen was combined with a ligand-based, lead-hopping method, and through this effort a novel compound, NIT, was identified. NIT demonstrated the capability to modulate catalytic activity of HIV protease (HIV-1p) in a manner that is independent of the presence of an active site inhibitor such as pepstatin A and is shown to act on an allosteric site other than the dimerization interface. Furthermore, it is equally potent against wild-type and multi-drug resistant mutant of HIV-1p. Molecular dynamics

simulations of the ligand-protein complex show that compound NIT remains stably bound to the Eye site throughout the trajectories and affects the dynamics of the flaps when compared to an apo-protease simulation. Furthermore, NIT is shown to be in contact with several highly conserved residues in the Eye site (Gly49, Gly78, Pro79, Thr80, and Pro81) throughout the simulations. These results are consistent with the crystallographic evidence of the existence of the Eye site,[2] and it shows this to be a viable site for allosteric modulation of the enzyme.

2.2 Introduction

Proteins are inherently dynamic and conformationally heterogenic. It is generally recognized that they exist in an ensemble of differently populated conformational states in equilibrium, where certain conformations play crucial roles in protein functions, such as enzymatic activity and molecular recognition.[3,4] Therefore, it may be possible to design ligands that specifically target certain conformational states of a protein, such as the "native" state (lowest energy, most populated state) and "lock" the enzyme in the catalytically inactive state.[5-7] An example of such modulation is the binding of the non-nucleoside reverse transcriptase inhibitor, nevirapine, to an allosteric site of HIV reverse transcriptase. The small molecule effects inhibitory activity through selectively binding and locking the enzyme in a catalytically inactive conformational state, thus shifting the equilibrium of the conformational ensemble and depopulating the catalytically active conformational state essential for the enzymatic activity.[8]

The aforementioned phenomenon can also be applied to other protein systems to modulate enzymatic activity. In this study, we focus on the clinically important HIV-1 protease (HIV-1p). HIV-1p is a C_2-symmetric homodimeric protease, shown in Figure 2.1 A. It is critical in the maturation of the infective HIV virion[9] as it cleaves the *gag* and *gag-pol* polyproteins to release the structural proteins (MA, CA, NC, and p6) and the enzymes reverse transcriptase, integrase, and protease.[10] Thus, it is an important target for AIDS treatments and has lead to several FDA-approved drugs that specifically target its catalytic site. The active site of HIV-1p catalyzes the hydrolysis of the substrate peptides. It is gated by a pair of β-hairpin glycine-rich loops, one from each monomeric HIV-1p, which are commonly referred to as the "flaps" (K45-M-I-G-G-I-G-G-F-I54). The mobility of the flaps is essential to HIV-1p activity because they control the access and positioning of substrate in the catalytic site during hydrolysis. Several studies based on crystallography,[11,12] EPR,[13,14] NMR,[15] and MD simulations[16-18] suggest that the flaps of HIV-1p exist in an ensemble of conformational states and can adopt a range of conformations (closed, semi-open, and open).[19-22] Therefore, the dynamics of the flap can be exploited as a way to modulate the enzymatic activity of HIV-1p.

A recent study by Perryman *et al.*[2] identified potential allosteric sites near the flap region of HIV-1p. That work used fragment-based crystallography of HIV-1p in various conformations. Of particular interest is the presence of the molecular probe 5-nitroindole (5NI) in the hydrophobic pocket adjacent to the flap. Importantly, the 2.10-Å crystal showed that the 5NI-bound HIV-1p remained in semi-open conformational state. The hydrophobic pocket corresponds neatly to the aforementioned Eye site of the apo HIV-1p (communication with Dr. Stout). Conversely, crystals of HIV-1p complexed with

competitive inhibitors all maintain the closed conformation; the Eye site was not present and there was no bound 5NI. The molecular probe 5NI forms hydrophobic contacts with Val32, Ile47, Ile54, Pro81, and Ile84 and a hydrogen bond with the Gly51 amide through 5NI's nitro group. These residues have been suggested to play a role in flap-recognition.[16] This is the first crystallographic confirmation that demonstrates the existence of the Eye site in the semi-open HIV-1p, supporting the notion that the Eye site is a viable site for small molecule targeting. Furthermore, 5NI fits well within our Eye site MPS pharmacophore model (Figure 2.1 B) and overlaps with 2 of the 3 aromatic pharmacophore elements as well as the hydrogen-bond acceptor element. In this article, we demonstrate that a nitro-containing compound identified from our MPS pharmacophore screen, which has topological similarity to 5NI, can modulate the activity of HIV-1p. Additionally, we show the viability of MPS pharmacophore models as a tool for ligand-based screening.

(A)   **(B)**
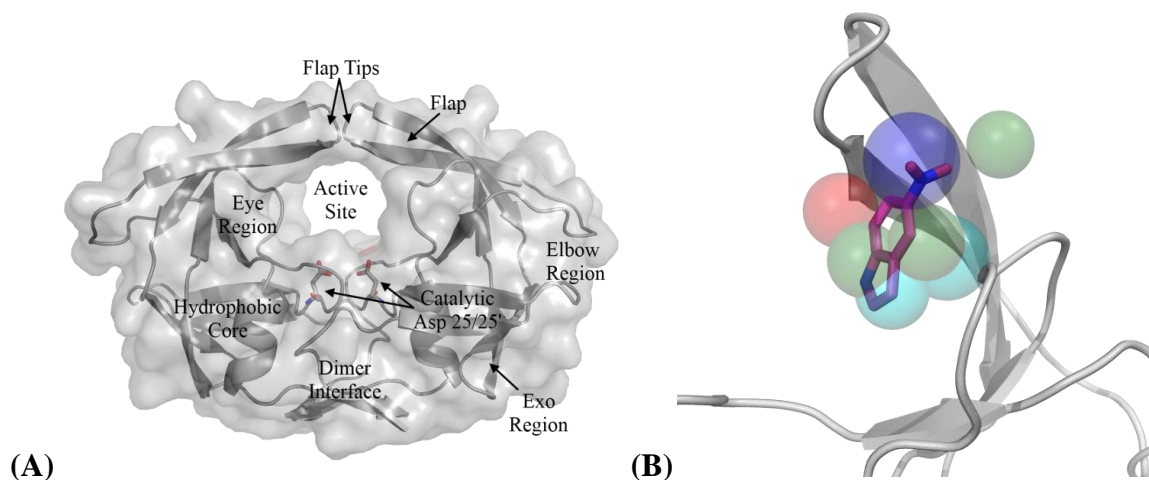
Figure 2.1. (A) A cartoon representation with protein surface of HIV-1p in the semi-open conformation (PDB: 1HHP). **(B)** 5-nitroindole (5NI, stick), soaked into apo HIV-1p in *P4₁* crystal lattice, matches part of the Eye MPS pharmacophore model. The pharmacophores are color-coded: hydrophobic (cyan), aromatic (green), hydrogen-bond donor (red) and hydrogen-bond acceptor (blue).

2.3 Methods

DiverseSolutions Chemical Similarity Search

DiverseSolutions[23-25] (**DVS**), available in SYBYL version 8.0,[26] was used to perform lead hopping based on the physical properties of the input structures. The DVS algorithm automated the construction of the DVS chemistry-space of the complete CCG library (v2007.10). Five 2D, physical descriptors were selected to construct the DVS chemistry-space: Gasteiger charges, number of hydrogen-bond acceptors, number of hydrogen-bond donors, polarizability, and the structure topology/connectivity. Our previous MPS pharmacophore model of the Eye site identified 78 CCG compounds and those were used as "seed." Compounds in bins within 1 unit radius, the most stringent setting, of the 76 "seeds" were selected. Compounds with molecular weight above 500 Da were removed and the remaining compounds were clustered by the MACCS fingerprint function of MOE.[27] A total of 77 compounds were selected for testing as new lead-hopped compounds for further study.

Markush Chemical-Similarity Search

UNITY,[28,29] a module of SYBYL suite (version 8.0), was used in the ligand-based chemical search. Four Markush searches were constructed based on the chemical structure and connectivity of the reference structure, Compound **1**. UNITY performed the query searches against four academic and commercial chemical libraries: ChemBridge (2007), ChemDiv (2007), MayBridge (2007), and CCG (2007.10) chemical libraries. Compounds that matched the queries were selected, and the 3D conformers were generated with OMEGA 2.3.2,[30] a module of the OpenEye suite. The selected compounds

were scored and ranked by ROCS (version 3.0.0) and EON (version 2.0.1), respectively, both modules of the OpenEye suite.[31] ROCS scores a chemical according to its shape similarity to the reference structure, while EON scores the charge distribution similarity of a chemical to the reference structure. A consensus score of the ROCS and EON scores with equal weight were used to rank the selected compounds. The top-200 compounds from each library were examined manually, and 43 compounds were selected for testing experimentally (CCG: 4 compounds; ChemBridge: 27 compounds; ChemDiv: 8 compounds; MayBridge 4 compounds).

Inhibitor Screening Assay

Pseudo wild-type (WT; G7K) and multi-drug resistant (MDR; L10I/L63P/A71V/ G73S/I84V/L90M) HIV-1p were kindly provided by Dr. Celia Schiffer of the University of Massachusetts. Pseudo WT protease was used to avoid autoproteolysis activity of HIV-1p. A resonance energy transfer (FRET)-based biochemical assay was used to assess the HIV-1p enzymatic activity. A fluorogenic peptide substrate was used, RE(EDANS)SQNYPIVQK(Dabcyl)R (Molecular Probes, Cat. No. H-2930) was used. This substrate contains a fluorophore EDANS and a chromophore Dabcyl, which quenches the excited fluorophore when the two chromophores are in close proximity.[32] EDANS has an excitation wavelength near 340 nm and emission wavelength near 490 nm, while Dabcyl has an excitation wavelength overlapping the emission range of EDANS. The peptide substrate turnover by the protease was monitored by the spectrometer SpectraMax M5, from the Molecular Devices. Top-read mode was used in the fluorescence detection, where excitation/emission wavelengths of EDANS at 340 and

490 nm, respectively, were monitored. A 475-nm cutoff filter was applied to reduce the noise signal. Screening assays were performed in triplicate in black, round-bottom, low-volume, 384-well plates (Corning No. 3676).

All purchased compounds were screened at 150 μM (WT HIV-1p at 30nM and substrate at 2μM). Stock solution of the tested compound dissolved in DMSO was diluted with milli-Q H$_2$O to ten-fold the desired final concentration. Final concentration of DMSO in the assay was kept below 2% v/v.[33] Milli-Q H2O and pepstatin A were used as the negative and positive controls, respectively. The ligand solution was mixed with the protease in the standard assay buffer (100 mM sodium acetate, 1 M NaCl, 1 mM EDTA, 1 mM DTT, 20% v/v Glycerol, 0.1% w/w CHAPS, 0.2% v/v PEG-400, pH = 4.7)[1,34-38] and incubated for 30 mins at room temperature. The enzymatic assay was initiated by introduction of the fluorogenic peptide substrate (diluted in assay buffer) and shaken for 15 secs inside the plate reader; the assay was monitored for 10 mins at 30°C. The kinetic data was fitted linearly to determine the rate of fluorogenic substrate turnover, measured as change in fluorescence intensity per unit time. Inhibitory activity was calculated by comparing the turnover rate against the negative control,

$$\% \text{ Remaining Protease Activity} = \frac{d(RFU)_{ligand}}{dt} / \frac{d(RFU)_{negative}}{dt} \times 100\%$$

where RFU is the raw fluorescence unit measured by the spectrometer.

Michaelis-Menten Kinetics

The Michaelis-Menten kinetics of the fluorogenic peptide substrate with WT or MDR HIV-1p was performed by varying the fluorogenic substrate concentration between 2.5 μM and 100 μM against a constant concentration of HIV-1p at 30 nM. The experimental conditions were identical to those mentioned above. As this experiment

requires the use of high concentrations of fluorogenic peptide substrate, which contains the chromophore Dabcyl, reabsorption of the fluorophore-emitted light by the chromophores in the solution will be significant. This effect, called the inner filter effect, will disproportionally affect and reduce the intensity of observed fluorescence at high substrate concentration (> 20 μM) while having little influence on the intensity of fluorescence at low substrate concentration. To correct this effect, a standard curve of EDANS fluorescence against an increasing concentration of the fluorogenic peptide substrate was obtained and used to correct the intensity of observed fluorescence. The data-analysis package SigmaPlot 11.0 was used to calculate the Michaelis constant (K_m) by fitting the data to an one-site saturation model with a non-linear regression method.

Dose-Dependent Inhibition Assay

For IC_{50} determination, the final concentrations of protease and fluorogenic peptide substrate were 30 nM and 5 μM, respectively. K_m of the substrate was determined to be 91 ± 11 μM and 207 ± 26 μM for WT and MDR HIV-1p, respectively, in our buffer and assay condition. IC_{50} and Hill Slope were obtained by fitting the kinetic data to a sigmoidal dose-response model using SigmaPlot. Under the assumption that the dose-response kinetics is appropriate to describe allosteric inhibition, the inhibition constant (K_i) of the tested compound was obtained through the Cheng-Pursoff equation

$$K_i = \frac{IC_{50}}{\left(1 + \frac{[S]}{K_m}\right)}$$

Dimerization Inhibition Analysis

In order to rule out the mechanism where the compounds inhibit dimerization, Zhang-Poorman kinetics analysis was performed.[37,39-41] Protease concentration was varied between 0.5 and 30 nM while substrate concentration was fixed at 10 μM. The kinetic data was plotted as $\sqrt{v_i}$ vs. $[E]_0/\sqrt{v_i}$, where v_i is the rate of substrate turnover and $[E]_0$ is the concentration of the protease. The kinetic data was linearly fitted and the mode of inhibition was determined by comparing the slope of the curves, where parallel lines indicate the compound modulates protease activity through the dimerization site while intercepting lines indicate the mode of inhibition does not involve the dimerization site.

$$\frac{[E]_0}{\sqrt{v_i}} = \frac{K_m\left(1 + \frac{[I]}{K_i^{com}}\right) + [S]}{[S]k_{cat}}\sqrt{v_i} + \sqrt{\frac{K_d K_m}{4[S]k_{cat}}\left(1 + \frac{[I]}{K_i^{dim}}\right)}$$

Cross-Competitive Inhibition

A variation of Yonetani-Theorell kinetic analysis was used to examine the mode of inhibition of our compounds.[42,43] Pepstatin A, a known competitive inhibitor for HIV-1p, was used in this kinetic assay. Pepstatin A concentration varies from 0 to 300 nM in assay with WT HIV-1p and 0 to 400 nM with MDR HIV-1p.[44] The concentrations of substrate and protease were kept at 5 μM and 30 nM, respectively, while other experimental conditions were same as above. The kinetic data was plotted as v_0/v_i versus $[pepstatin]/IC_{50pepstatinA}$. The kinetic data was fitted linearly to determine the interaction factor that defines the type of interaction (agonistic, antagonistic, or mutually exclusive) between the known competitive inhibitor and the tested compound with its unknown mode of inhibition.

Molecular Docking

FRED[45] (version 2.2.5), a module of the OpenEye suite, was used to dock NIT into the semi-open HIV protease (PDB: 1HHP) or the closed form,[46] with the two scoring functions, ShapeGauss[47] and ChemGauss3, were chosen for exhaustive searching and optimization, respectively. Ligand and protein were setup according to the default settings.

Dynamics Simulations

Unrestrained all-atom molecular dynamic (MD) and Langevin dynamic (LD) simulations were performed with the FF99SB force field[48] and the AMBER10 suite of programs.[49] An apo HIV-1p in semi-open conformation (PDB: 1HHP[50]) was obtained from the PDB.[51] C_2-symmetric homodimer of HIV-1p was generated with PyMOL version 1.2.[52] The ligand NIT was placed into the Eye site of HIV-1p by docking with Glide version 5.5[53] from Schrödinger, using Standard Precision and default settings. R.E.D.[54] version 3.4 and Ante_R.E.D. version 2.0 (http://q4md-forcefieldtools.org/RED/) were used to derive the RESP charge values for the ligand NIT. R.E.D. calculated the ESP charges at HF/6-31G* level with Gaussian09, which were then fitted to the ligand through a two-stage RESP fitting. Force field parameters of NIT were built with GAFF force field[55] and the RESP charges using the Antechamber module of AMBER.[56]

Five independent, 20-ns MD simulations were performed using different random-number seeds. The NIT-protease complex was solvated with a truncated octahedral TIP3P water box[57] with a buffer distance of 12 Å and closeness parameter of 0.5. The system charge was neutralized with Cl⁻ counter-ions. A 10-Å cut-off for van der Waals

interactions was used and particle mesh Ewald for long-range electrostatics[58] was employed. The simulations were run in the NPT ensemble and SHAKE algorithm[59] was used to constrain all bonds to hydrogen atom to allow a 2 fs time step. To avoid flap collision, we applied the following equilibration protocol.[60] For the solvated system, hydrogen atoms were first minimized, followed by the side chains and then all atoms. The system was then equilibrated first with a gradual heating of water from 10 to 310 K over 50 ps and then a water equilibration with protein atoms restrained for 250 ps at 310 K. This was followed by a full system heating from 10 to 310 K over 180 ps and a full system equilibration with protein unrestrained at 310 K for 400 ps. The production phase was run for 20 ns at 310 K.

Five independent, 20-ns LD simulations were also performed. Hydrogen atoms were constrained with SHAKE while a 999 Å cut-off distance for non-bonded interactions was used. Generalized Born approach was used to implicitly model aqueous solvation for the LD simulations.[61] Default dielectric values, where interior = 1 and exterior = 78.5, were used. The time step and the collision frequency of the simulation were 1 fs and 1 ps^{-1}, respectively. Simulations began with hydrogen minimization, followed by side-chain and then all-atom minimizations. Equilibration was done in six stages: the system was gradually heated from 100 K to the final temperature at 300 K in the first two equilibration steps. Restraints were placed on all heavy atoms and gradually removed over the first four equilibration steps using force constants from 2.0 to 0.1 kcal/mol·Å^2, where the first three steps were done over 10 ps and the fourth step over 50 ps. Only the backbone atoms were restrained at 0.1 kcal/mol·Å^2 over 50 ps in the fifth equilibration step. In the sixth equilibration step, the restraints were removed, and all

atoms were allowed to move for 300 ps at 300 K. The production phase was run for 20 ns. Analysis of the trajectories was performed using the PTRAJ module of AMBER suite.

Essential Dynamics Analysis

Essential Dynamics analysis of the MD trajectory data was applied to compare dynamics of the protein structure in simulations.[62,63] PTRAJ was used to perform the orthogonal transformation calculation on the covariance matrix of the backbone heavy atoms and solve for the eigenvectors and the associated eigenvalues of the MD trajectories.[64] The eigenvectors were compared by calculating the dot product of the corresponding vectors in each set of eigenvector. The dot-product values were then rescaled to between 0 and 100, where 0 corresponds to a dot-product of 1.0, i.e. strongly correlated, and 100 corresponds to a dot-product of -1.0, i.e. strongly anti-correlated. The rescaled dot-product values were added to a reference PDB structure for visualization in VMD 1.8.9.[65]

<div align="center">

2.4 Screening for Eye Compounds

</div>

We have a particular interest in the flap-tip recognition cleft, also known as the "Eye" site,[1] consisting of residues Val32, Ile47, Gly48, Gly49, Ile50, Ile54, Val56, Gly78, Pro79, Thr80, Pro81, and Ile84 (Figure 2.1 A). Upon substrate binding, each flap closes down and positions its flap tip (residues 49-52) into the highly conserved region on the opposite-side monomer. We proposed that by specifically targeting this site the

catalytic efficiency of the protease can be modulated through altering the equilibrium of the flap conformational states.[1] Using the multiple protein structures (MPS) method[66-69] and an ensemble of HIV-1p in semi-open conformations generated from a MD simulation, we previously constructed a novel pharmacophore model of the Eye site of HIV-1p that incorporates the flexibility of the flap (Figure 2.2 A).

(A)　　　　　　　　　　　　　　　　　　(B)
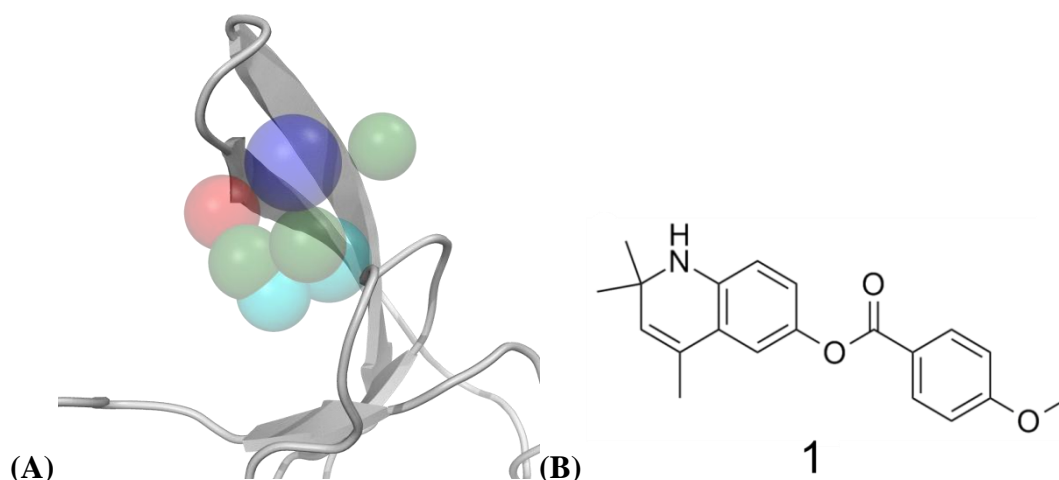
Figure 2.2. (A) Pharmacophore model of the HIV-1 protease allosteric site, the Eye site. Elements of the pharmacophore model are color coded according to chemical property: red for hydrogen-bond donor, blue for hydrogen-bond acceptor, green for aromatic and cyan for hydrophobic. **(B)** Structure of Eye-pharmacophore derived compound **1** with inhibitory activity against HIV-1p.

Testing Compounds Matching the Eye Pharmacophore

The Eye site pharmacophore model has been used to virtually screen against the Center of Chemical Genomics (CCG) library from the University of Michigan. The best inhibitor, 2,2,4-trimethyl-1,2-dihydroquinolin-6-yl 4-methoxybenzoate (compound **1**), had dose-dependent activity against HIV-1p (Figure 2.2 B). This compound was the first experimentally tested, active inhibitor that was designed to target the newly discovered allosteric site, rather than the traditional catalytic site.[1]

In this work, we started by pursuing additional CCG compounds in the original hit list that fulfill the pharmacophore requirements (Appendix 1). Of the 93 CCG compounds in the list, 56 were purchasable from ChemBridge, ChemDiv, and Enamine. All compounds were tested at 150 µM against the wild-type HIV-1p in the standard inhibition screening assay; however, none of the compounds in this list showed significant inhibitory activity. All wells retained > 80% catalytic activity when compared to the negative control (Table 2.1).

Expand Chemical Scaffold Diversity

To expand the number of chemical scaffolds with physical characteristics similar to the compounds that match the Eye-site pharmacophore model, we applied a ligand-based, drug design strategy and performed a property-based lead hopping. This property-based, lead hopping identifies a diverse subset of compounds which allows a better representation of the full range of chemical diversity from a much larger population of compounds. Thereby, it helps to identify diversity-voids or missing diversity that the pharmacophore model did not pick up, and it avoids the time and expense of synthesizing or screening 'redundant' compounds. Lead hopping was performed with DiverseSolutions (**DVS**). DVS algorithm constructs a chemistry-space by using orthogonal metrics as the axes that best represent the physical characteristics of the compounds. This chemistry-space is partitioned into bins, or "cells"; all compounds with similar properties, i.e. similar position on the metrics, will be placed within the same bin, and thus, each bin is considered a subset of diverse compounds. The metrics are selected

and scaled for optimal distribution of the bins within this space automatically by the DVS

algorithm (Figure 2.3 A).

Table 2.1. Anti-protease activity of compounds matching Damm Eye pharmacophore model

CCG #	Vender ID	Activity (%)	MW	CCG #	Vender ID	Activity (%)	MW
11749	CB 5562086	96 ± 6	230	32726	CD C881-1035	94 ± 14	336
15562	T0500-0102	95 ± 13	264	25407	CD 6658-0012	81 ± 8	337
15534	CD 1465-0022	97 ± 9	286	2755	CB 5626286	--	339
9795	CB 5524366	68 ± 4	288	31042	CD C700-1363	88 ± 8	339
15497	CB 5210344	95 ± 5	296	17743	CD 3448-7570	94 ± 13	342
11711	CB 5561288	--	297	23868	CD 5867-3931	--	342
33887	CD K402-0900	83 ± 4	297	10728	CB 5542560	--	343
27314	CD C200-1280	91 ± 6	298	22387	CD 5228-4182	86 ± 14	343
17350	CD 3343-2780	96 ± 6	299	4820	CB 6504928	93 ± 8	345
25664	CB 7828364	99 ± 3	304	30807	CD C677-0232	90 ± 7	345
17498	CD 3379-0757	97 ± 10	306	24648	CD 6143-0142	87 ± 6	346
9502	CB 5491484	83 ± 8	308	34929	CD 6737-0622	101 ± 7	346
9796	CB 5524374	--	314	30113	CD C610-0367	86 ± 7	347
15509	CD 1323-0109	86 ± 6	315	30878	CD C684-0061	84 ± 8	349
21913	CD 5056-0111	109 ± 12	316	5666	CB 5307415	82 ± 31	351
10077	CB 5529870	98 ± 20	316	23977	CD 5926-0037	79 ± 10	351
10587	CB 5539991	--	318	17756	CD 3448-9010	102 ± 6	351
33840	CD K402-0245	89 ± 12	323	9375	CB 5486930	86 ± 6	352
9555	CB 5493267	73 ± 12	323	8272	CB 5466385	105 ± 17	353
33642	CD K280-0545	100 ± 7	324	31038	CD C700-1355	89 ± 8	353
25197	CD 6456-0643	85 ± 7	326	31041	CD C700-1362	89 ± 7	353
34745	CB 5224354	99 ± 9	328	35026	CD C700-1271	92 ± 8	353
22386	CD 5228-4175	85 ± 7	331	31707	CD C742-0059	84 ± 8	353
32745	CD C884-0070	89 ± 10	333	11538	CB 5557939	87 ± 7	354
11244	CB 5552988	88 ± 7	333	31043	CD C700-1364	91 ± 7	357
9924	CB 5526815	87 ± 7	334	12617	CB 5600269	83 ± 8	358
25421	CD 6703-1500	89 ± 7	334	17727	CD 3448-5797	94 ± 8	358
4854	CB 6585837	88 ± 6	335	21532	CD 4896-4230	108 ± 7	359

CB: ChemBridge; CD: ChemDiv; T: Enamine
All compounds were tested at 150 μM
Compounds suffered solubility or auto-fluorescence issue are unmarked

(A)
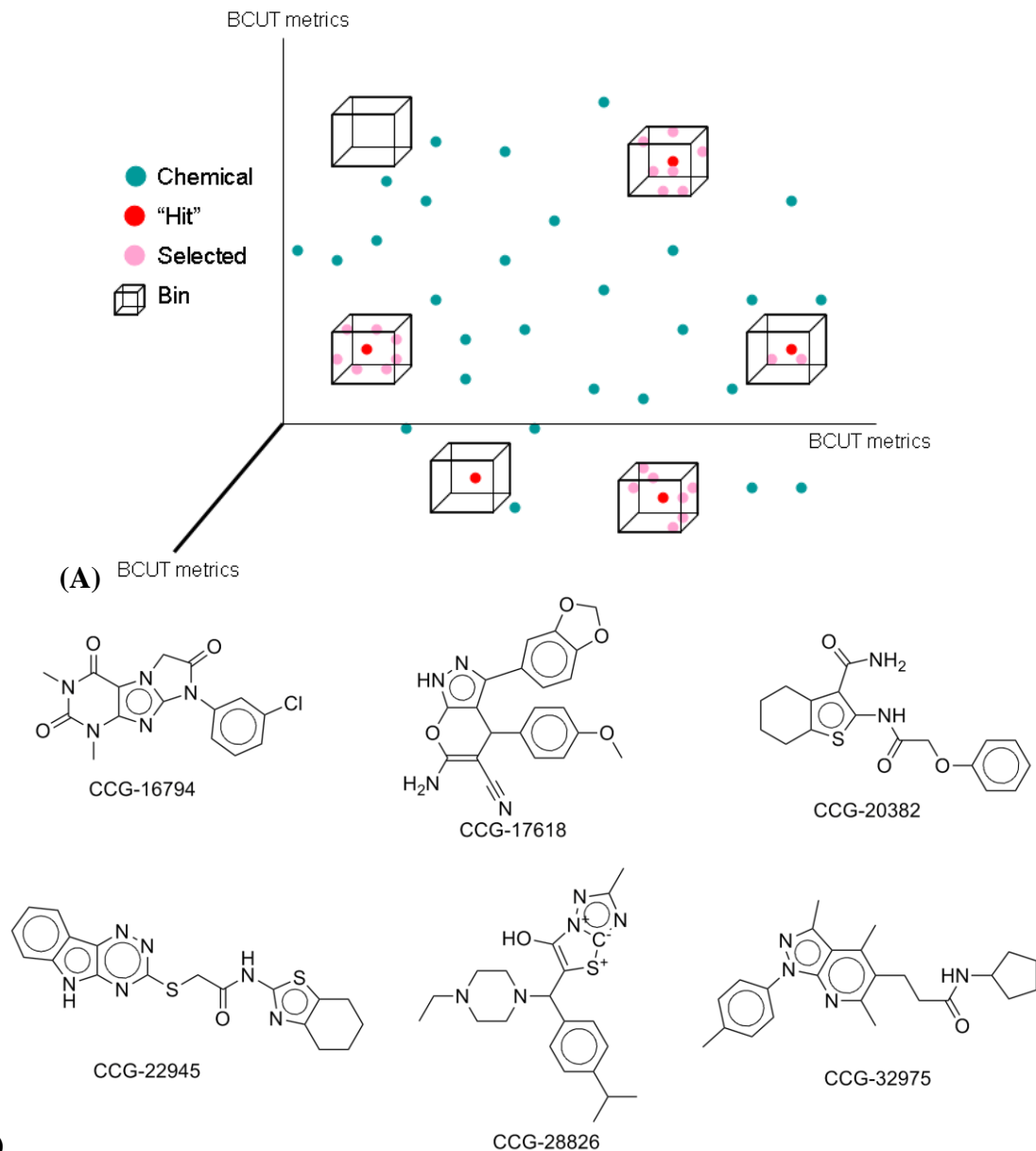
(B)

Figure 2.3. (A) Schematic representation of DiverseSolutions (DVS) chemistry space and binning of physically similar compounds. The axes are represented by a scaled quantitative physical property measurement. Each compound is placed in this chemistry space based on the quantified physical properties. Compounds are pooled into a single bin if they share similar physical properties. "Seeds" are placed into this chemistry space and bins within a certain "radius" of the "seed" will be considered to have similar physical properties as the "seed". **(B)** Compounds selected based on DVS lead hopping that match the physical characteristics of compounds selected by Damm Eye pharmacophore model.

For the CCG library (version 2007.10; ~ 90,000 compounds), five 2D, physical descriptors were selected as metrics to construct the DVS chemistry-space: Gasteiger charges, number of hydrogen-bond acceptors, number of hydrogen-bond donors, polarizability, and structure topology/connectivity. The previously identified CCG compounds were used as the "seeds" in this DVS chemistry-space.

To perform lead hopping, bins within a given distance of the bin with "seed" chemical are collected. In this study, 406 CCG compounds were selected from the bins within 1 unit radius, the most stringent setting. Compounds identified by the DVS method are structurally different from the "seeds" but have similar physical characteristics (Figure 2.3 B). Of these, 389 compounds have molecular weight below 500 Da. After removing problematic compounds and clustering the compounds with MACCS fingerprint function of MOE, 77 compounds were ordered from CCG library for testing (Table 2.2). However, none of the tested compounds showed significant inhibitory activity against HIV-1p, and thus, we focused on the one chemical scaffold that has strong inhibitory activity against the protease, compound **1**.

Structure-Activity Relationship of Compound 1

Compound **1** showed significant inhibitory activity against HIV-1p in the FRET-based biochemical assay. We sought to expand on the chemical space around compound **1** and establish a simple structure-activity relationship (SAR). The search of analogs of compound **1** that contain the core 2,2,4-trimethyl-1,2-dihydroquinolin-6-yl moiety was achieved through the ZINC 7 database.[70] ZINC has a comprehensive chemical database containing commercially available compounds from various chemical vendors. In

Table 2.2. Anti-protease activity of DVS-searched compounds					
CCG ID	Activity (%)	MW	CCG ID	Activity (%)	MW
11749*	96 ± 6	230	10042	56 ± 6	376
28407	45 ± 7	279	28832	74 ± 17	377
17350	67 ± 3	299	32976	--	377
50440	67 ± 7	302	1715	--	377
10587	--	318	14308	--	382
11429	--	318	25865	--	383
6394	72 ± 6	319	20465	--	387
1518	--	324	17617	64 ± 8	388
49421	71 ± 13	326	17618	82 ± 12	388
19572	103 ± 7	328	14780	--	388
20382	84 ± 24	330	28802	64 ± 5	389
34984	71 ± 9	332	16226	113 ± 5	390
14584	95 ± 18	333	1757	--	390
4854	--	335	32975	--	391
2755*	--	339	16935	--	391
18494	103 ± 8	340	4514	73 ± 8	392
25811	104 ± 7	342	28834	70 ± 7	393
12059	105 ± 17	343	28841	85 ± 9	393
11928	--	344	29760	71 ± 6	393
56317	103 ± 8	345	1775	--	394
34066	86 ± 10	345	13117	106 ± 13	394
16794	--	346	11653	--	395
44696	74 ± 8	346	14637	--	397
9678	--	347	22945	83 ± 9	397
14656	92 ± 17	350	1714	--	397
20512	93 ± 7	352	13693	91 ± 13	398
17596	89 ± 6	355	31704	--	398
30784	75 ± 11	356	20783	95 ± 9	399
17627	78 ± 8	359	14407	77 ± 14	400
17628	89 ± 7	359	28826	94 ± 6	401
25812	--	360	9349	--	407
15685	--	360	30782	71 ± 4	410
17629	76 ± 7	362	19810	75 ± 8	413
14712	90 ± 17	363	19881	71 ± 3	418
14738	74 ± 10	363	28936	80 ± 6	431
29752	75 ± 8	366	15962	--	460
12013	--	368	4023	--	472
28800	92 ± 8	373	29776	69 ± 7	491
19584	129 ± 9	374			

* Compounds from KLD original 93 Eye compound list
  Compounds tested at 150 µM

addition, the database has an integrated substructure search engine to identify compounds with similar connectivity to the query compounds. A ZINC substructure search with a 90% similarity cutoff was used to identify analogs of compound **1**. In addition, we searched the CCG library for analogs with MOE v2007 using the MACCS fingerprint at

85% similarity. A total of 43 compounds with connectivity similar to compound **1** were identified and they were purchased for testing (Table 2.3). However, the tested analogs were all weak inhibitors of proteolytic activity of HIV-1p, and this suggested that the available compounds with 1,2-dihydroquinolin-6-yl moiety, with the exception of compound **1**, were likely ineffective at targeting the Eye site of HIV-1p.

2.5 Identifying New Scaffolds through Ligand-Based Screening Methods

Identification of Compounds through Pharmacophore Screening

Although compounds with the core scaffold of compound **1** (2,2,4-trimethyl-1,2-dihydroquinolin-6-yl moiety) did not appear to be particularly effective at inhibiting HIV-1p, compound **1** itself fits the Eye site pharmacophore model well. Hence, the chemistry space around this particular scaffold could be further explored and expanded to identify new scaffolds that would also fit the Eye site pharmacophore model and target the allosteric site of HIV-1p.

We again applied a ligand-based drug discovery strategy; instead of property-based lead hopping method like DVS, we used a structure-based, lead-hopping method that applies the key features of an input structure. The chemical features of compound **1** were used to generate a series of Markush chemical-similarity searches with the following general features: (1) sp^2 N-H connecting to an aromatic ring (analog to quinolin-6-yl moiety) that can be either a 5- or 6-member ring, (2) aromatic ring of any decoration (analog to *p*-methoxy benzoate moiety), and (3) a 3-atom linker of connecting

91

Table 2.3. Anti-protease activity of analogs of compound 1 against HIV-1p

ID	Structure	Activity (%)	ID	Structure	Activity (%)
IB 1S-07881 (MW 283)		94 ± 3	IB 1S-44276 (MW 329)		80 ± 5
IB 3S-52852 (MW 285)		83 ± 5	CB 5803990 (MW 337)		88 ± 7
CCG 8324 (MW 287)		--	CB 6598249 (MW 337)		--
CB 5493403 (MW 293)		100 ± 4	CB 5406001 (MW 338)		--
CB 6593620 (MW 294)		86 ± 5	CB 5885306 (MW 338)		--
IB 2S-18800 (MW 294)		--	IB 1S-11006 (MW 338)		73 ± 8
IB 2S-15238 (MW 295)		101 ± 7	CB 6708464 (MW 340)		79 ± 5
CB 6347154 (MW 299)		81 ± 4	CB 5480770 (MW 343)		104 ± 2
CB 6997621 (MW 301)		86 ± 6	IB 1S-05888 (MW 343)		61 ± 5
IB 1S-02470 (MW 307)		--	IB 1S-40780 (MW 343)		--
IB 1S-03621 (MW 307)		82 ± 4	CB 6616468 (MW 353)		91 ± 2
IB 1S-04132 (MW 307)		--	IB 1N-03786 (MW 353)		96 ± 9
CB 6953658 (MW 313)		--	CB 6596434 (MW 354)		75 ± 5
CB 5310024 (MW 323)		96 ± 5	IB 1S-07849 (MW 358)		--
CB 5373175 (MW 323)		--	IB 1S-15598 (MW 358)		--

CB = ChemBridge; CCG = Center of Chemical Genomics; IB = IB Screen; T = Enamine
All compounds were tested at 150 μM
Compounds suffered solubility or autofluorescence issue are marked as n/a

Table 2.3 (cont'd). Anti-protease activity of analogs of compound 1 against HIV-1p

ID	Structure	Activity (%)	ID	Structure	Activity (%)
CCG 5524 (MW 323)		29 ± 10	CCG 8482 (MW 358)		--
IB 2S-16330 (MW 323)		99 ± 3	T0502-1854 (MW 364)		104 ± 3
CB 7735913 (MW 325)		83 ± 9	IB 1S-16456 (MW 372)		92 ± 3
IB 1S-12036 (MW 328)		97 ± 5	CCG 3414 (MW 374)		--
IB 1S-13537 (MW 328)		90 ± 4	CB 5470681 (MW 387)		103 ± 4
IB 1S-19089 (MW 328)		93 ± 8	CCG 3426 (MW 429)		57 ± 6
CB 5660491 (MW 328)		--			

CB = ChemBridge; CCG = Center of Chemical Genomics; IB = IB Screen; T = Enamine
All compounds were tested at 150 μM
Compounds suffered solubility or auto-fluorescence issue are unmarked

(1) and (2) (analog to the ester linkage). Four two-dimensional Markush search queries were constructed based on the chemical features and connectivity of the reference structure, Compound **1** (Figure 2.4 A). UNITY, a module of SYBYL suite (version 8.0), was used in the ligand-based chemical search against the CCG library (v2007.10, ~ 90,000 compounds). Since the chemical diversity of CCG library was relatively limited, we expanded the chemical space by searching additional chemical libraries of three commercial vendors: ChemBridge (EXPRESS-Pick v2007, ~ 50,000 compounds), ChemDiv (2007, ~ 120,000 compounds), and MayBridge (v2007, ~ 56,000 compounds) chemical libraries (Table 2.4). As a result, 7,230 compounds marched the Markush search queries and a filtering method was needed to prioritize the matched compounds for testing.

(A)

(B)
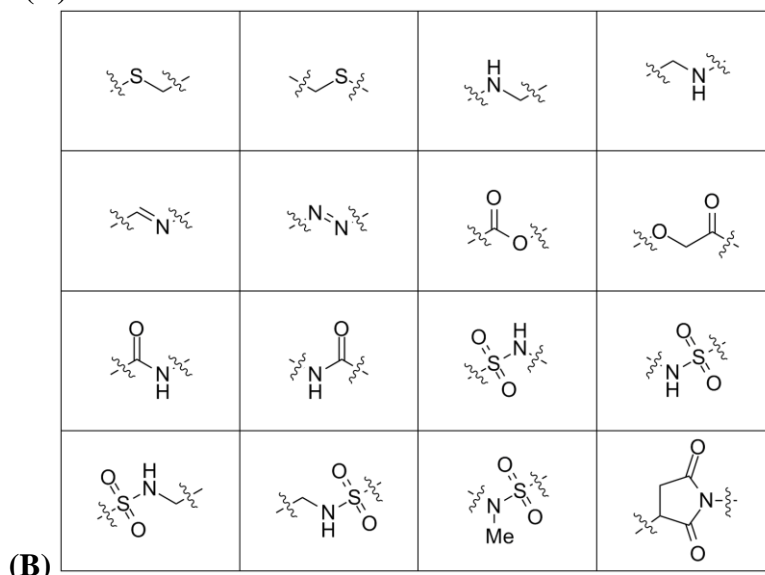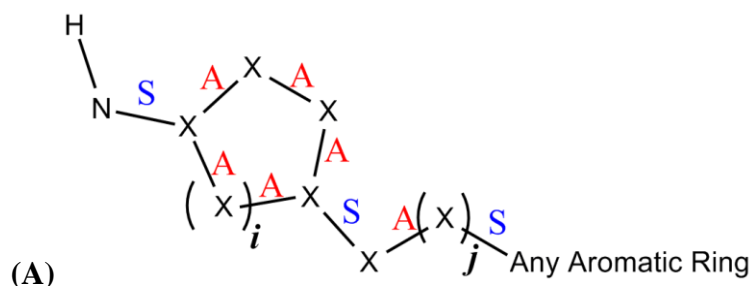
Figure 2.4. (A) Summary of Markush search queries based on the chemical features and connectivity of Compound **1**. The search queries have these designations: *A*, any bond order; *S*, single bond only; *x*, any heavy atom; *i* and *j*, number of heavy atom. **(B)** List of linker moieties used in the filtering of Markush searched results.

Table 2.4. Number of compound returned from Markush query searches

i	*j*	CCG	ChemBridge	ChemDiv	MayBridge
5	1	809	709	279	32
	2	1346	1439	281	95
6	1	392	291	51	16
	2	679	589	116	16

CCG: version 5-Oct-2007; ChemBridge: version 2007;
ChemDiv: version 2007; MayBridge: version 2007

To prioritize and rank the compounds that matched the Markush search queries, the selected compounds were docked, scored, and ranked by ROCS version 3.0.0 and EON version 2.0.1, respectively, both modules of the OpenEye suite. ROCS scores a

94

chemical according to its shape complementary to the reference structure, while EON

scores based on the similarity of charge distribution of a chemical to the reference

structure. A consensus score of the ROCS and EON scores with equal weight were used

to rank the selected compounds. The top-200 compounds from each library were

examined manually, and only those with fused double aromatic ring (similar to quinolin-

6-yl moiety) and unique linkers (Figure 2.4 B) were selected. Finally, 48 compounds (9

from CCG, 27 from ChemBridge, 8 from ChemDiv, and 4 from MayBridge) were

purchased for experimental testing (Table 2.5).

Table 2.5. Anti-protease activity of Markush compounds

ID	Activity (%)	MW	ID	Activity (%)	MW
CB 5202063	80 ± 5	367	CB 9058855	68 ± 6	328
CB 5228968	--	306	CB 9112729	65 ± 7	324
CB 5343019	--	329	CB 9118639	72 ± 9	333
CB 5373492	65 ± 2	286	CB 9123846	79 ± 12	338
CB 5482954	--	352	CB 9154080	--	325
CB 5522020	88 ± 8	280	CCG 18156 [a]	52 ± 9	318
CB 5566364	76 ± 5	296	CCG 21102 [a]	42 ± 4	340
CB 5570164	--	339	CCG 24371 [a]	65 ± 6	400
CB 5660337	79 ± 2	310	CCG 29660 [a]	65 ± 8	439
CB 5660491	--	328	CCG 30993 [a]	62 ± 6	390
CB 5738653	84 ± 5	244	CD D071-0025	80 ± 7	343
CB 5797113	77 ± 7	279	CD D298-0425	82 ± 5	315
CB 5979646	37 ± 6	357	CD E245-0514	92 ± 5	336
CB 7386606	79 ± 9	342	CD E966-0066	--	358
CB 7910527	88 ± 4	317	CD E977-0196	88 ± 5	334
CB 7929697	71 ± 7	369	CD G428-0017	81 ± 11	354
CB 7932774	75 ± 4	325	CD G645-0045	91 ± 14	284
CB 7935081	64 ± 4	261	CD G851-1179	93 ± 6	314
CB 7936269	79 ± 10	256	MB BTB 09244	95 ± 8	297
CB 7936917	58 ± 5	282	MB HTS 12769	92 ± 8	274
CB 7937150	82 ± 7	328	MB HTS 12771	84 ± 7	280
CB 7940247	89 ± 5	299	MB HTS 12793	89 ± 12	347
CB 7989211	74 ± 6	304	MB SEW 04939	--	406
CB 9022153	78 ± 10	331			
CB 9058373	--	240			

CB: ChemBridge; CD: ChemDiv; IB: IBScreen; MB: MayBridge
All compounds were tested at 150 µM
[a] Compound not purchasable from chemical vendors

Several compounds from the Markush set appeared to have some inhibitory activity against HIV-1p (CCG-18156, CCG-21102, CCG-24371, CCG-29660, CCG-30993, CB-5373492, CB-7935081, CB-7936917) with > 25% inhibition at 150 µM; hence, we also searched for compounds similar to these relatively active ones for testing. A total of 23 additional compounds with molecular weight < 500 Da and the required Markush features were found using the substructure search function of ChemSpider and ZINC online database. These compounds were purchased from various chemical vendors and tested (Table 2.6). A total of 66 purchasable compounds were hand-picked based on the consensus score result and the uniqueness of the scaffold. One compound, ChemBridge 5979646, was identified to have good inhibitory activity against HIV-1p and most importantly, it shared chemical features with compound 5NI, the fragment found to bind the Eye site in crystallographic studies.[2] This compound was specifically selected for more in depth study.

Table 2.6. Anti-protease activity of Markush-like compounds.

ID	Activity (%)	MW	ID	Activity (%)	MW
CB 5470681	80 ± 6	386	CD D043-0196	92 ± 7	311
CB 5977108	80 ± 4	312	CD G428-0524	95 ± 9	385
CB 7770331	82 ± 6	364	CD K906-2240	82 ± 4	375
CD C224-0681	122 ± 7	346	IB 3S-76897	--	345
CD C698-0346	98 ± 7	361	IB 3S-77617	104 ± 7	283
CD C698-0347	107 ± 13	361	IB 3S-80036	107 ± 6	299
CD C698-0348	100 ± 7	359	IB 3S-80366	--	361
CD C698-0378	96 ± 5	359	IB 3S-84057	95 ± 4	247
CD C698-0391	97 ± 24	375	IB 3S-90451	98 ± 5	259
CD C698-0447	79 ± 6	373	IB 5S-32389	--	350
CD C698-0540	93 ± 5	375	T0502-1854	81 ± 9	364
CD D043-0170	87 ± 6	318			

CB: ChemBridge; CD: ChemDiv; IB: IBScreen; MB: MayBridge; T: Enamine
All compounds were tested at 150 µM

2.6 Characterization of Active Eye Compound

Dose-Dependent Inhibition of NIT on HIV-1p

We confirmed that ChemBridge 5979646, 4-nitro-2-(2-thioxo-2,3-dihydrobenzothiazol-6-yl)isoindoline-1,3-dione (**NIT**, Figure 2.5 A), inhibited wild-type (WT) HIV-1p activity in a dose-dependent manner and has a $K_i = 96 \pm 3$ µM (Figure 2.5 B). This also translates into a binding affinity of approximately -5.5 kcal/mol and a respectable binding efficacy ~ 0.23 kcal/mol·heavy-atom. To eliminate the possibility of NIT as a promiscuous aggregator, we performed a dose-dependence assay at three HIV-1p concentrations (15, 30, 45 nM);[71] no significant change in Hill slope value was observed (Figure 2.5 B, insert).

We investigated the effectiveness of compound NIT against a multiple-drug resistant HIV-1p mutant (MDR; L10I/L63P/A71V/G73S/I84V/L90M). Using the competitive inhibitor pepstatin A (pepA) as the control, we showed that the MDR mutations cause only a 2.5-fold increase in K_i of pepA compared to WT HIV-1p ($K_i^{MDR} = 0.23 \pm 0.06$ µM and Hill Slope ~ 1.0; $K_i^{WT} = 0.094 \pm 0.012$ µM and Hill Slope ~ 1.0). Significantly, the MDR mutations do not appear to impact the K_i of compound NIT at all (Figure 2.5 B; $K_i^{MDR} = 91 \pm 6$ µM and Hill Slope ~ 1.5; $K_i^{WT} = 96 \pm 3$ µM and Hill Slope ~ 1.6). The MDR mutant appears to be as sensitive to NIT as the WT. Furthermore, NIT has a Hill Slope value of ~ 1.5 in the dose-dependent assay against both WT and MDR HIV-1p. This implies NIT may interact with more than one binding site on the protease, which is consistent with the presence of two Eye sites on each HIV-1p.
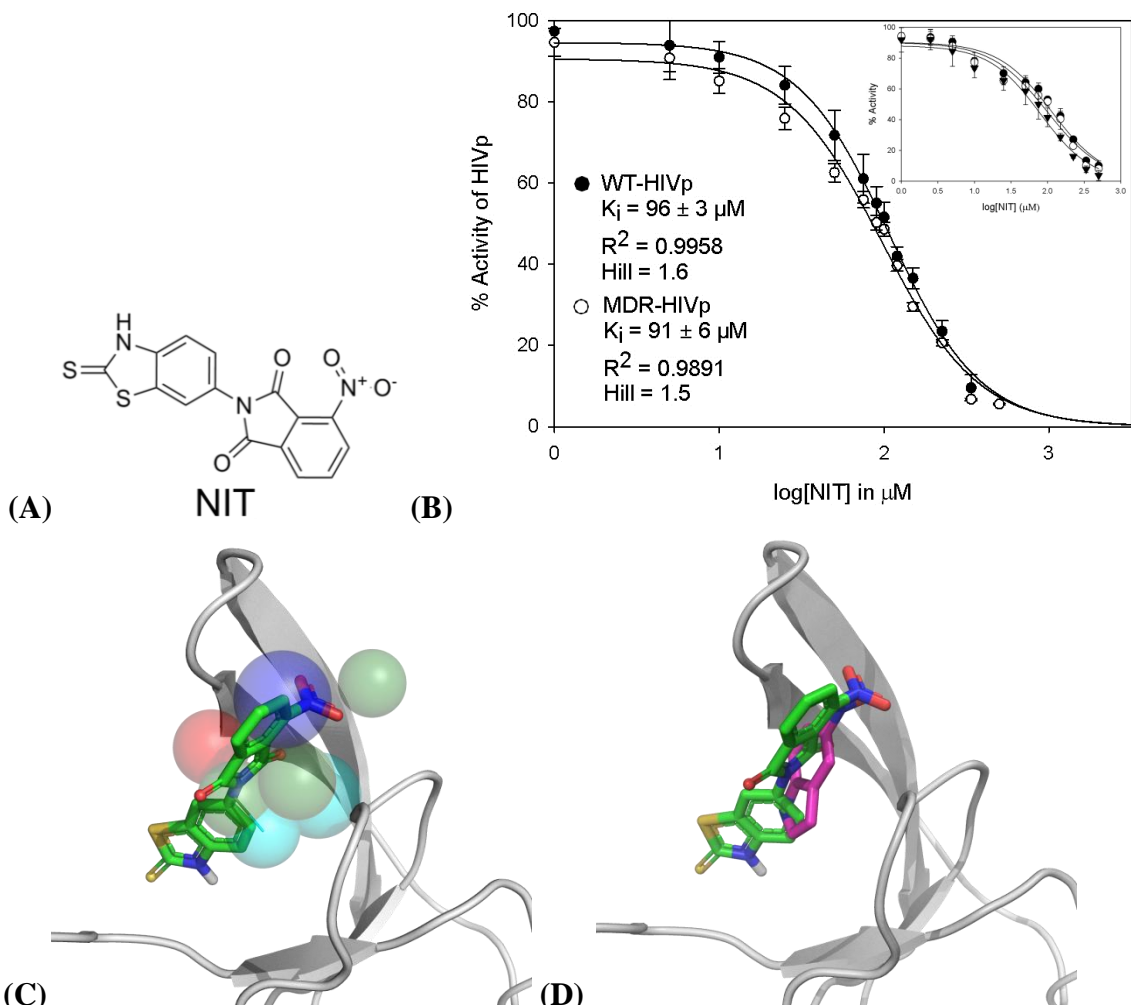
Figure 2.5. (A) Compound **NIT** is identified through Markush chemical-similarity search queries built from compound **1**. **(B)** Dose-dependence inhibition curves of compound **NIT**. It has similar inhibitory activity against the wild-type (● K_i = 96 ± 3 μM) and MDR HIV-1p (○ K_i = 92 ± 6 μM) at 30 nM protease concentration. The insert figure is the dose-dependence inhibition curves of NIT at different wild-type HIV-1p concentrations (▼ 15 nM; ○ 30 nM; ● 45 nM) and have very similar Hill slope values. **(C)** FRED docking pose of compound NIT (green) overlaying with the Eye site pharmacophore model.[1] The pharmacophores are color-coded according to chemical property: hydrophobic (cyan), aromatic (green), hydrogen-bond donor (red) and hydrogen-bond acceptor (blue). **(D)** FRED docking pose of NIT (green) overlaying with the molecular probe 5NI (purple) in the eye site.[2] Both 5NI and NIT have similar hydrophobic contacts with the Eye site.

Docking of NIT to the Eye site of semi-open HIV-1p with FRED shows that NIT overlays some of the Eye pharmacophores (Figure 2.5 C). The nitro group matches the hydrogen donor pharmacophore, while the isoindoline-1,3-dione (IID) moiety overlaps

with the aromatic pharmacophores. In addition, the 4-nitro-IID (4NIID) moiety of compound NIT is topologically similar to the molecular probe, 5NI, that was observed to occupy the Eye site of an apo HIV-1p crystal structure in semi-open conformation.[2] Docking of NIT to the Eye site with FRED generated poses with contacts to HIV-1p similar to 5NI (Figure 2.5 D). The 4NIID moiety forms hydrophobic contacts with Val32, Ile47, Ile54, Pro81, and Ile84 and a hydrogen bond with the amide of Ile50 through the nitro group, similar to the interactions seen in the 5NI-protease complex crystal structure. The benzothiazole-2-thione moiety of NIT forms hydrophobic contacts with Val32, Ile47, and Leu76 and hydrogen bonds with Asp30 and Gly48. These additional interactions may explain the slight difference in orientation of the 4NIID moiety of compound NIT, relative to the posing of 5NI in the crystal structure (Figure 2.5 D).

NIT Affects Michaelis-Menten Kinetics

To determine the effect of compound NIT on the kinetics of the HIV-1p, we determined the protease activity at various substrate concentrations (2.5 to 100 μM) using several concentrations of NIT. We obtained the K_m and V_{max} parameters of the HIV-1p through fitting the measured initial velocities to the Michaelis-Menten kinetics model using non-linear, least-square regression. We observed dose-dependent changes in the K_m and V_{max} of WT HIV-1p after the introduction of NIT. The result is consistent with mixed competitive inhibition kinetics, where the apparent K_m increases as NIT concentration increases, while V_{max} decreases concurrently (Table 2.7). The same trend is also observed in the MDR HIV-1p. This suggests NIT may be able to bind to both apo and substrate-

bound HIV-1p and may bind to an allosteric site as a mixed competitive inhibitor rather than binding to the catalytic site as a competitive inhibitor.

Table 2.7. Effects of compound NIT on WT and MDR HIV-1p Michaelis-Menten kinetics.

		Compound NIT (µM)		
		0	100	150
Wild-type	V_{max} (nM/s)	44 ± 3	35 ± 3	34 ± 3
	K_m (µM)	91 ± 8	144 ± 18	184 ± 26
MDR	V_{max} (nM/s)	42 ± 6	30 ± 2	30 ± 4
	K_m (µM)	171 ± 36	246 ± 21	359 ± 66

Ruling Out NIT as Dimerization Inhibitor

Although compound NIT may not act on the HIV-1p through the catalytic site, it is possible that NIT can modulate HIV-1p activity through an alternative site such as the dimer interface. To examine if NIT modulated the observed HIV-1p activity through disrupting the dimer interface, we determined the rate of protease-catalyzed hydrolysis (v_i) of the fluorogenic substrate at various protease concentrations for both WT and MDR strains of HIV-1p, ranging from 0.5 nM to 30 nM. The resulting initial velocities of the reaction were plotted as $[E]_0/\sqrt{v_i}$ vs. $\sqrt{v_i}$ according to the Zhang-Poorman equation. We observed lines with increasing slope as the ligand concentration increased (Figure 2.6). The same variation in slope was also observed when pepstatin A, a competitive inhibitor, was used as a non-dimer inhibitor control (Figure 2.6; insert). Hence, we determined that NIT does not act as a dimerization inhibitor to HIV-1p as it does not generate the typical parallel lines seen in classical dimer inhibitors in the Zhang-Poorman plot but has an increasing slope that is typically seen non-dimer inhibitors.[37]
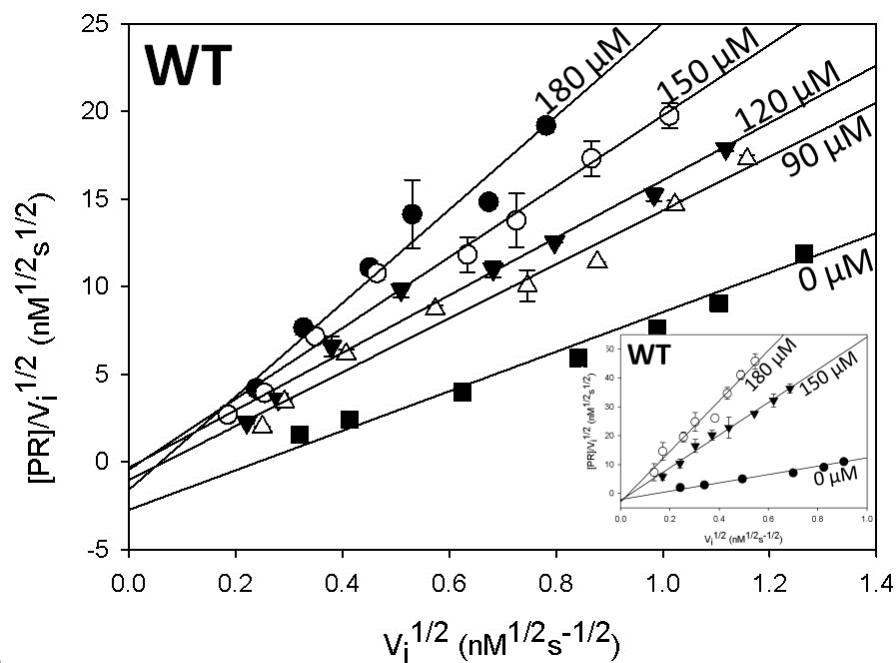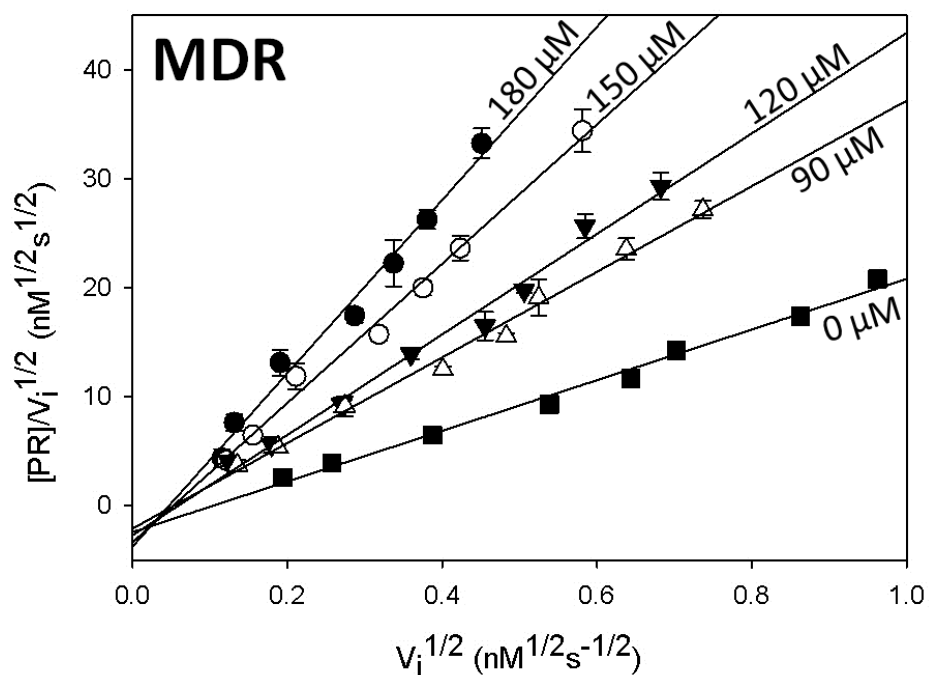
Figure 2.6. Zhang-Poorman analysis of compound NIT (0.5-30 nM HIV-1p; 5 μM substrate) with **(A)** Wild-Type and **(B)** MDR HIV-1p. Compound NIT concentrations: ■ 0 μM; Δ 90 μM; ▼ 120 μM; ○ 150 μM; ●, 180 μM. Non-parallel linear fits of compound NIT at various concentrations indicates the small molecule does not act as a dimerization inhibitor. The small insert in WT shows the Zhang-Poorman plot of pepstatin A, a non-dimerization inhibitor (● 0 nM; ▼ 150 nM; ○ 300nM).

Interaction of NIT and Competitive Inhibitor Pepstatin A

As compound NIT does not affect HIV-1p activity through the dimerization interface and appears as a mixed competitive inhibitor to HIV-1p in the Michaelis-Menten kinetics model, this indicates that NIT can bind to the protease while the substrate is bound to the protease catalytic site. To test how compound NIT interacts with the protease in relation to the substrate, we performed a cross-competitive assay of NIT with pepstatin A. We used the Yonetani-Theorell plot in the form of *Equation (1)* to evaluate the binding mode of the small molecule.[42,43,72]

By rearranging *Equation (1)* and plotting the data into *Equation (2)*, we obtained factor γ, which represents the degree of mutual influence of the two inhibitors on the binding of each other. In principal, the Yonetani-Theorell plot of v_0/v_i vs. *[I$_1$]* at various fixed *[I$_2$]* will give a series of parallel lines when the two ligands bind to the enzyme in a mutually exclusive manner ($\gamma = \infty$). If the two ligands bind to the enzyme simultaneously, the lines will intercept ($\infty > \gamma > 0$), and the type and strength of mutual interference (facilitation when $0 < \gamma < 1$; hindrance when $1 < \gamma < \infty$) can be assessed by the numerical value of γ. In the case when $\gamma = 1$, the two ligands bind in an independent manner.

$$\frac{V_o}{V_i} = 1 + \frac{[NIT]}{IC50_{NIT}} + \frac{[pepA]}{IC50_{pepA}} + \frac{1}{\gamma}\frac{[NIT][pepA]}{IC50_{NIT}IC50_{pepA}} \quad (1)$$

$$\frac{V_o}{V_i} = \left(1 + \frac{[NIT]}{IC50_{NIT}}\right) + \frac{[pepA]}{IC50_{pepA}}\left(1 + \frac{1}{\gamma}\frac{[NIT]}{IC50_{NIT}}\right) \quad (2)$$

Compound NIT at fixed concentrations was assayed against the competitive inhibitor pepstatin A at various concentrations and the kinetic data was visualized with the Yonetani-Theorell plot. Intercepting lines of fixed NIT concentration were observed, indicating that NIT does not bind mutually exclusively to the substrate hydrolysis site where pepstatin A binds competitively. NIT yields a γ value that is approximately 1.0 ± 0.2 with both WT and MDR HIV-1p (Figure 2.7), indicating that it likely binds independently to a separate site and does not interfere with the binding of competitive inhibitors to the substrate hydrolysis site.

In summary, NIT exhibits mixed inhibition kinetics in the Michaelis-Menten kinetics model, binds independently to a separate site, does not compete with pepstatin A for the substrate hydrolysis site, and does not modulate HIV-1p activity through the dimerization interface. These kinetics data support that NIT binds to an allosteric site and modulates HIV-1p catalytic activity through a non-competitive binding mode. Additionally, NIT is topologically similar to compound **1** that fits the MPS pharmacophore,[1] while it is structurally similar to 5NI that overlays to part of the MPS pharmacophore model and is seen crystallographically bound to the Eye site.[2] These data strongly suggest that NIT targets the Eye site of HIV-1p and modulates the protein in a novel allosteric manner.

Dynamics Studies of NIT-Protease Complex

To further examine the interactions between compound NIT and the Eye site of HIV-1p, five independent runs of 20-ns MD simulations were performed (100 ns total). Compound NIT was initially docked into one of the two Eye sites of the homodimeric

103
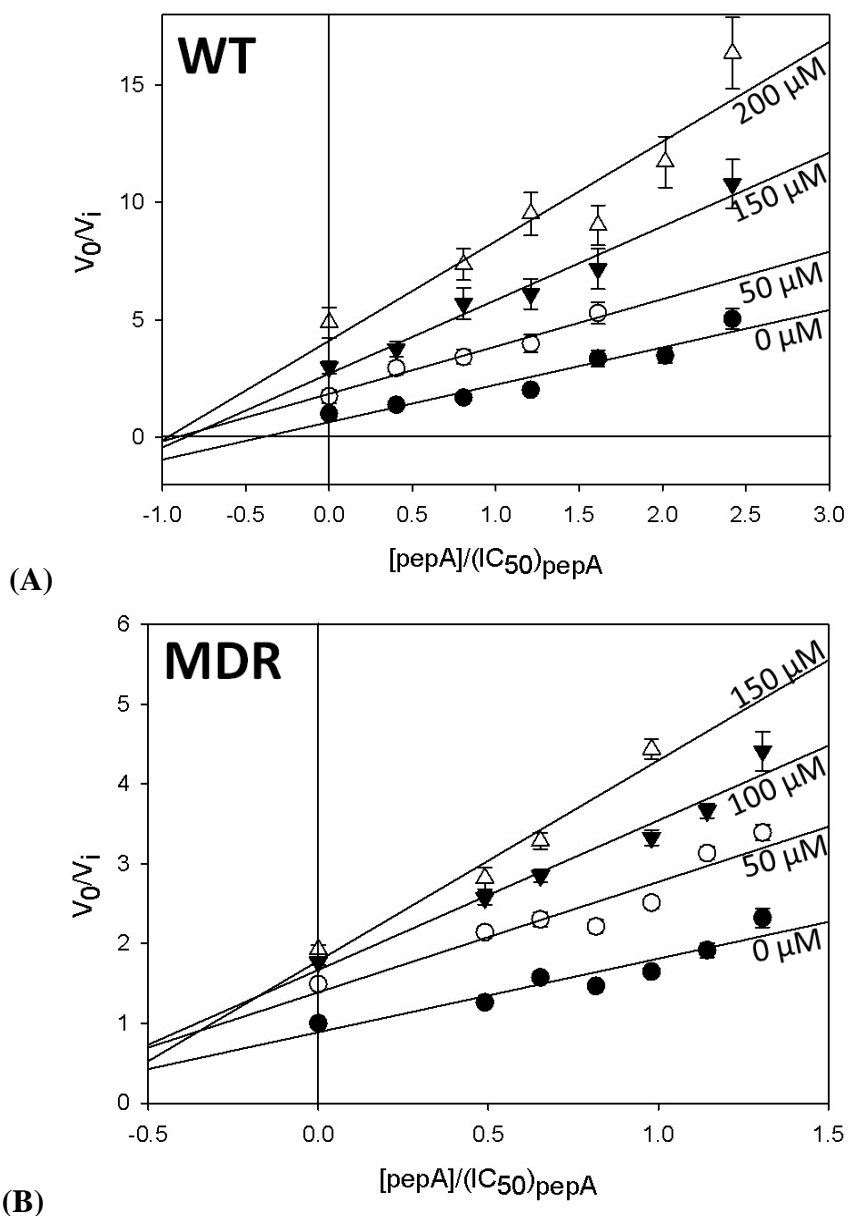
Figure 2.7. Yonetani-Theorell Inhibition analysis of compound NIT (30 nM HIV-1p, 5 μM substrate). **(A)** wild-type protease with compound NIT at concentrations: ●, 0 μM; ○, 90 μM; ▼, 150 μM; and Δ, 200 μM. **(B)** MDR protease with compound NIT at concentrations: ●, 0 μM; ○, 50 μM; ▼, 100 μM; and Δ, 150 μM. The interaction term γ determined from the curves is approximately 1 for both wild-type and MDR HIV-1p.

HIV-1p with Glide, and the docked conformation of NIT with the best score was used as the starting conformation in HIV-1p for simulations. During the MD simulations, NIT readjusted its position and remained in the Eye site throughout most of the simulation

trajectories. In one of the trajectories, NIT briefly left the Eye site for 2 ns but returned to the Eye site and remained there for the rest of the trajectory. This behavior was seen in our previous study of compound **1**. The median RMSD of NIT throughout the simulation trajectories when compared to the starting frame of the trajectory is 5.1 Å and 50% of the trajectories has RMSD ranged between 3.9 Å to 6.1 Å (Figure 2.8, Table 2.8). NIT remained in the site with a median center-of-mass (COM) distance of 4.7 Å when measuring the distance between all atoms of NIT and the Eye site residues. The COM distance is between 3.3 Å to 5.9 Å in 50% of the analyzed trajectories (Figure 2.9). This demonstrates that NIT binds stably to the Eye site of HIV-1p.

(A)   (B)

Figure 2.8. **(A)** Root-mean-square distance of five independent NIT-protease MD simulations throughout the production runs. The median RMSD of all five trajectories is 1.96 Å, with standard deviation of 0.32 Å and variance of 0.11 Å. **(B)** Atomic fluctuations by residue of five independent NIT-protease MD simulations. The maximum fluctuations occur at the flap region (Ile47 – Phe53) on both monomers with the maximum fluctuation (> 2.0 Å) centered at residue Ile50.

Table 2.8. Statistics of RMSD of NIT-protease MD simulation trajectories

	Run 1	Run 2	Run 3	Run 4	Run 5	Average
Median (Å)	1.99	1.68	1.86	1.98	2.16	1.93
St. Dev. (Å)	0.40	0.30	0.20	0.28	0.43	0.32
Variance (Å)	0.16	0.09	0.04	0.08	0.18	0.11
Maximum (Å)	3.52	2.85	2.65	2.72	4.05	3.16
Minimum (Å)	1.09	1.25	1.32	1.12	1.17	1.19
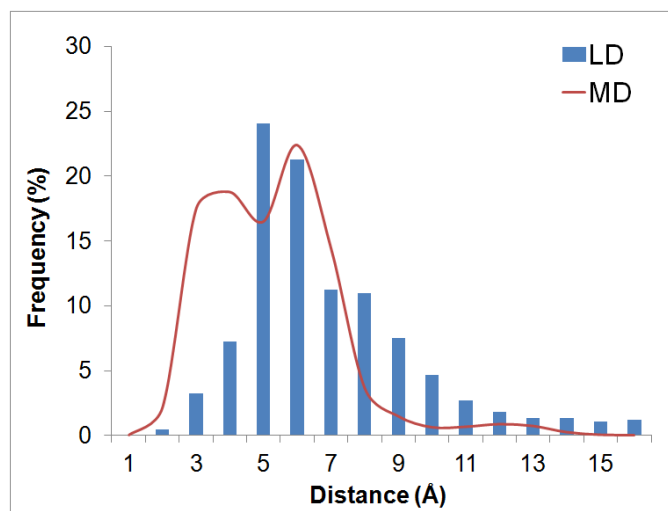
Figure 2.9. Distribution of distance (Å) between the center-of-mass of NIT and HIV-1p Eye site in MD and LD simulations. In MD simulations (line), NIT remained in HIV-1p monomer A Eye site throughout all trajectories. 50% of the population has center-of-mass distance between the ligand and the Eye site ranging from 3.3 Å to 5.9 Å with median distance at 4.7 Å. In LD simulations (bars) where NIT began in one of the Eye sites, NIT gradually distributed between the two Eye sites of the HIV-1p dimer, where 50% of the population has center-of-mass distance ranging from 4.7 Å to 7.7 Å with median distance at 5.6 Å.

Additionally, five independent runs of 20-ns Langevin dynamics (LD) simulations of the NIT-protease complex were conducted. As there is no dampening effect from the explicit solvent molecules, NIT and HIV-1p alike had more atomic fluctuations compared to the MD simulations (Figure 2.10, Table 2.9). Whereas compound NIT mostly remained in the Eye site of the HIV-1p throughout the trajectory of the MD simulations, NIT in all five independent LD simulations would exit the Eye site, sample conformations along the protein surface, and rebind to one of the 2 Eye sites of HIV-1p (Figure 2.11). Two possible paths of NIT transition from one Eye site to another are observed; one path moves along across the active site without full dissociating from the HIV-1p as seen previously in Damm *et al.*[1] The other path involves a full exit of the HIV-1p, sampling the surface of the flaps and rebinding the Eye site on the opposite monomer

(Figure 2.11 B). This indicates that NIT is able to sample both of the available Eye sites. Overall, compound NIT is within the Monomer 1's Eye site during ~ 40% of the trajectories at 7.0 Å COM distance cutoff and within the Monomer 2's Eye site ~ 28% at 7.0 Å COM distance cutoff. Combined, at a 7.0 Å COM distance cutoff, NIT occupied Eye sites in HIV-1p Monomer 1 or 2 during ~ 68% of the trajectory. The remaining time, it mostly sampled along the surface of the flap region. The COM distance ranged between 4.7 Å and 7.7 Å and a median distance of 5.6 Å for 50% of the analyzed trajectories. The high frequency of sampling at the Eye sites demonstrates that it is viable for ligand binding (Figure 2.9).
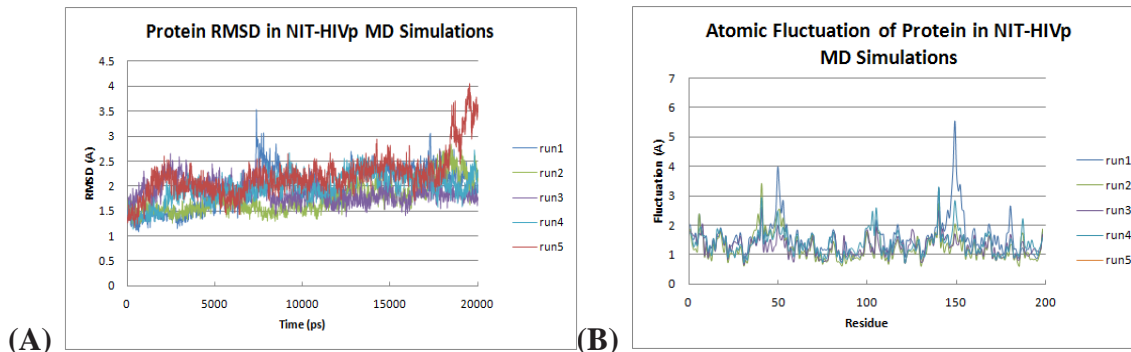

(A) **(B)**
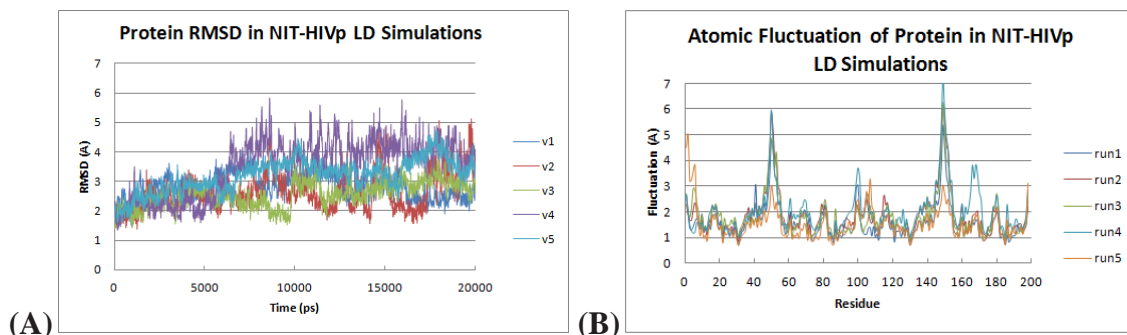
Figure 2.10. (A) Root-mean-square distance of five independent NIT-protein LD simulations throughout the production runs. The median RMSD of all five trajectories is 2.72 Å, with standard deviation of 0.51 Å and variance of 0.27 Å. **(B)** Atomic fluctuations by residue of five independent NIT-protease LD simulations. The maximum fluctuations occur at the flap region (Ile47 – Phe53) on both monomers with the maximum fluctuation (> 5.0 Å) centered at residue Ile50.

Table 2.9. Statistics of RMSD of NIT-protease LD simulation trajectories

	Run 1	Run 2	Run 3	Run 4	Run 5	Average
Median (Å)	2.84	2.53	2.54	3.27	2.44	2.72
St. Dev. (Å)	0.45	0.64	0.45	0.57	0.46	0.51
Variance (Å)	0.20	0.41	0.20	0.32	0.21	0.27
Maximum (Å)	4.07	5.13	4.04	4.74	4.06	4.41
Minimum (Å)	1.47	1.36	1.41	1.60	1.30	1.43
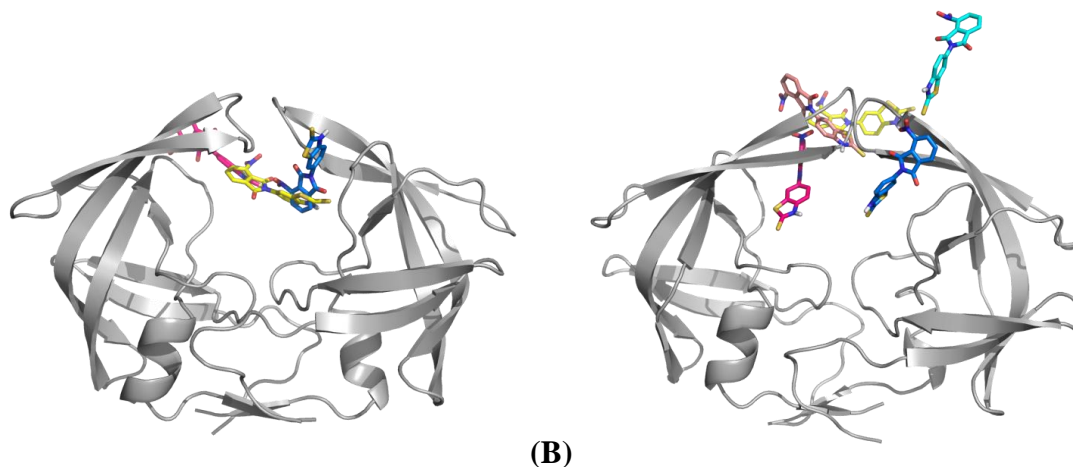
(A)   **(B)**

Figure 2.11. Paths of NIT transition between two Eye sites of HIV-1p. **(A)** NIT traveled across the active site from one Eye site to another without complete dissociation from the protease. **(B)** NIT dissociates from one Eye site and travels along to surface of the flaps before binding to the opposite Eye site.

Throughout the MD trajectories, NIT interacted primarily (> 50% of the trajectory) with the flap residues (Ile47, Gly48, Gly49, Ile50, Ile54) and with the bottom of the Eye site (Val32, Gly78, Pro79, Thr80, Pro81, Ile84). These residues were frequently with a heavy-atom cutoff distance of 4.1 Å (Figure 2.12). Of the residues that NIT consistently interacted with, Gly49, Gly78, Pro79, Thr80, and Pro81 are known to be highly conserved;[73,74] most importantly, Thr80 has such a strong influence on the flap region's flexibility that a mutation of this invariant residue results in a deleterious effect on HIV-1p catalytic activity.[38] Four residues that NIT interacts with are known to have conservative mutations (Val32(I), Ile47(V/A), Gly48(V), Ile50(V/L)) that maintain the hydrophobic nature of the side chains.[73,74] Ile54 and Ile84, which has relatively low frequency of contact with the ligand (Figure 2.12), are known to have non-conservative mutations although the most common clinically observed mutations maintain the hydrophobic character of the side chain (Ile54(V), Ile84(V/A)).
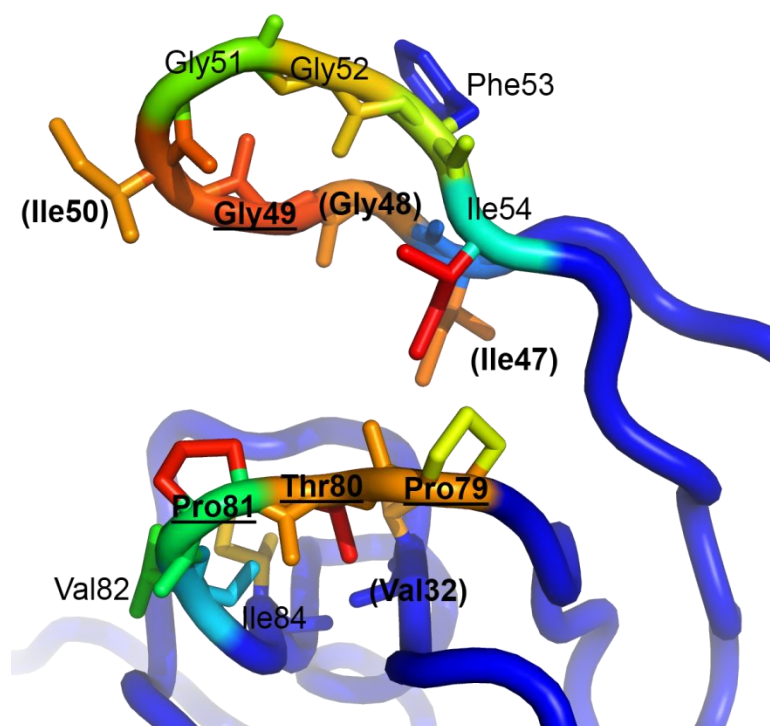
Figure 2.12. Residues in contact with compound NIT. Residues with over 30% of the trajectory in contact with NIT (heavy-atom cutoff distance 4.1 Å) during MD simulations are colored, where blue is below the threshold 30% and red is the maximum 80% of the simulations. Residues with bolded and underlined name are found to be highly conserved residues and those with name in bold and bracket are found to have conservative mutations in the clinic.

To examine the change in the dynamic behaviors of HIV-1p due to the binding of NIT to the Eye site, essential dynamics analysis was applied to analyze the MD trajectories of NIT-protease complex and an apo protease. Essential dynamics analyzes the protein dynamics by performing an orthogonal transformation of a covariance matrix of the MD trajectories and isolates the collective modes of motion of the protein. Individual mode of motion, or eigenvector, contains the collective motion of the residues of the examined protein while the associating eigenvalue quantifies the contribution of the eigenvector in the examined trajectories. Each of these eigenvectors can be visualized and inspected separately to identify the main modes of collective motion from the more

109

localized fluctuations. The first few eigenvectors, or modes of motion, usually describe the global dynamic motions of the protein (i.e., domain movement and change in protein conformation) while the higher order modes of motion usually describe the local dynamic motions (i.e., residue side-chain movements).

To compare the essential dynamics of NIT-protease complex to apo-protease, a dot-product was calculated between the eigenvectors as a measurement of the degree of overlap between the examined vectors. Visual inspection of the MD and comparison of the 1st eigenvector of both systems, as shown in Figure 2.13, indicates a significant conformational change to the flap in contact with NIT, for which the motion of the flap in NIT-protease complex and in apo protease is strongly anti-correlated (the rest of the protein remains relatively correlated). Furthermore, both 1st eigenvectors have significant contribution to the overall dynamic motion of the protease in each respective trajectory; the 1st eigenvector of NIT-protease complex has a normalized eigenvalue of 0.403 while apo-protease has a normalized eigenvalue of 0.540. This strongly suggests that the influence of the extensive interactions with compound NIT in the Eye site has a profound effect on the dynamic motion the flap throughout the trajectory, as one might expect.

Another observation is the change in the flap openness of HIV-1p in the MD trajectories through binding of NIT to the Eye site (Table 2.10). The degree of flap openness can be described by the C_α distance measured between catalytic Asp25 and flap-tip Ile50. The flap-opening distance in 50% of the trajectories ranges between 13.7 and 15.8 Å with a median of 14.8 Å, and the opposite monomer with no NIT bound to the Eye site has a similar flap opening distance between 13.0 Å to 16.1 Å with a median of 14.3 Å (Table 2.10). This is compared to the apo HIV-1p MD trajectory where 50% of
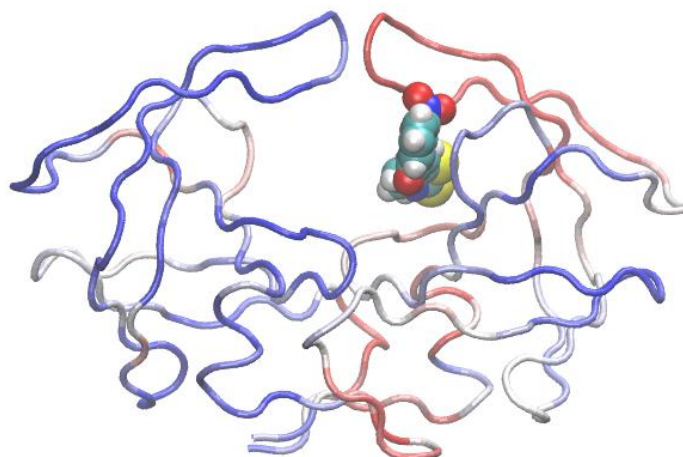
Figure 2.13. NIT binding influences the flap movement of HIV-1p. Essential dynamics of NIT-protease and apo HIV-1p MD simulations were calculated to describe the motions of the protease. Their lowest mode of motion, the 1st eigenvectors, were compared by calculating the dot-product of vectors on each residue in the eigenvectors. The dot-products were then scaled and viewed as B-factors. Blue color indicates strong correlation while red color indicates strong anti-correlation. The presence of NIT in the Eye site induces localized change in flap movement.

the trajectory has a flap opening distance of 11.9 Å to 17.3 Å and a median of 15.1 Å. The flaps under the influence of NIT remain in the semi-open conformation states comparable to the semi-open conformational states observed in apo HIV-1p (median of 14.8 Å with NIT and 14.3 Å without NIT vs. 15.1 Å in apo HIV-1p) and has a small range of motion compared to the apo HIV-1p (RMS distance ~ 2.7 Å for NIT-protease complex vs. ~ 3.8 Å for apo HIV-1p).

We also observed a change in the width of the Eye site relative to the apo HIV-1p, which can be measured by the C_α distance between Gly51 and Thr80 (Table 2.10). In the NIT-protease complex simulations, the Eye site with NIT bound has a width that ranges between 11.3 Å and 14.4 Å and a median width of 12.8 Å, while the opposite Eye site with no bound NIT has width between 10.6 Å to 13.6 Å and a median of 12.1 Å. The two Eye sites of the apo HIV-1p MD trajectory have a similar width distribution that ranges

Table 2.10. Distribution flap openness and Eye site width of NIT-protease MD simulations

	Flap openness (Asp25 C_α - Ile50 C_α) Å				
	2.5%	25.0%	50.0%	75.0%	97.5%
A (NIT)	11.0	13.7	14.8	15.8	24.4
B (none)	11.4	13.0	14.3	16.1	20.1
apo	10.0	11.9	15.1	17.3	25.7
	Eye site width (Gly51 C_α - Thr80 C_α) Å				
	2.5%	25.0%	50.0%	75.0%	97.5%
A (NIT)	8.8	11.3	12.8	14.4	17.6
B (none)	8.6	10.6	12.1	13.6	15.6
apo	4.7	8.4	9.7	11.9	17.0

between 8.4 Å and 11.9 Å and a median width of 9.7 Å. As such, the width of the Eye site in the NIT-protease complex simulations is notably larger than the Eye site in the apo HIV-1p simulation (median of 12.8 Å with NIT and 12.1 Å without NIT vs. 9.7 Å in apo HIV-1p). This demonstrates the presence of NIT affects the flap mobility and flap conformation through its binding to the Eye site.

2.7 Conclusion

We have described the discovery of a novel small molecule that probes the allosteric Eye site of HIV-1p. Compound **NIT**, with a novel scaffold, was identified as a molecular probe that can bind to the Eye site of HIV-1p and modulate its catalytic activity. NIT demonstrates a novel mechanism in modulating the HIV-1p proteolytic activity. It has a mixed competitive inhibition character in the Michaelis-Menton enzymatic kinetics, demonstrating its potential to act on the HIV-1p through a mechanism other than competing for the active site. It is likely that NIT modulates HIV-

1p proteolytic activity through the allosteric Eye site of the protease, as shown in the Yonetani-Theorell experiment. MD simulations of a NIT-protease complex show that NIT remains stably bound in the Eye site and affects the dynamics of the β-hairpin flaps when compared to an apo HIV-1p simulation. Moreover, compound NIT shares similar chemical features and binding modes in the Eye site of HIV-1p with the molecular probe 5NI that co-crystallized with apo HIV-1p in a semi-open conformation. These data support compound NIT's ability to allosterically modulate HIV-1p proteolytic activity through the binding to the Eye site of HIV-1p and effecting the protease through altering its dynamics.

2.8 Acknowledgements

2.9 References

1. Damm, K. L.; Ung, P. M.-U.; Quintero, J. J.; Gestwicki, J. E.; Carlson, H. A. A Poke in the Eye: Inhibiting HIV-1 Protease through its Flap-Recognition Pocket. *Biopolymers* **2008**, *89*, 643-652.

2. Perryman, A. L.; Zhang, Q.; Soutter, H. H.; Rosenfeld, R.; McRee, D. E.; Olsen, A. J.; Elder, J. E.; Stout, C. D. Fragment-Based Screen Against HIV Protease. *Chem. Biol. Drug Des.* **2010**, *75*, 257-268.

3. Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. The Energy Landscape and Motions of Proteins. *Science* **1991**, *254*, 1598-1603.

4. Boehr, D. D.; Nussinov, R.; Wright, P. E. The Role of Dynamic Conformational Ensembles in Biomolecular Recognition. *Nat. Chem. Bio.* **2009**, *5*, 789-796.

5. Carlson, H. A. Protein Flexibility and Drug Design: How to Hit a Moving Target. *Curr. Opin. Chem. Biol.* **2002**, *6*, 447-452.

6. Carlson, H. A. Protein Flexibility is an Important Component of Structure-Based Drug Discovery. *Curr. Pharm. Des*. **2002**, *8*, 1571-1578.

7. Teague, S. J. Implications of Protein Flexibility for Drug Discovery. *Nat. Rev. Drug Discov.* **2003**, *2*, 527-541.

8. Kroeger Smith, M. B.; Rouzer, C. A.; Taneyhill, L. A.; Smith, N. A.; Hughes, S. H.; Boyer, P. L.; Janssen, P. A. J.; Moereels, H.; Koymans, L.; Arnold, E.; Ding, J.; Das, K.; Zhang, W.; Michejda, C. J.; Smith, R. H. Molecular Modeling Studies of HIV-1 Reverse Transcriptase Nonnucleoside Inhibitors: Total Energy of Complexation as a Predictor of Drug Placement and Activity. *Protein Sci.* **1995**, *4*, 2203-2222.

9. Kohl, N. E.; Emini, E. A.; Schleif, W. A.; Davis, L. J.; Heimbach, J. C.; Dixon, R. A.; Scolnick, E. M.; Sigal, I. S. Active Human Immunodeficiency Virus Protease is Required for Viral Infectivity. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 4686-4690.

10. Flexner, C. HIV Drug Development: the Next 25 Years. *Nat. Rev. Drug Dis.* **2007**, *6*, 959-966.

11. Wlodawer, A.; Miller, M.; Jaskólski, M.; Sathyanarayana, B. K.; Baldwin, E.; Weber, I. T.; Selk, L. M.; Clawson, L.; Schneider, J.; Kent, S. B. H. Conserved Folding in Retroviral Proteases: Crystal Structure of a Synthetic HIV-1 Protease. *Science* **1989**, *245*, 616-621.

12. Prabu-Jeyabalan, M.; King, N. M.; Nalivaika, E. A.; Heilek-Snyder, G.; Cammack, N.; Schiffer, C. A. Substrate Envelope and Drug Resistance: Crystal Structure of Ro1 in Complex with Wild-Type Human Immunodeficiency Virus Type 1 Protease. *Antimicrob. Agents Chemother.* **2006**, *50*, 1518-1521.

13. Ding, F.; Layten, M.; Simmerling, C. Solution Structure of HIV-1 Protease Flaps Probed by Comparison of Molecular Dynamics Simulation Ensembles and EPR Experiments. *J. Am. Chem. Soc.* **2008**, *130*, 7184-7185.

14. Galiano, L.; Ding, F.; Veloro, A. M.; Blackburn, M. E.; Simmerling, C.; Fanucci, G. E. Drug Pressure Selected Mutations in HIV-1 Protease Alter Flap Conformations. *J. Am. Chem. Soc.* **2009**, *131*, 430-431.

15. Ishima, R.; Louis, J. M. A Diverse View of Protein Dynamics from NMR Studies of HIV-1 Protease Flaps. *Proteins* **2008**, *70*, 1408-1415.

16. Perryman, A. L.; Lin, J.-H.; McCammon, J. A. Restrained Molecular Dynamics Simulations of HIV-1 Protease: the First Step in Validating a New Target for Drug Design. *Biopolymers* **2006**, *82*, 272-284.

17. Sadiq, S. K.; De Fabritiis, G. Explicit Solvent Dynamics and Energetics of HIV-1 Protease Flap Opening and Closing. *Proteins* **2010**, *78*, 2873-2885.

18. Karthik, S.; Senapati, S. Dynamic Flaps in HIV-1 Protease Adopt Unique Ordering at Different Stages in the Catalytic Cycle. *Proteins* **2011**, *79*, 1830-1840.

19. Hornak, V.; Okur, A.; Rizzo, R. C.; Simmerling, C. HIV-1 Protease Flaps Spontaneously Open and Reclose in Molecular Dynamics Simulations. *Proc. Nat. Am. Sci. U.S.A.* **2006**, *103*, 915-920.

20. Hornak, V.; Okur, A.; Rizzo, R. C.; Simmerling, C. HIV-1 Protease Flaps Spontaneously Close to the Correct Structure in Simulations Following Manual Placement of an Inhibitor into the Open State. *J. Am. Chem. Soc.* **2006**, *128*, 2812-2813.

21. Hornak, V.; Simmerling, C. Targeting Structural Flexibility in HIV-1 Protease Inhibitor Binding. *Drug Discov. Today* **2007**, *12*, 132-138.

22. Deng, N.-J.; Zheng, W.; Gallicchio, E.; Levy, R. M. Insights into the Dynamics of HIV-1 Protease: A Kinetic Network Model Constructed from Atomistic Simulations. *J. Am. Chem. Soc*. **2011**, *133*, 9387-9394.

23. Menard, P. R.; Mason, J. S.; Morize, I.; Bauerschmidt, S. Chemistry Space Metrics in Diversity Analysis, Library, Design, and Compound Selection. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1204-1213.

24. Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28-35.

25. Masek, B. B.; Shen, L.; Smith, K. M.; Pearlman, R. S. Sharing Chemical Information without Sharing Chemical Structure. *J. Chem. Inf. Comput. Sci.* **2008**, *48*, 256-261.

26. SYBYL version 8.0; Tripos International: St. Louis, MO, **2008**.

27. Molecular Operating Environment (MOE), version 2010.10; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, **2010**.

28. Martin, Y. C. 3D Database Searching in Drug Design. *J. Med. Chem.* **1992**, *35*, 2145-2154.

29. Hurst, T. Flexible 3D Searching: The Directed Tweak Technique. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190-196.

30. OMEGA version 2.3.2; OpenEye Scientific Software: Santa Fe, NM, **2009**.

31. EON version 2.0.1; ROCS version 3.0.0; OpenEye Scientific Software: Santa Fe, NM, **2009**.

32. Matayoshi, E. D.; Wang, G. T.; Krafft, G. A.; Erickson, J. Novel Fluorogenic Substrates for Assaying Retroviral Proteases by Resonance Energy Transfer. *Science* **1990**, *247*, 954-958.

33. Jordan, S. P.; Zugay, J.; Darke, P. L.; Kuo, L. C. Activity and Dimerization of Human Immunodeficiency Virus Protease as a Function of Solvent Composition and Enzyme Concentration. *J. Biol. Chem.* **1992**, *267*, 20028-20032.

34. Toth, M. V.; Marshall, G. R. A Simple, Continuous Fluorometric Assay for HIV Protease. *Int. J. Pept. Prot. Res*. **1990**, *36*, 544-550.

35. Woon, T. C.; Brinkworkth, R. I.; Fairlie, D. P. Inhibition of HIV-1 Proteinase by Metal Ions. *Int. J. Biochem.* **1992**, *24*, 911-914.

36. York, D. M.; Darden, T. A.; Pedersen, L. G.; Anderson, M. W. Molecular Modeling Studies Suggest that Zinc Ions Inhibit HIV-1 Protease by Binding at Catalytic Aspartates. *Envir. Health Persp.* **1993**, *101*, 246-250.

37. Bowman, M. J.; Byrne, S.; Chmielewski, J. Switching between Allosteric and Dimerization Inhibition of HIV-1 Protease. *Chem. Biol.* **2005**, *12*, 439-444.

38. Foulkes, J. E.; Prabu-Jeyabalan, M.; Cooper, D.; Henderson, G. J.; Harris, J.; Swanstrom, R.; Schiffer, C. A. Role of Invariant Thr80 in Human Immunodeficiency Virus Type 1 Protease Structure, Function, and Viral Infectivity. *J. Virol*. **2006**, *80*, 6906-6916.

39. Zhang, Z. Y.; Poorman, R. A.; Maggiora, L. L.; Heinrikson, R. L.; Kezdy, F. J. Dissociative Inhibition of Dimeric Enzymes: Kinetic Characterization of the Inhibition of HIV-1 Protease by Its COOH-Terminal Tetrapeptide. *J. Biol. Chem.* **1991**, *266*, 15591-15594.

40. Schramm, H. J.; Boetzel, J.; Buttner, J.; Fritsche, E.; Gohring, W.; Jaeger, E.; Konig, S.; Thumfart, O.; Wenger, T.; Nagel, N. E.; Schramm, W. The Inhibition of Human Immunodeficiency Virus Protease by 'Interface Peptides'. *Antiviral Res*. **1996**, *30*, 155-170.

41. Breccia, P.; Boggetto, N.; Perez-Fernandez, R.; Van Gool, M; Takahashi, M.; Rene, L.; Prados, P.; Badet, B.; Reboud-Ravaux, M.; de Mendoza, J. Dimerization Inhibitors of HIV-1 Protease Based on a Bicyclic Guanidinium Subunit. *J. Med. Chem.* **2003**, *46*, 5196-5207.

42. Yonetani, T.; Theorell, H. Studies on Liver Alcohol Hydrogenase Complexes. 3. Multiple Inhibition Kinetics in the Presence of Two Competitive Inhibitors. *Arch. Biochem. Biophys.* **1964**, *106*, 243-251.

43. Martinez-Irujo, J. J.; Villahermosa, M. L.; Mercapide, J.; Cabodevilla, J. F.; Santiago, E. Analysis of the Combined Effect of Two Linear Inhibitors on a Single Enzyme. *Biochem. J.* **1998**, *329*, 689-698.

44. Chang, M. W.; Giffin, M. J.; Muller, R.; Savage, J.; Lin, Y. C.; Hong, S.; Jin, W.; Whitby, L. R.; Elder, J. H.; Boger, D. L.; Torbett, B. E. Identification of Broad-Based HIV-1 Protease Inhibitors from Combinatorial Libraries. *Biochem. J.* **2010**, *429*, 527-532.

45. FRED version 2.2.5; OpenEye Scientific Software: Santa Fe, NM, **2009**.

46. Sham, H. L.; Zhao, C.; Stewart, K. D.; Betebenner, D. A.; Lin, S.; Park, C. H.; Kong, X.-P.; Rosenbrook, W.; Herrin, T.; Madigan, D.; Vasavanonda, S.; Lyons, N.; Molla, A.; Saldivar, A.; Marsh, K. C.; McDonald, E.; Wideburg, N. E.; Denissen, J. F.; Robins, T.; Kempf, D. J.; Plattner, J. J.; Norbeck, D. W. A Novel, Picomolar Inhibitor of Human Immunodeficiency Virus Type 1 Protease. *J. Med. Chem.* **1996**, *39*, 392-397.

47. McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian Docking Functions. *Biopolymers* **2003**, *68*, 76-90.

48. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins* **2006**, *65*, 712-725.

49. Case, D. A.; Darden, T. A.; Cheatham, III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvary, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER 10*, University of California, San Francisco.: **2008**.

50. Spinelli, S.; Liu, Q. Z.; Alzari, P. M.; Hirel, P. H.; Poljak, R. J. The Three-Dimensional Structure of the Aspartyl Protease from the HIV-1 Isolate BRU. *Biochimie* **1991**, *73*, 1391-1396.

51. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235-242.

52. PyMOL Molecular Graphics System, Version 1.2 Schrödinger, LLC.

53. Glide, version 5.5, Schrödinger, LLC, New York, **2009**.

54. Dupradeau, F.-Y.; Pigache, A.; Zaffran, T.; Savineau, C.; Lelong, R.; Grivel, N.; Lelong, D.; Rosanski, W.; Cieplak, P. The R.E.D. Tools: Advances in RESP and ESP Charge Derivation and Force Field Library Building. *Phys. Chem. Chem. Phys.* **2010**, *12*, 7821-7839.

55. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157-1174.

56. Wang, J.; Wang, W.; Kollman P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247-260.

57. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926-935.

58. Darden, T. A.; York, D. M.; Pedersen, L. G. Particle Mesh Ewald. An N.log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089-10092.

59. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of *n*-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327-341.

60. Meagher, K. L.; Carlson, H. A. Solvation Influences Flap Collapsein HIV-1 Protease. *Proteins* **2005**, *58*, 119-125.

61. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical Treatment of Salvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127-6129.

62. García, A. E. Large-Amplitude Nonlinear Motions in Proteins. *Phys. Rev. Lett.* **1992**, *68*, 2696-2699.

63. Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential Dynamics of Proteins. *Proteins* **1993**, *17*, 412-425.

64. Mongan, J. Interactive Essential Dynamics. *J. Comp.-aided Mol. Des.* **2004**, *18*, 433-436.

65. Humphrey, W.; Dalke, A.; Schulten, K. VMD − Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33-38.

66. Carlson, H. A.; Masukawa, K. M.; McCammon, J. A. Method for Including the Dynamic Fluctuations of a Protein in Computer-Aided Drug Design. *J. Phys. Chem. A* **1999**, *103*, 10213-10219.

67. Carlson, H. A.; Masukawa, K. M.; Jorgensen, W. L.; Lins, R. D.; Briggs, J. M.; McCammon, J. A. Developing a Dynamic Pharmacophore Model for HIV-1 Integrase. *J. Med. Chem.* **2000**, *43*, 2100-2114.

68. Meagher, K. L.; Carlson, H. A. Incorporating Protein Flexibility in Structure-Based Drug Discovery: Using HIV-1 Protease as a Test Case. *J. Am. Chem. Soc.* **2004**, *126*, 13276-13281.

69. Meagher, K. L.; Lerner, M. G.; Carlson, H. A. Refining the Multiple Protein Structure Pharmacophore Method: Consistency Across Three Independent HIV-1 Protease Models. *J. Med. Chem.* **2006**, *49*, 3478-3484.

70. Irwin, J. J.; Shoichet, B. K. ZINC – a Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177-182.

71. Shoichet, B. K. Interpreting Steep Dose-Response Curves in Early Inhibitor Discovery. *J. Med. Chem.* **2006**, *49*, 7274-7277.

72. Asante-Appiah, E.; Chan, W. W. Analysis of the Interactions between an Enzyme and Multiple Inhibitors using Combination Plots. *Biochem. J.* **1996**, *320*, 17-26.

73. Shafer, R. W.; Rhee, S. Y.; Pillay, D.; Miller, V.; Sandstrom, P.; Schapiro, J. M.; Kuritzkes, D. R.; Bennett, D. HIV-1 Protease and Reverse Transcriptase Mutations for Drug Resistance Surveillance. *AIDS* **2007**, *21*, 215-223.

74. Johnson, V. A; Brun-Vézinet, F.; Clotet, B.; Günthard, H. F.; Juritzkes, D. R.; Pillay, D.; Schapiro, J. M.; Richman, D. D. Update of the Drug Resistance Mutations in HIV-1: December 2010. *Top HIV Med.* **2010**, *18*, 156-163.

CHAPTER 3

Computational Study of Allosteric Inhibitors of HIV-1 Protease

3.1 Abstract

Chang *et al.*[1] proposed that a recently discovered HIV-1 protease (HIV-1p) inhibitor acts through an allosteric site, the Elbow site, based on enzyme kinetics and a simple standard docking calculation. Comprehensive standard docking against multiple conformations of HIV-1p indicated that another allosteric site, the Eye site, may also be a potential site for ligand binding. However, the docking results were not sufficient in deducing the most probable ligand binding site. To better understand the protein-ligand interaction, Langevin dynamics (LD) was used. LD simulations indicated that HIV-1p in closed form transforms into semi-open form spontaneously. Using MM/PBSA to estimate binding free energy, we found that binding to the Eye site was more favorable than in the Elbow site for this allosteric inhibitor. However, binding to both sites is highly favorable, and we cannot completely rule out the possibility of the inhibitor acting at the Elbow site. Therefore, the Eye site should be considered as an equally potential binding site as the Elbow site for the ligand.

3.2 Introduction

In a high-throughput screening study, Chang *et al.*[1] identified a new class of compounds that disrupt the enzymatic activity of HIV-1 protease (Figure 3.1 A). This compound is somewhat pseudo-symmetric as it contains a benzoxazole as the central scaffold and flanked by two thiophene moieties through amide linkage. In this study, we refer this molecule as compound **2**.

Using enzyme kinetic data, Chang *et al.* demonstrated that this compound acted on HIV-1p through a non-competitive mechanism as it did not compete with the well-characterized competitive inhibitor pepstatin A.[2,3] Although the authors supported the non-competitive nature of compound **2** with thorough kinetics experiments, these experiments could not provide concrete evidence as to where compound **2** may actually bind. The authors suggested the Elbow site (Figure 3.1 B) as the allosteric site[4,5] where compound **2** potentially binds based on a simple molecular docking study on a HIV-1p crystal structure with closed conformation, PDB: 2HS1.[6]

Chang *et al.* suggested the Elbow site based on two clues. First, computational studies have suggested this site may influence the flap movement.[4,7-9] Another clue was that the inhibition constant (K_i) estimated by AutodDock Vina, a docking software, is very similar to the experimentally determined K_i for compound **2**, both around 5 μM. However, the authors did not examine binding to the Eye site (Figure 3.1 B) in their docking study, likely because this site is not available in the HIV-1p crystal structure that they examined. To assess the potential of compound **2** binding to the Eye site, we performed molecular docking with several docking methods on HIV-1p crystal structures

in both open and closed conformations. We then assess the interactions between compound **2** and HIV-1p in both conformations with Langevin Dynamics (LD) simulations and estimate the binding free energy with MM/PBSA.
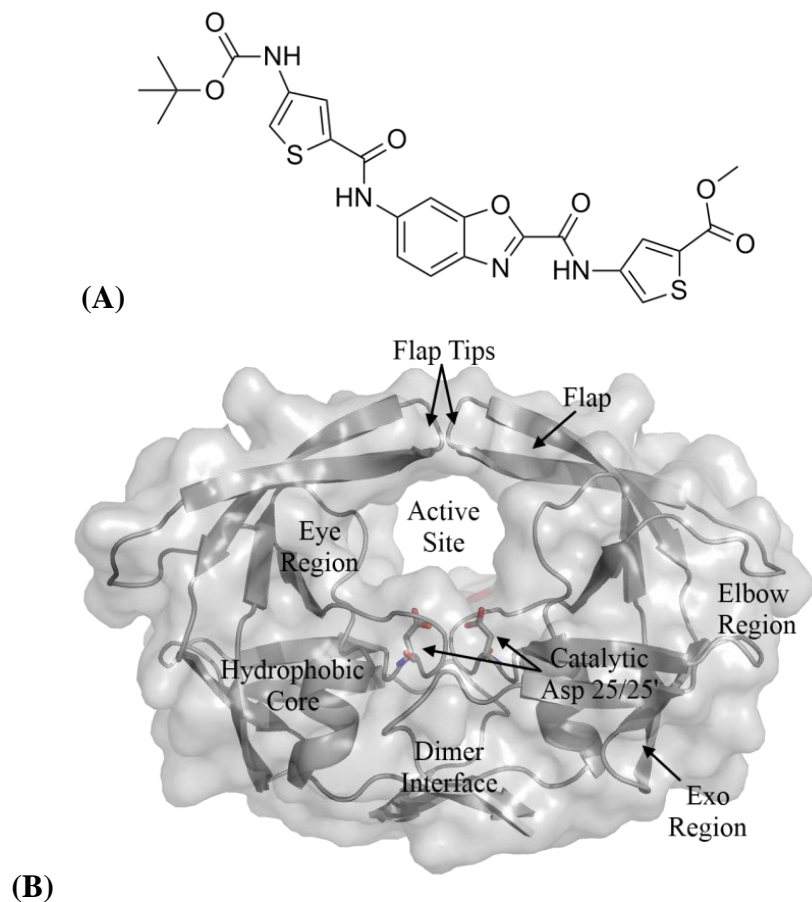

(A)

(B)

Figure 3.1. (**A**) Compound **2** described by Chang *et al*. (**B**) A cartoon representation with protein surface of HIV-1p in the semi-open conformation (PDB ID: 1HHP).

3.3 Computational Methods

Dynamics Simulations

Unrestrained all-atom Langevin dynamic (LD) simulation was performed with the FF99SB force field[10] and the AMBER10 suite of programs.[11] An apo HIV-1p in semi-open conformation, PDB: 1HHP,[12] was obtained from the PDB.[13] C_2-symmetric homodimer of HIV-1p was generated with PyMOL version 1.2. The ligand was placed into the Eye or Elbow site of HIV-1p using AutoDock 4.2.3[14] (AD4). Ligand parameters were built with GAFF force field[15] and the AM1-BCC charges[16,17] from the Antechamber module of AMBER.[18]

Ten independent, 10-ns LD simulations were also performed. The SHAKE algorithm[19] was used to restrain hydrogen atoms. A 999 Å cut-off distance for non-bonded interactions was used. Generalized Born approach was used to implicitly model aqueous solvation for the LD simulations.[20] Model II of a modified Generalized Born approach[21] was used to implicitly model aqueous solvation. Default dielectric values were used, interior = 1 and exterior = 78.5. The time step and the collision frequency of the simulation were 1 fs and 1 ps^{-1}, respectively.

Simulations began with hydrogen atom minimization, followed by side-chain and then all-atom minimizations. Equilibration was done in six stages: the system was gradually heated from 100 K to the final temperature at 300 K in the first two equilibration steps. Restraints were placed on all heavy atoms and gradually removed over the first four equilibration steps using force constants from 2.0 to 0.1 kcal/mol·Å^2, where the first three steps were done over 10 ps and the fourth step over 50 ps. Only the

backbone atoms were restrained at 0.1 kcal/mol·$Å^2$ over 50 ps in the fifth equilibration step. In the sixth equilibration step, the restraints were removed and all atoms were allowed to move for 300 ps at 300 K. The production phase was run for 10 ns. For the 10 runs of each system, independent velocities were obtained by using default random number seeds, which lead to different random trajectories. Analysis of the trajectories was performed using the PTRAJ module of AMBER suite. MM/PBSA was performed to estimate the partial binding free energy of the ligand-receptor complex; the entropy term, TΔS, was not calculated due to the complexity involved in the normal mode calculation. Default settings were used. Poisson-Boltzmann function was used instead of the Generalized-Born estimation in desolvation free energy calculation.

Molecular Docking

Molecular docking was performed on HIV-1p in semi-open form (PDB: 1HHP) and closed forms (PDB: 1PRO[22] and 2HS1[6]). Several docking methods were used: AD4, AutoDock Vina 1.02[23] (Vina), and MOE.[24] Default settings were used on AD4, Vina, and MOE. To obtain the reference docked pose of compound **2** similar to the published docked pose by Chang *et al.*, the ligand was docked to the Elbow site of 2HS1 and the best pose, which has very similar predicted binding free energy and strong resemblance to the published pose, was chosen to be the reference.

3.4 Results and Discussion

Molecular Docking

To compare the binding of compound **2** to the Elbow versus the Eye sites of the HIV-1p, we performed molecular docking with AD4, Vina, and MOE. Three HIV-1p crystal structures were used in the docking study: the semi-open conformation in 1HHP and the closed conformations of 1PRO and 2HS1. Since Chang *et al.* used 2HS1 in their original study, we performed molecular docking on 2HS1 as a "control". 1PRO and 2HS1 do not have an open Eye site because they are in closed conformations and the Eye site is filled with flap-tip residues.[25] Therefore, docking was done only to their Elbow sites. The asymmetric unit of 1PRO and 2HS1 is the full HIV-1p dimer; hence, both Elbow sites in each crystal structure were docked individually to compare any possible variation in the structure. Conversely, the asymmetric unit of 1HHP, which possesses both Elbow and Eye sites, is a monomer, hence docking was performed once on both sites in the symmetric dimer.

When docking compound **2** to the Eye site of a semi-open HIV-1p (1HHP), all three docking methods placed the ligand inside the Eye site and ranked this pose as the best-scored docking pose (Figure 3.2 A). In general, these docked poses occupy the same space within the Eye site and overlap with the Eye site pharmacophore model proposed by Damm *et al.*[25] Notably, the most favorable poses placed the hydrophobic core, the benzoxazole scaffold, in close proximity to the hydrophobic/aromatic elements of the pharmacophore model. Perryman *et al.*[5] performed fragment-based X-ray crystallography and obtained a semi-open form HIV-1p with resolved 5-nitroindole electron density
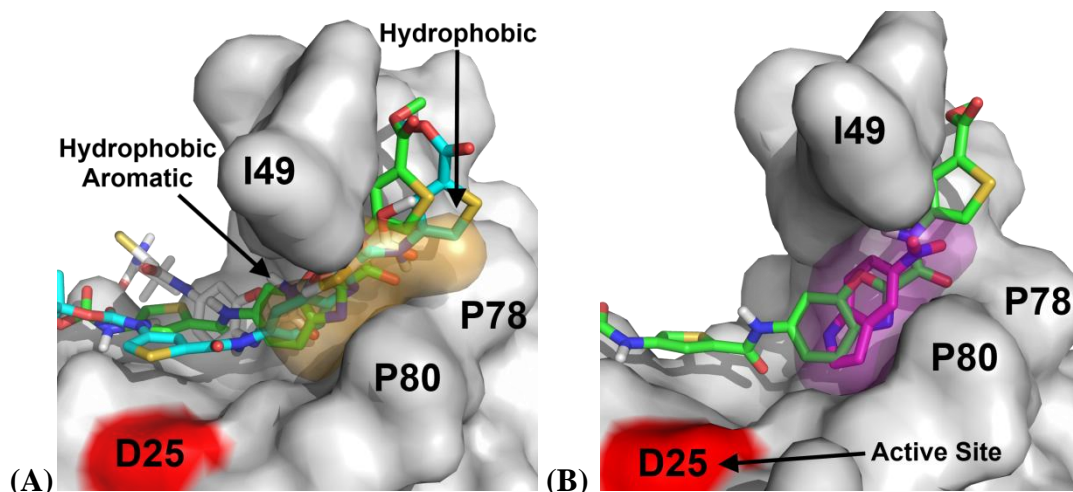
125

Figure 3.2. (A) Docked poses of compound **2** in the Eye site of semi-open HIV-1 protease crystal structure, which also overlaps with the Eye site pharmacophore[25] is shown as orange transparent surface. **(A)** 5-nitroindole,[5] colored in purple, overlaps with the benzoxazole scaffold of compound **2** docked poses by AD4. Protein is shown in white surface. Docked poses are generated by: AD4 (green), Vina (white), and MOE (cyan). The RMSD of the poses are: AD4 vs. Vina, 6.5 Å; AD4 vs. MOE, 9.9 Å; Vina vs. MOE, 11.1 Å.

within the Eye site. The best-scored docked poses of compound **2** overlap with 5-nitroindole, placing the benzoxazole scaffold in the same location or in close proximity to the 5-nitroindole (Figure 3.2 B).

Reproducing the work of Chang *et al*., we used Vina to generate an initial docked pose of compound **2** in the Elbow site of 2HS1. Visual comparison with the figures in their paper suggested that the best-scored docked pose was an appropriate reference to compare to their work. All three docking methods (AD4, Vina, and MOE) were able to generate docked poses similar to the reference docked pose in the Elbow sites of 2HS1, 1PRO, and 1HHP. As shown in Figure 3.3, the poses are all placed within the volume of the reference docked pose. The actual docked poses were structurally different from one another as each software applied different methods for pose generation. Furthermore, the local structure of each crystal structure varies; hence, all docked poses were different

2HS1

2HS1	RMSD (Å)
AD4 vs. Vina	2.1
AD4 vs. MOE	10.3
Vina vs. MOE	9.9

(A)

1PRO

1PRO	RMSD (Å)
AD4 vs. Vina	15.8
AD4 vs. MOE	10.4
Vina vs. MOE	11.1

(B)

1HHP

1HHP	RMSD (Å)
AD4 vs. Vina	1.7
AD4 vs. MOE	11.0
Vina vs. MOE	10.8

(C)
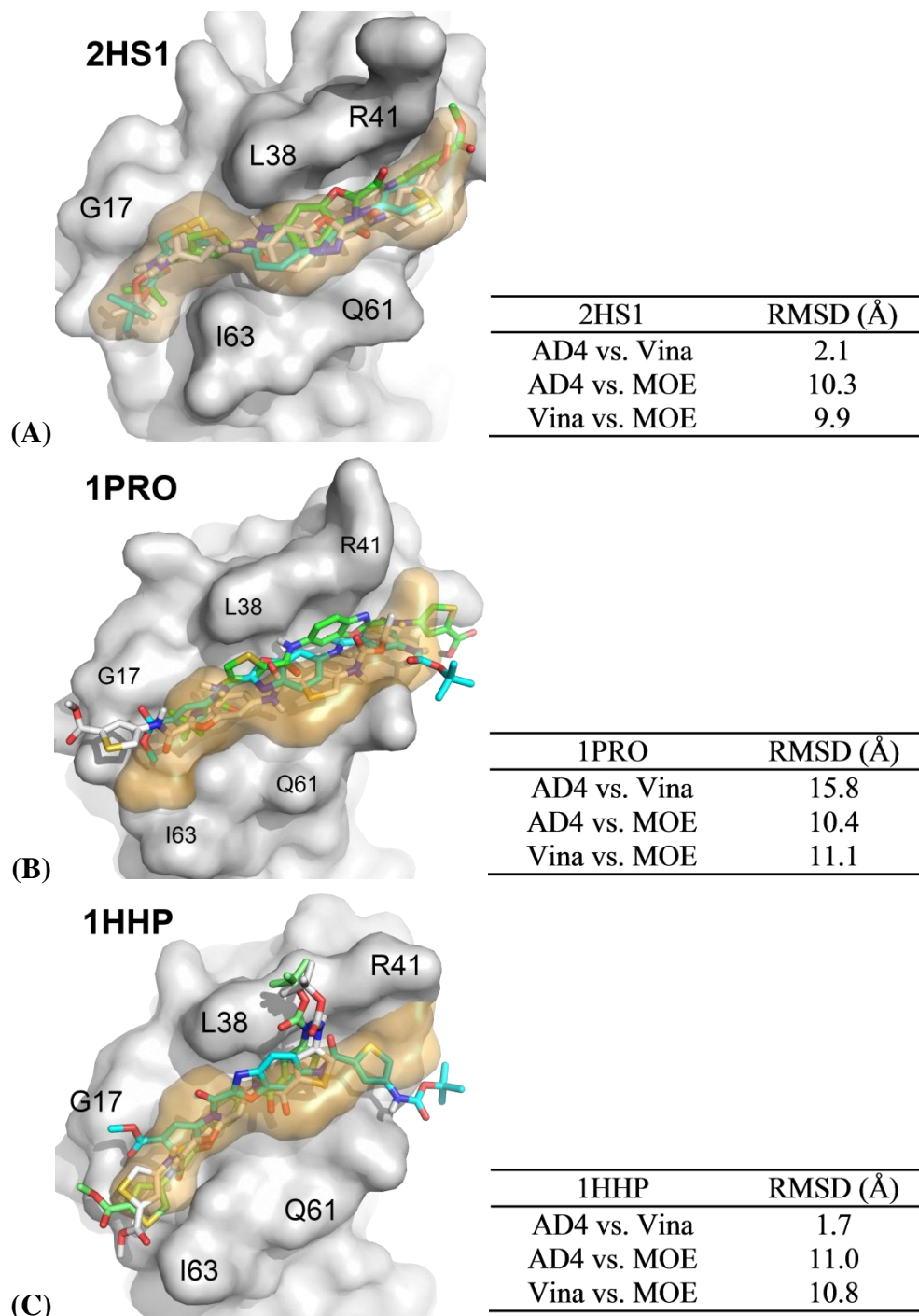
Figure 3.3. Docked poses of compound **2** in the Elbow site of HIV-1 protease crystal structures: **(A)** 2HS1, **(B)** 1PRO and **(C)** 1HHP. To obtain a reference pose of compound **2** similar to the published docked pose by Chang *et al*., the ligand was docking to the Elbow site of 2HS1 with Vina and the best pose with strong resemblance to the published pose was used as the reference (shown in orange surface). Protein is shown as white surface. The software used to generate the docked poses are: AD4 (green), Vina (white), and MOE (cyan).

from the reference pose in 2HS1.

In the docking of compound **2** to the Elbow site of 2HS1, the best-scored pose generated by AD4, Vina, and MOE all have same orientation as the reference pose by *Chang et al.* (Figure 3.3 A). Furthermore, the benzoxazole scaffold and the thiophene moieties overlap with the corresponding moieties in the reference pose. The AD4 and Vina poses are structurally similar as the RMSD of the two structures is relatively small, ~ 2.1 Å. This is not unexpected since AD4 and Vina share some of the ideas and approaches of the docking-pose generation algorithms. Although all three docked poses are close in proximity, the RMSDs of these two docked poses are larger when compared to the MOE-generated docked pose (~ 10 Å), due to both translational/rotational and conformational sampling about the amide bonds originated from the docking-pose generation algorithms used in MOE docking.

When docking of compound **2** to the Elbow site of a closed form HIV-1p, 1PRO, the relative position of the docking results are mostly within the volume of the reference pose, but these results are more divergent from one another (Figure 3.3 B). The orientation of the MOE-generated docked ligand is flipped when compared to that of AD4 and Vina poses. Furthermore, the RMSD difference of AD4 and Vina poses is significant (> 15 Å), largely due to a major shift in the position of the Vina pose.

The Elbow site of semi-open form HIV-1p in 1HHP is structurally different from that of the closed forms of HIV-1p such as 2HS1 and 1PRO. The semi-open conformation alters the conformation of the flap loops that make up the Elbow site. The loops have been shown to be anti-correlated to the flap[4,9] and the opening of the flaps induces a narrower Elbow site. The docking results are appropriate given this structural

relationship. The most favorable poses are less consistent and have conformations that place the ligand outside of the volume of the reference pose (Figure 3.3 C). In addition, the docked poses from all three docking software have reversed orientation when compared to the reference pose. These results suggest the Elbow site in semi-open form and closed form HIV-1p has significant conformational changes in the local structure of the site. Thus, direct comparison of the docked poses generated by standard docking methods may not be ideal.

The scores of the docked poses are listed in Table 3.1. In the docking to 1HHP, the docked poses of compound **2** to the Eye site were consistently scored better than the 1HHP Elbow site. This is likely because the inhibitor fits more poorly in this Elbow site which is narrower than the open Elbow site of 1PRO and 2HS1. For docking of compound **2** to the Elbow sites, the results of the scoring vary significantly among the docking software and the binding sites. Also, the docking scores make no attempt to include the free energy change of the protein in different conformations because these terms are not known. All poses in the open Elbow sites of 2HS1 and 1PRO have estimated binding free energies of ~ 7.3 ± 1.8 kcal/mol. It is very difficult to choose a single pose as the "answer". Furthermore, all Elbow sites are structurally different, and the result from docking to only one Elbow site, as Chang *et al.* did, is insufficient to conclude the binding mode of a ligand. Therefore, with only a few static receptor structures for sampling, molecular docking alone is insufficient to predict which site, Elbow or Eye, is the preferred binding site for compound **2**.

Table 3.1. Docking scores of 2-protease complex

	Closed HIV-1p (2HS1)		Closed HIV-1p (1PRO)		Semi-open HIV-1p (1HHP)	
	Elbow 1	Elbow 2	Elbow 1	Elbow 2	Elbow	Eye
AD4	-7.3	-9.2	-6.6	-8.4	-5.8	-7.6
AD Vina	-7.1	-7.0	-5.5	-7.1	-5.3	-6.5
MOE	-7.3	-7.7	-7.5	-7.1	-5.8	-7.3
Average	-7.2	-7.9	-6.5	-7.5	-5.6	-7.1
	± 0.2	± 1.2	± 1.0	± 0.8	± 0.3	± 0.6
			-7.3 ± 1.8			

Protein Dynamics

If we were to focus solely on 1HHP, binding to the Eye site would clearly be the most favorable pose. However, the local variation of a static receptor structure in 2HS1 and 1PRO can strongly effect the docking results. Therefore, the dynamic nature of the ligand-receptor interaction should be considered for this problem. To understand the interactions between compound **2** and the Elbow or Eye site, 10 independent 10-ns Langevin Dynamics (LD) simulations were performed for each unique complex. The best docked pose of the ligand with AD4 was used as the initial coordinates of the ligand. The system with compound **2** docked to the Elbow site of 1PRO was named **2**-Elbow[1PRO], whereas the systems with compound **2** docked to the Elbow and Eye sites of 1HHP were named **2**-Elbow[1HHP] and **2**-Eye[1HHP], respectively.

In **2**-Eye[1HHP] LD simulations, the presence of compound **2** in the Eye site did not adversely affect the overall stability of the **2**-Eye[1HHP] protein (Figure 3.4 A). On the other hand, the conformations of the flaps appear to be affected by the presence of compound **2** in the Eye site, in which the ligand interacts with the flap directly. Throughout the trajectories of **2**-Eye[1HHP] simulations, compound **2** usually remained in the Eye site
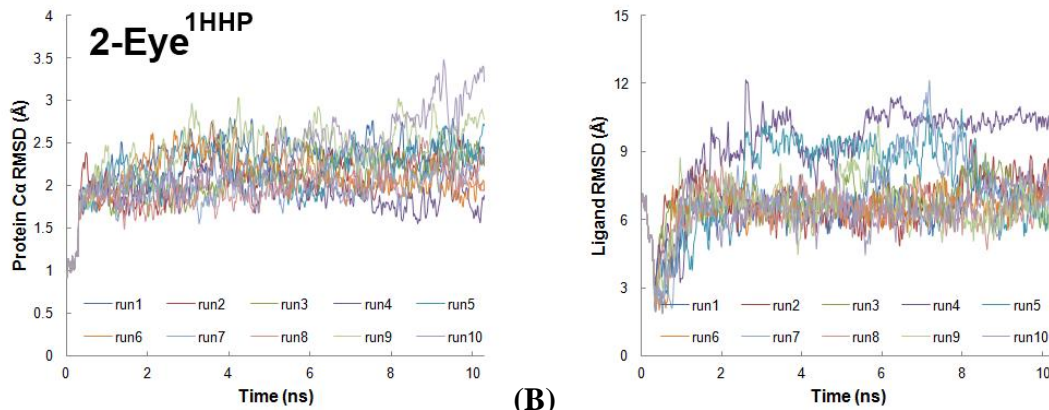
(A)

(B)

Figure 3.4. Structural stability of **2**-protease complexes in Langevin dynamics. Compound **2** was located in the Eye site of HIV-1p in semi-open form (1HHP). **(A)** The figure illustrates the RMSD of protein C_α during the simulations relative to its starting position. **(B)** The figure shows the RMSD of compound **2** from its starting position in the Eye site. In general, compound **2** has exited the binding site when it has RMSD > 8 Å.

(Figure 3.4 B) and interacted extensively with the flap region, occasionally with the flap tip of the opposing unit. Event of compound **2** exiting the Eye site was also observed; ligand RMSD > 8 Å in Figure 3.4 B. It may also drift into the active site and remained there, similar to the orthosteric inhibitors, e.g. pepstatin A.

In **2**-Elbow[1PRO] and **2**-Elbow[1HHP] simulations, we observed that the structure of the protease remained largely stable throughout the simulations (Figure 3.5). While compound **2** usually remained in the Elbow site, it was observed to "slide" inside the Elbow groove during the simulations, which generated large RMSD in some of the trajectories. Significantly, the conformation of HIV-1p in **2**-Elbow[1PRO], which began the simulation with closed conformation, relaxed and spontaneously converted from closed to semi-open form. In 8 of the 10 runs, the handedness of the flaps reversed and the flap openness, measured by the flap-active site distance (D25 C_α - I50 C_α), increased from closed form ~ 14.0 Å (average of 2HZ1 and 1PRO measurements) to semi-open form ~

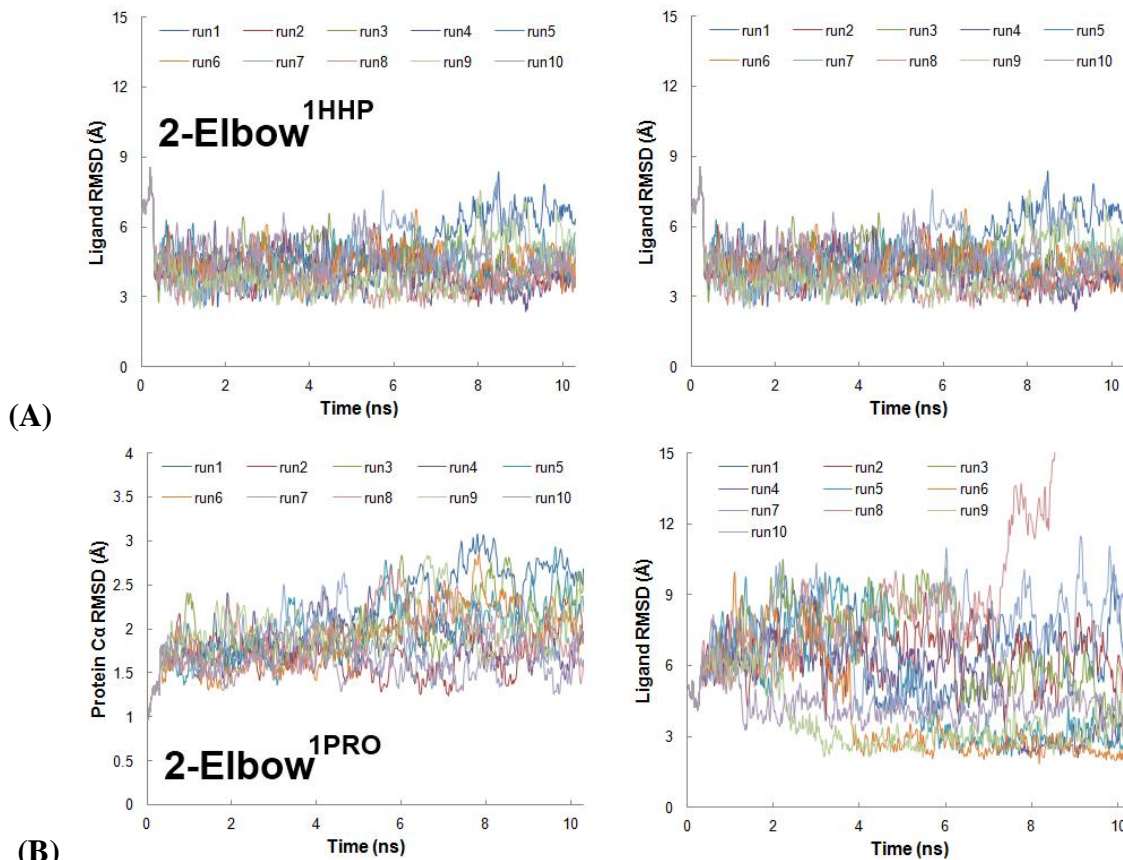**Figure 3.5.** Structural stability of **2**-protease complexes in LD. **(A)** Compound 2 in the Elbow site of HIV-1p in semi-open form (1HHP). **(B)** Compound **2** in the Elbow site of HIV-1p in closed form (1PRO). The left panel illustrates the RMSD of protein Cα during the simulations to its starting position. The right panel shows the RMSD of compound **2** to its starting position in the binding site. In general, compound **2** has exited the binding site when it has RMSD > 8 Å.

20.0 ± 3.7 Å, slightly beyond the 17.2 Å distance observed in the semi-open 1HHP. HIV-1p in **2**-Elbow^{1HHP} and **2**-Eye^{1HHP} mostly remained in semi-open conformation. **2**-Elbow^{1HHP} had a flap openness ~ 20.8 ± 3.2 Å, similar to the flap openness of the relaxed **2**-Elbow^{1PRO} (~ 20.0 ± 3.7 Å). The flap openness of **2**-Eye^{1HHP} was slightly smaller than **2**-Elbow^{1HHP} and **2**-Elbow^{1PRO}, ~ 17.0 ± 3.2 Å. Presumably this narrower flap openness is due to the presence of compound **2** in the Eye site, which interacts with the flap region and may thus limit the range of motion that the flap can achieve.

Perryman *et al.*4 proposed that the structure of the Elbow region may influence the flap motion due to the anti-correlation relationship of the Elbow gap distance (G40 C_α − Q61 C_α distance) and the flap-active-site distance. Based on this hypothesis, Chang *et al.* reasoned that the binding of compound **2** to the Elbow site may allosterically control the flap movement and thus inhibit the catalytic activity of HIV-1p. In our simulations, we did not observe strong correlation between the two distances. The Elbow gap distance of **2**-Elbow[1PRO], **2**-Elbow[1HHP], and **2**-Eye[1HHP] did not have a noticeable difference, and all have an Elbow gap distance ~ 7.2 ± 1.6 Å. Notably, ~ 33% of the 2-Elbow[1PRO] trajectories have conformations with Elbow gap distance ~ 9.8 ± 1.6 Å, a distance close to the Elbow gap distance observed in the Elbow site of closed form HIV-1p (2HS1, 9.1 / 12.4 Å and 1PRO, 10.2 / 10.8 Å).

Taken together, the results of flap-active-site distance and the Elbow gap distance suggest that the presence of compound **2** in the Elbow site did not alter the natural conversion from a closed to semi-open/open conformation in HIV-1p and did not have significant influence on the flap/Elbow movement. Conversely, compound **2** in the Eye site appeared to interact extensively with the flap and alter their openness. To better characterize the interactions between compound **2** and HIV-1p, the partial binding free energy of the **2**-protease complexes was estimated by MM/PBSA, in which the entropy term T∆S was not included due to the complexity of the normal mode analysis. Since compound **2** dissociated or drifted away from Elbow or Eye site in some of the trajectories, only frames with the appropriate ligand-bound state were selected for the free energy calculation (Table 3.2).

Table 3.2. MM/PBSA partial binding free energy estimations of 2-protease complexes

	No. of Frames	Median Free Energy (kcal/mol)
2-Eye[1HHP]	2480	-22.4 ± 2.6
2-Elbow[1HHP]	4425	-18.3 ± 1.1
2-Elbow[1PRO]	3100	-19.0 ± 2.4

2-Eye[1HHP] has an estimated binding free energy of -22.4 ± 2.6 kcal/mol, which is more favorable than both **2**-Elbow[1PRO] (-19.0 ± 2.4 kcal/mol) and **2**-Elbow[1HHP] (-18.3 ± 1.1 kcal/mol). There is a ~ 3.0 kcal/mol difference in the estimated binding free energy, suggesting that compound **2** would favored the Eye site over the Elbow site. However, the error bars are large for the MM/PBSA method, and we cannot completely rule out the Elbow site. **2**-Elbow[1PRO] and **2**-Elbow[1HHP] have similar estimated binding free energy. This supports the observation that the protease in **2**-Elbow[1PRO] underwent conversion from closed to semi-open/open conformation, in which **2**-Elbow[1PRO] would adopt conformations similar to **2**-Elbow[1HHP] and thus should have ligand-protein interaction equivalent to **2**-Elbow[1HHP] complex. In addition, since 2-Elbow[1PRO] adopted two conformations with different Elbow gap distances (~ 66% at ~ 7.2 ± 1.6 Å and ~ 34% at 9.8 ± 1.6 Å), the estimated binding free energy has a wider standard deviation than that of **2**-Elbow[1HHP].

3.5 Conclusion

The discovery of a new class of allosteric inhibitor of HIV-1p is a significant advancement in AIDS research. Chang *et al.* utilized enzyme kinetics methods to deduce the non-competitive nature of the new inhibitor and used molecular docking method to

hypothesize the binding mode of these compounds. In this study, we suggested an alternative binding site for the inhibitor, the Eye site. This allosteric site directly influences the motion of the flaps, an important feature of HIV-1p that is linked to substrate recognition and hydrolysis. We supported this alternative binding site through extensive molecular docking and protein dynamics studies. We found that the ligand interacts with the Eye site favorably and alters the protein's conformational behavior of the flaps. However, the inhibitor does not appear to significantly modulate the protein dynamics through the Elbow site.

3.6 Acknowledgements

3.7 References

1. Chang, M. W.; Giffin, M. J.; Muller, R.; Savage, J.; Lin, Y. C.; Hong, S.; Jin, W.; Whitby, L. R.; Elder, J. H.; Boger, D. L.; Torbett, B. E. Identification of Broad-Based HIV-1 Protease Inhibitors from Combinatorial Libraries. *Biochem. J.* **2010**, *429*, 527-532.

2. Seelmeier, S.; Schmidt, H.; Turk, V.; von der Helm, K. Human Immunodeficiency Virus Has an Aspartic-Type Protease that Can Be Inhibited by Pepstatin A. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 6612-6616.

3. Dunn, B. M.; Gustchina, A.; Wlodawer, A.; Kay, J. Subsite Preferences of Retroviral Proteinases. *Methods Enzymol.* **1994**, *241*, 254-278.

4. Perryman, A. L.; Lin, J.-H.; McCammon, J. A. Restrained Molecular Dynamics Simulations of HIV-1 Protease: the First Step in Validating a New Target for Drug Design. *Biopolymers* **2006**, *82*, 272-284.

5. Perryman, A. L.; Zhang, Q.; Soutter, H. H.; Rosenfeld, R.; McRee, D. E.; Olsen, A. J.; Elder, J. E.; Stout, C. D. Fragment-Based Screen Against HIV Protease. *Chem. Biol. Drug Des.* **2010**, *75*, 257-268.

6. Kovalevsky, A. Y.; Liu, F.; Leshchenko, S.; Ghosh, A. K.; Louis, J. M.; Harrison, R. W.; Weber, I. T. Ultra-High Resolution Crystal Structure of HIV-1 Protease Mutant Reveals Two Binding Sites for Clinical Inhibitor TMC114. *J. Mol. Biol.* **2006**, *13*, 161-173.

7. Perryman, A. L.; Lin, J.-H.; McCammon, J. A. HIV-1 Protease Molecular Dyanamics of a Wild-Type and of the V82F/I84V Mutant: Possible Contributions to Drug Resistance and a Potential New Target site for Drugs. *Protein Sci.* **2004**, *3*, 1108-1123.

8. Tozzini, V.; McCammon, J. A. A Coarse Grained Model of the Dynamics of Flap Opening in HIV-1 Protease. *Chem. Phys. Lett.* **2005**, *413*, 123-128.

9. Tozzini, V.; Trylska, J.; Chang, C,-e; McCammon, J. A. Flap Opening Dynamics in HIV-1 Protease Explored with a Coarse-Grained Model. *J. Struct. Biol.* **2007**, *157*, 606-615.

10. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins* **2006,** *65*, 712-725.

11. Case, D.A.; Darden, T.A.; Cheatham, III, T.E.; Simmerling, C.L.; Wang, J.; Duke, R.E.; Luo, R.; Crowley, M.; Walker, R.C.; Zhang, W.; Merz, K.M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvary, I.; Wong, K.F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S.R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D.H.; Seetin, M.G.; Sagui, C.; Babin, V.; Kollman, P.A. *AMBER 10*, University of California, San Francisco.: **2008**.

12. Spinelli, S.; Liu, Q. Z.; Alzari, P. M.; Hirel, P. H.; Poljak, R. J. The Three-Dimensional Structure of the Aspartyl Protease from the HIV-1 Isolate BRU. *Biochimie* **1991**, *73*, 1391-1396.

13. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235-242.

14. Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olsen, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30*, 2785-2791.

15. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A. Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157-1174.

16. Jakalian, A; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comp. Chem.* **2000**, *21*, 132-146.

17. Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comp. Chem.* **2002**, *23*, 1623-1641.

18. Wang, J.; Wang, W.; Kollman P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247-260.

19. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of *n*-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327-341.

20. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical Treatment of Salvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127-6129.

21. Onufriev, A.; Bashford, D.; Case, C. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins* **2004**, *55*, 383-394.

22. Sham, H. L.; Zhao, C.; Stewart, K. D.; Betebenner, D. A.; Lin, S.; Park, C. H.; Kong, X.-P.; Rosenbrook, W.; Herrin, T.; Madigan, D.; Vasavanonda, S.; Lyons, N.; Molla, A.; Saldivar, A.; Marsh, K. C.; McDonald, E.; Wideburg, N. E.; Denissen, J. F.; Robins, T.; Kempf, D. J.; Plattner, J. J.; Norbeck, D. W. A Novel, Picomolar Inhibitor of Human Immunodeficiency Virus Type 1 Protease. *J. Med. Chem.* **1996**, *39*, 392-397.

23. Trott, O.; Olson, A. J. Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461.

24. Molecular Operating Environment (MOE), version 2010.10; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, **2010**.

25. Damm, K. L.; Ung, P. M.-U.; Quintero, J. J.; Gestwicki, J. E.; Carlson, H. A. A Poke in the Eye: Inhibiting HIV-1 Protease through its Flap-Recognition Pocket. *Biopolymers* **2008**, *89*, 643-652.

CHAPTER 4

Identifying Binding Hot Spots on Protein Surface by Mixed-Solvent MD:

HIV-1 Protease as a Test Case

4.1 Abstract

Mixed-solvent molecular dynamics (MixMD) simulations use full protein flexibility and allow competition among water molecules and small organic probes to achieve accurate hot-spot mapping on the protein surface. In this study, we performed MixMD simulations with the homodimeric HIV-1 protease. The simulations used three probe-water solutions (acetonitrile-water, isopropanol-water, and pyrimidine-water), first at 50% w/w concentration and later at 5% v/v. MixMD of 5% v/v organic solutions have superior signal-to-noise ratio and far fewer spurious hot spots than 50% w/w MixMD. Furthermore, very intense and well defined organic probe occupancies were observed in only a handful of sites on the protein surface. The Eye site, an allosteric site underneath the Gly^{49}Ile^{50}Gly^{51} flap of HIV-1 protease that has been confirmed by the presence of a 5-nitroindole fragment in a crystal structure, was successfully mapped by all probes. Two additional, solvent-exposed, allosteric sites were also mapped by MixMD: the Exo site

139

(between the Gly^{16}Gly^{17} and Cys^{67}Gly^{68} loops) and the Face site (between Glu^{21}Ala^{22} and Val^{84}Ile^{85} loops). The Exo site was observed to overlap with crystallographic fragments such as acetate and DMSO. Analysis of crystal structures of HIV-1 protease in different symmetry groups confirmed the two surface, allosteric sites are common surfaces for crystal contacts.

4.2 Introduction

As one of the cornerstones in pharmaceutical research, structure-based drug design (SBDD) utilizes high-quality X-ray crystal structures of proteins as template for rational drug design. SBDD has lead to successful FDA-approved pharmaceutics.[1-3] A crucial step in SBDD is the identification of the potential sites, on the target protein for selective, high-affinity ligand binding. Active sites are straight forward to identify, but allosteric binding sites and the protein-protein interfaces are more difficult. This is especially crucial for new protein targets with little prior knowledge of the potential binding sites. Furthermore, identification of allosteric sites may provide novel mechanisms for protein modulation with pharmaceutical applications.

Binding sites are generally characterized by small pockets, or binding hot spots, on the protein surface that have high propensity for ligand binding.[4-7] These hot spots are typically lined by solvent-exposed, hydrophobic amino acid residues. Such composition has a profound impact on the properties of the protein surface; it allows for more favorable van der Waals interactions with hydrophobic molecules over polar water

140

molecules, while loosely bound water molecules on the hydrophobic protein surface can also be displaced with minimal entropic penalty. Hence, binding hot spots tend to attract and bind organic fragments or molecules more favorably than water molecules through a combination of enthalptic and entroptic contributions, allowing them to effectively compete against the bulk solvent (~ 55 Molar of water).

Two powerful experimental approaches were developed to validate the presence of binding hot spots on protein surface that bind small organic fragments: the X-ray crystallography-based multiple-solvent crystal structure (MSCS) method[8-12] and fragment binding detected by nuclear magnetic resonance ("SAR by NMR").[13,14] Both methods successfully mapped the binding hot spots of the target proteins with small organic solvents (acetone, acetonitrile, isopropanol, DMSO, etc) and correctly identified these hot spots as the substrate binding site of the target proteins.

Although the experimental methods are very powerful, limitations exist that render them not applicable to all protein systems. NMR studies are limited to small proteins. Some proteins may not crystallize at all, or the protein crystal may become unstable when soaked in an organic solvent. To avoid these physical restrictions, several computational methods have been developed to locate binding hot spots using static X-ray crystal structures.[15-20] These methods have had varying degrees of success and further development is needed to overcome limitations. In particular, numerous local free energy minima are common on the protein surface, and this is difficult to identify the true hot spots that would actually be observed experimentally. Such limitations may stem from the structural uncertainty in the crystals or the lack of full protein dynamics in the computational study. However, the most significant factor that limits these methods is the

inability to incorporate solvation effects and the probe-water competition at the protein surface.

To enhance binding hot-spot sampling, methods that model probe-protein interactions in dynamic simulations have been developed.[21-26] In general, these methods perform molecular dynamics simulations of the target protein in a box of aqueous solution of small organic probes and identify the binding hot spots on the protein surface that are frequented by probes for an extended period of time. The MacKerell group has developed the site-identification by ligand competitive saturation (SILCS) method that simulates the dynamics of target proteins in a benzene/propane/water mixture to generate a map of binding hot spots,[22-24] where binding free energy is estimated from the binding propensities of the probes.[25] However, SILCS requires the use of artificial repulsive interactions to avoid aggregation of the highly hydrophobic probes. Seco *et al.*[21] simulated proteins in an isopropanol/water solvent box and estimated the binding affinity, which was achieved by quantifying the clustering of probes as the ratio of observed probe density to the expected probe density. The ratio was then converted to an estimated binding free energy through an inverse Boltzmann relationship. Careful evaluation of the atomic contributions to binding free energy was attempted to avoid redundant inclusion of interdependent interactions, which may limit the accuracy of the method as it can lead to overestimation of binding affinity. Similar to the approach used by Seco *et al.*, Bakan *et al.*[25] assessed the druggability of the protein surface by estimating the probe binding affinities. Instead of using a single probe type, a mixture of organic fragments that are commonly found in approved drugs (isopropanol, isopropylamine, acetate, and acetamide) was used. The proportion of these fragments was estimated based on their

frequency of occurrence in drug molecules.[27] However, this approach relies on the estimation of probe density and has not been validated by comparison to MSCS, an experimental counterpart to all solvent-probing methods.

To overcome the shortcoming of these methods, the mixed-solvent molecular dynamics (MixMD) was developed[28] (Lexa, in press). MixMD utilizes a solution of water and miscible organic molecules (acetonitrile, isopropanol, and pyrimidine) to probe hot spots. This avoids the artifacts of highly hydrophobic organic molecules that require additional interaction terms to avoid aggregation or formation of separated solvent layers. The locations of the probes were converted into probe-occupancy maps that can then be examined similar to electron density maps. The accuracy of MixMD has been confirmed by direct comparison between the MixMD and MSCS density, the experimentally counterpart of all hot-spot probing methods. In our first study, hen egg-white lysozyme (HEWL)[10] was simulated in a 50% w/w acetonitrile aqueous solution. MixMD successfully reproduced the acetonitrile electron density found in the substrate binding pocket of HEWL in the MSCS, and it identified this site as the only significant binding hot spot[28] when full protein flexibility was allowed. In addition, "SAR by NMR" agrees with the MixMD and MSCS results in which the binding site is the only significant binding site for various solvent probes.[13] In follow-up work, acetonitrile and isopropanol were used as probes against several proteins: HEWL, elastase, p53 core, RNase A, subtilisin, and thermolysin (Lexa & Carlson, in press). Various choices for protocol were examined and the computed occupancy maps were compared directly to MSCS electron density to choose the best options for reproducing the correct experimental hot spots.

There was concern that the 50/50 solvent concentration we used in the original MixMD setup may disrupt the structure of proteins with no disulfide bonds[24] to help stabilize the protein fold.[29] In this paper, we will demonstrate that 50/50 MixMD method does not disrupt the HIV-1 protease (HIV-1p) dimer on the timescale examined here. Furthermore, we will extend and refine the current MixMD method to utilize 5% v/v organic probes in place of the 50% w/w. We chose to introduce a smaller concentration of organic solvent because it could be directly verified experimentally by NMR and X-ray crystallography, as 5% v/v organic solvent is usually not detrimental to the stability of the protein structure or crystal integrity. We were pleasantly surprised to find that the lower concentration also significantly reduced the presence of spurious minima on the surface of the protein.

HIV-1p was chosen as a test case because it is a clinically important target that is well-studied and has ample amount of experimental and crystallographic information. It is a C_2-symmetric, homodimeric protein that contains two monomeric units held together by non-covalent interactions at the dimer interface. This aspartyl protease is crucial in the lifecycle of HIV-1 as it recognizes and cleaves the *gag* and *gag-pol* polyprotein precursors and thus releases viral proteins that are essential in the assembly of viable viral particles.[30,31]

The access to the substrate binding site of HIV-1p is controlled by a pair of flexible glycine-rich β-hairpin, or flaps (K45-M-I-G-G-I-G-G-F-I54), one from each monomeric HIV-1p.[32-36] Crystal structures of HIV-1p in semi-open conformation have the flaps elevated from the orthosteric site when compared to those in closed conformation.[37,38] In addition, the flaps have different handedness in semi-open and

closed conformations. In closed conformation, the tip of the flap (I50) docks to the flap-tip recognition pocket located on the opposite monomer, also known as the Eye site. This conformation masks the Eye site, and thus, it has not been widely reported. In A semi-open conformation, the flaps elevate and the handedness of the flaps is reversed, thus opening up the Eye site for binding. Damm *et al.*[39] pioneered the Eye site as a novel allosteric site of HIV-1p and developed an Eye site pharmacophore model that helped identify small molecules targeting this site. Importantly, the Eye site has been confirmed by a fragment-soaked crystal structure of apo HIV-1p in semi-open conformation,[40] in which the fragment 5-nitroindole was resolved and found overlapping with the pharmacophore model of the Eye site.

In addition to the orthosteric, substrate-binding site and the allosteric Eye site, an additional allosteric site, the Exo site, has been described by Perryman *et al.*[40] Crystallographic evidence indicates that it is a shallow cleft formed by the elbow/cantilever/fulcrum components of the protease (consist of residues Gly16, Leu63, Ile64, Glu65, Lys70, and Ala71). Computational studies have shown that the width of this site is anti-correlated to the openness of the flap.[41,42] Experiments show that constraining the movement of the Exo site with disulfide cross-linking diminishes the HIV-1p catalytic activity.[43] Additionally, Perryman *et al.*[40] also proposed a Flap site (consists of residues Trp42, Lys43, Pro44, Lys45, Met46, Lys55, Val56, Arg57) on the solvent-exposing side of the flap structure based on crystallographic observation.

As HIV-1p has a wealth of binding sites with different functions and experimental information for verification, it is well suited for use as a test case for MixMD validation and development. Three solvent probes are used, and the solvent at different

concentrations is studied for its effect on protein structure integrity and probing efficiency and accuracy. In particular, consensus of hot spots is required. More than one probe must map the same hot-spot for us to consider it relevant.

4.3 Computational Methods

Identification of binding hot spots was accomplished by first performing a set of independent MixMD simulations of HIV-1p solvated in an organic-aqueous solution. Then, the end of the trajectories were combined, and the distribution of organic probes was calculated as an occupancy grid. The results were then visualized as probe occupancy maps on the protein surface, similar to viewing an electron density map. The occupancy positions with strong intensity were identified as potential binding hot spots and compared to crystal structures of HIV-1p for validation.

Mixed-Solvent Molecular Dynamics Simulations

The crystal structures of the semi-open form (PDB: 1HHP[37]) and closed form (PDB: 1PRO[44]) of HIV-1p were used in the simulations, in which the ligand in 1PRO was removed. Hydrogens were added to the protein with *tLeAP*, a module of AMBER11,[45] and the protein was parameterized with FF99SB force field.[46] SHAKE[47] was applied to restrain all bonds to hydrogen atoms. A 10-Å cutoff for van der Waals and Particle Mesh Ewald[48] for electrostatics were used. Cl⁻ ions were added to neutralize the

system charge. A 2-fs time step was used. Temperature was regulated through an Andersen thermostat.[49]

Amber parameters for acetonitrile (ACN) were used.[50] Parameters for isopropanol (IPA) and pyrimidine (1P3) were based on the OPLS-AA parameters for pure alcohols[51] and heterocycles,[52] respectively. This choice was based on an in-depth exploration of available solvent parameters (Lexa, in preparation). For 50% w/w organic solvent, the protein was solvated in an 18-Å, pre-equilibrated box of organic solvent and TIP3P water.[53] For 5% v/v organic solvent, the protein was set up in a layered manner, which solvates the protein with a small shell of organic solvent followed by solvation with a large box of pure TIP3P water. Control of solvent concentration is achieved through adjusting the ratio of organic probe and water molecules.

The MixMD system underwent 250 cycles of steepest-descent minimization followed by 4750 cycles of conjugate-gradient minimization with the protein fixed. Each system was gradually heated from 10K to 300K over 80 ps while the protein was restrained by a harmonic force constant of 10 kcal/mol·Å^2. Restraints on protein heavy atoms were gradually removed over 350 ps while the temperature was maintained at 300 K. This was followed by unrestrained equilibration for 1.4 ns. Five independent, 20-ns production runs of MixMD were performed using the GPU-enabled *pmemd*, a module of AMBER 11.[45,54]

Essential Dynamics

Principle component analysis (PCA) of the MD trajectory data, also called Essential Dynamics[55,56] (ED), was used to monitor and compare dynamics of the protein

structure in MixMD simulations. The routine *ptraj* was used to perform the matrix calculation on backbone heavy atoms and generate the eigenvalues and the associated eigenvectors of the MD trajectories.[57] Vectors of each residue were superimposed onto the C_α atom of the residue and the results were visualized using VMD in the form of a porcupine plot.[58,59] We sought to quantitatively compare eigenvectors from different MD trajectories with a dot product of the two eigenvectors. The vector on each of the *n* number of C_α atoms in an eigenvector was compared to the corresponding vector in another eigenvector, and the dot-product of these two component vectors was calculated. The median of dot-products on the vectors of the *n* number of C_α atoms was used to describe the global similarity of the compared eigenvectors.

Solvent Occupancy Maps

The results of MixMD simulations were analyzed using *ptraj*, for which the last 10 ns of each of the 5 individual runs were combined for solvent occupancy analysis. The trajectory was fitted to the non-flap region of the reference crystal structure by C_α RMS. The *ptraj* "grid" command, with a 0.5 Å x 0.5 Å x 0.5 Å spacing over the entire volume, was used to calculate the occupancies of water and organic probe at each grid point. To effectively compare the occupancy maps from different sources, the maps were normalized by converting the raw data into standard score (Z-score) with the equation

$$z_i = \frac{x_i - \mu}{\sigma}$$

where x_i is the occupancy at a grid point i, μ and σ are the mean and standard deviation of occupancy of all grid points, respectively. The normalized solvent occupancy maps were visualized with PyMOL,[60] similar to viewing electron density maps. The contour levels

of probe occupancy represent the distance between the raw occupancy at the grid points and the mean occupancy in units of the standard deviation (σ). Hence, these are the same as viewing contours of electron densities.

4.4 Results and Discussion

MixMD with High-Concentration Organic Solutions

Validation of MixMD Using HIV-1p in Semi-open Form

A 50% w/w of organic-aqueous solution was used for the initial MixMD simulations. To generate pre-mixed, organic-aqueous solutions used in the MixMD simulation of HIV-1p, the 1HHP structure was solvated in a small box of organic probes with TIP3P water at the appropriate molecular ratio (Table 4.1) was employed.

Table 4.1. Ratio of Organic Solvent and Water in 50% w/w Organic-Aqueous Solutions

	Acetonitrile[1] : water[2]	Isopropanol[3] : water	Pyrimidine[4] : water
Weight : Weight	50 : 50	50 : 50	50 : 50
Mol : Mol	1.220 : 2.775	0.832 : 2.775	0.624 : 2.775
Molecule : Molecule	2,766 : 6,113	829 : 2,779	622 : 2,751

[1] Acetonitrile: 41.05 g/mol; [2] Water: 18.02 g/mol
[3] Isopropanol: 60.10 g/mol; [4] Pyrimidine: 80.09 g/mol

Using the pre-mixed 50% w/w organic solutions, five independent 20-ns trajectories of MixMD of apo HIV-1p were generated, which totaled 250 ns for each organic solution system. To confirm that the solution in the HIV-1p MixMD remained well mixed throughout the simulations, the radial distribution functions of the solutions in all resultant trajectories were calculated at 10-ns intervals. All intervals of MixMD

ACN-H$_2$O
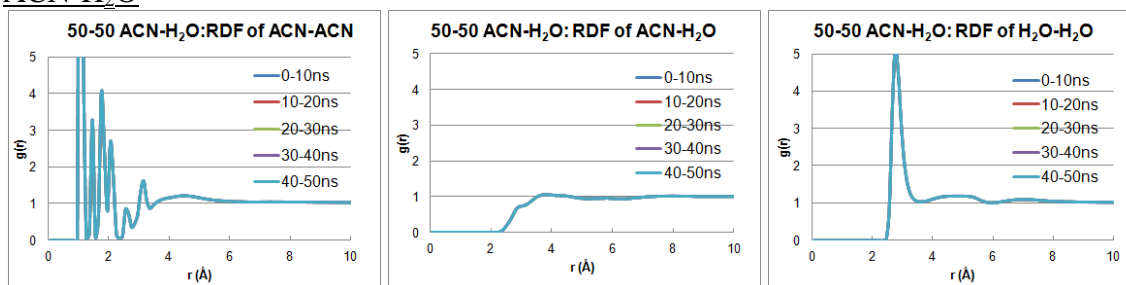
IPA-H$_2$O
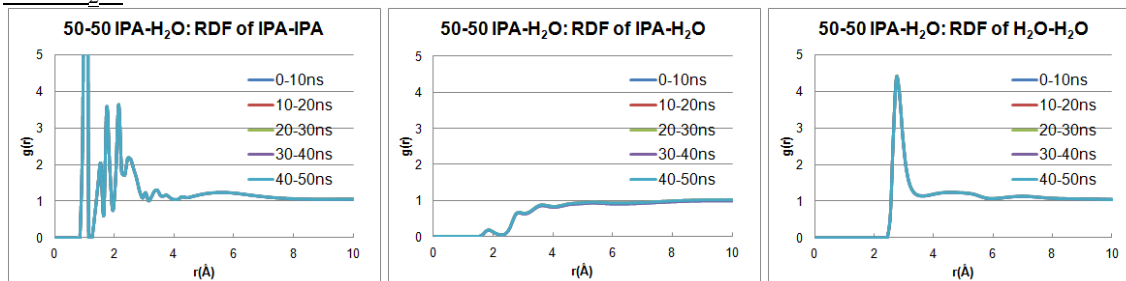
1P3-H$_2$O
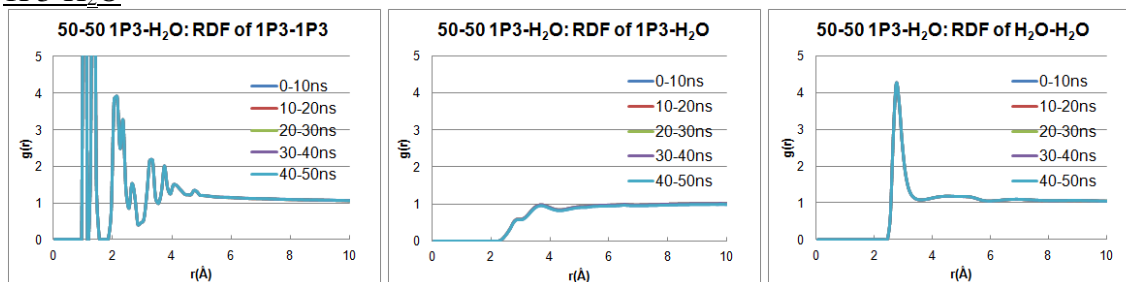
Figure 4.1. Radial distribution functions of 50% w/w probe-water MixMD at different time intervals. For ACN, IPA, and 1P3, radial distribution functions of all MixMD at different time intervals are virtually identical and converge to unity.

trajectories in all three 50% w/w organic solutions were confirmed to have appropriately converged to unity, indicating that even mixing of solvents was achieved (Figure 4.1).

There were concerns that unrestrained MixMD simulation of a protein without structure-stabilizing disulfide bridges may be susceptible to denaturation due to the introduction of small organic probes.[24] To ensure that the effects of the explicit organic solution are not detrimental to the structural dynamics of the protein, global and local analysis of the apo HIV-1p structure were performed. C$_\alpha$ RMSD was calculated for the core of the dimer (all of HIV-1p except residues 45-54 in each monomer), and it shows

150

that the protein remains stable throughout the MixMD simulations. The core RMSD of 50% w/w ACN-water, IPA-water, and 1P3-water MixMD simulations are 1.7 ± 0.3 Å, 1.6 ± 0.3 Å, and 1.8 ± 0.7 Å, respectively, similar to that (1.5 ± 0.3 Å) of a 40-ns pure-water MD simulation of HIV-1p. At a local level, the flaps of HIV-1p always maintain the handedness in semi-open conformation. Measured by the Asp25-Ile50 C_α distance, the flap-openness of 50% w/w ACN-water (19.2 ± 3.9 Å), IPA-water (14.7 ± 2.0 Å), and 1P3-water (16.1 ± 2.7 Å) MixMD are similar or larger than the flap-openness (15.2 ± 3.8 Å) of HIV-1p in a pure-water MD simulation. These values are closer to the openness in semi-open conformation (~17.2 Å in PDB:1HHP) than to the openness in closed conformation (~12.6 Å in PDB:1PRO). Two of the 1P3-water MixMD simulations experienced flap opening event, hence the larger core RMSD (Figure 4.2), whereas the IPA-water MixMD trajectories experienced closing down of the flaps while maintaining the semi-open flap handedness, hence the shorter Asp25-Ile50 C_α distance seen with IPA-water MixMD.

In addition to using the simple structure analyses mentioned above to examine the protein dynamics, the collective motions of the protein can be elucidate through ED.[55,56] ED performs the principal axis transformation of the covariance matrix of structural fluctuations of the protein and extracts the eigenvectors as the collective motions with the largest contribution to the variance of the atomic fluctuations. These collective motions are expected to correspond to large-scale atomic displacements that may be important to protein functions. Herein, we applied ED to analyze the MD trajectory of each MixMD system, and the resultant eigenvectors were compared to the ED of apo HIV-1p generated from a 40-ns pure-water MD simulation.
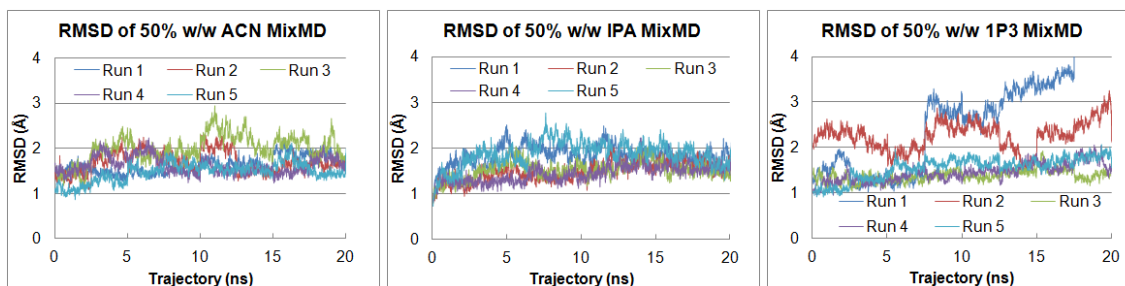
Figure 4.2. RMSD of the core of the HIV-1p dimer in 50% w/w MixMD. The trajectories are stable throughout the 20-ns simulation, with the exception of one opening event in 1P3. A similar opening event has been observed in pure-water simulations.[46]

To quantitatively compare the structural movements of two trajectories, we developed a global similarity factor to describe the degree of overlap between the compared ED eigenvectors. This factor is equivalent to the median of the dot-products of two eigenvectors, which was generated by calculating the dot-product of the vectors on each of the C_α atoms in the compared eigenvectors. Through comparison, we observed that the first several eigenvectors of all MixMD systems had good correlation (global similarity factor > 0.5) to the corresponding eigenvectors of apo HIV-1p in pure-water MD (Table 4.2, Figure 4.3). A full comparison of all low modes in the pure-water simulation was made to all low modes of the mixed-solvent simulations to ensure that the values in Table 4.2 are indeed the optimal comparison. Furthermore, these eigenvectors often encompassed the most significant dynamic motions of the protein,[55] which primarily described the opening-closing and shearing motions of the flaps; the first 3 modes of ED have cumulative eigenvalues that usually describe > 70% of all protein motions (Table 4.3). The ED comparison strongly suggests that the application of 50% w/w organic-aqueous solutions in MixMD does not adversely affect the general protein dynamics.

Table 4.2. Global similarity factor of ED eigenvectors from HIV-1p in pure-water MD and 50% w/w MixMD

50% w/w MixMD	Pure-Water MD				
	$1^{st} - 1^{st}$	$2^{nd} - 2^{nd}$	$3^{rd} - 3^{rd}$	$4^{th} - 4^{th}$	$5^{th} - 5^{th}$
ACN-water	0.734	0.516	0.605	0.353	0.487
IPA-water	0.556	0.786	0.557	0.518	0.111
1P3-water	0.760	0.180	0.600	0.053	0.056

Table 4.3. Normalized eigenvalues of the first five modes of ED of pure-water MD and 50% w/w MixMD

Eigenvector	Normalized Eigenvalue (Cumulative Percentage)			
	Pure Water	ACN 50%	IPA50%	1P3 50%
1	0.54 (54%)	0.29 (29%)	0.39 (39%)	0.54 (54%)
2	0.14 (68%)	0.18 (47%)	0.21 (60%)	0.14 (68%)
3	0.08 (75%)	0.14 (60%)	0.11 (71%)	0.09 (78%)
4	0.06 (81%)	0.08 (68%)	0.06 (77%)	0.04 (82%)
5	0.04 (86%)	0.07 (75%)	0.05 (82%)	0.04 (86%)
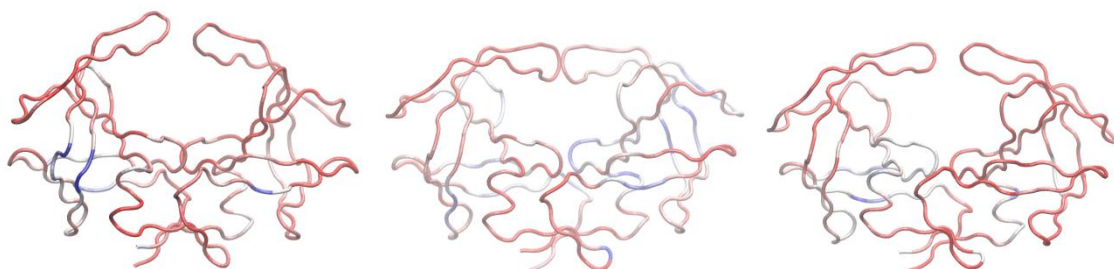
(A) 50% w/w ACN-Water   **(B)** 50% w/w IPA-Water   **(C)** 50% w/w 1P3-Water

Figure 4.3. Dot-product of the 1st essential dynamics eigenvectors of apo HIV-1p in pure water and in 50% w/w organic-aqueous solution MixMD. **(A)** 1st mode of ED from ACN-water MixMD and from pure-water MD have a global similarity factor of 0.734, **(B)** while IPA-water MixMD and pure-water MD is 0.556, and **(C)** 1P3-water MixMD and pure-water MD is 0.760. The illustrated HIV-1p is the averaged structure of 5 independent trajectories of MixMD. Color Red indicates the vectors are correlated and have positive dot-product. Color Blue indicates the vectors are anti-correlated and have negative dot-product. Color White indicates the vectors have dot-product equals 0, i.e. no correlation.

Identification of Binding Hot Spots

MixMD uses hydrophobic solvent molecules to identify sites on protein surface that have high propensity for ligand binding. The frequency of the hydrophobic probes occupying to a site would be an indication of the affinity for binding. To visualize the sites with high binding frequency, probe occupancy maps were calculated to show a consensus representation of the probe positions throughout the MixMD trajectories. As is appropriate for the technique, only the final 20 ns of each of the five independent, 20-ns simulations was used in the MixMD analysis. The probe-occupancy map was generated by integrating the occurrence of organic probe molecules within the grid space surrounding the protein throughout the trajectories. To enable quantitative comparison of the probe-occupancy maps, the occupancy of the map was normalized by converting the raw occupancy into a standard score (Z-score), in which the intensity of the occupancy is represented as the number of standard deviations (σ) above the basal level of occupancy. Hence, the intensity of the probe occupancy can be visualized as a surface of contour level that represents the minimum σ value of probe occupancy within the enclosed volume. To focus upon the locations that were most frequently sampled by the probes (i.e., the hydrophobic binding hot spots), the contour level was adjusted so that only the probe occupancies with strong intensity would remain.

In the MixMD of HIV-1p in 50% w/w ACN-water, IPA-water, and 1P3-water solutions, we observed numerous weak probe occupancies along the protein surface that disappeared as the contour threshold was increased from $\sigma = 5$ to 8. Among the intense probe occupancies that remained, several patches of protein surfaces were consistently mapped by all three organic probes, and these volumes were named the Exo, the Eye, and

the Face sites (shown in Figure 4.4). At the same contour level, IPA and 1P3 appear to have larger occupancy volume than ACN within the Eye site (Figure 4.4). Such observation may be due to the differences in size and shape of the probe molecules, in which the linear and compact ACN is smaller in size compared to the relatively globular IPA and the flat, disk-shaped 1P3. Furthermore, we observed that IPA has two small spherical occupancies within the Eye site whereas 1P3 has one flat and elongated occupancy within the Eye site. This difference may result from the conformational space sampled by the flaps during the 50% w/w MixMD simulations (Figure 4.2). HIV-1p in IPA-water solution sampled conformations similar to the closed form while maintaining the handedness of a semi-open form, which may reshape the Eye site and result in the two observed hot spots. In 1P3-water solution, the flaps sampled the open conformation and may result in an elongated Eye site, thus producing a flat and elongated hot spot in the Eye site.

Weak Probe Occupancies

Regardless of the choice of probe in the 50% w/w organic solution MixMD, at low probe occupancy contour levels, e.g. $\sigma = 5$, there are significant numbers of tiny probe occupancies, or "noise," all over the protein surface in the probe occupancy maps, as seen in Figure 4.4, similar to results observed in other methods.[21-23] Although many of the weak occupancies can be eliminated simply by increasing the occupancy contour threshold, e.g. $\sigma = 8$, they may misguide or distract researchers from studying the genuine binding hot spots on the protein. These irrelevant probe occupancies may hinder the identification of genuine binding hot spots on the protein surface. Such spurious probe

occupancies may result from the use of excessive quantity of organic probe in the MixMD; excess probe molecules may linger and occupy the local free energy minima on the protein surface because the genuine binding hot spots are already occupied by other probes. Another drawback of the 50% w/w MixMD method is the narrow margin that differentiates intense and weak probe occupancies. As observed in Figure 4.4, most of the potential binding hot spots are a few σ units from the basal level of occupancy (< 8). Increasing the probe occupancy contour level by as few as 3 to 4 σ units effectively eliminated most of the probe occupancies on the protein surface. This narrow range of margin renders the distinction of intense occupancies, i.e. plausible binding hot spots, from the spurious occupancies more subjective than we desire.

MixMD with Low-Concentration Organic Solutions

One of the pitfalls of the current MixMD simulation method is the use of a large quantity of organic probe. At a ratio of 50% w/w to water, it is not experimentally appropriate because many proteins are known to be sensitive to additives in the buffers, e.g., salts, surfactants, buffering agents and organic solvents. For example, HIV-1p is known to be sensitive to organic solvent added to the assay buffer; proteolytic activity diminishes when DMSO concentration is greater than 5%.[61] To circumvent the issues that can arise from using high concentration of organic probe to solvate the protein for MixMD simulations, a lower concentration of organic solution MixMD was pursued.
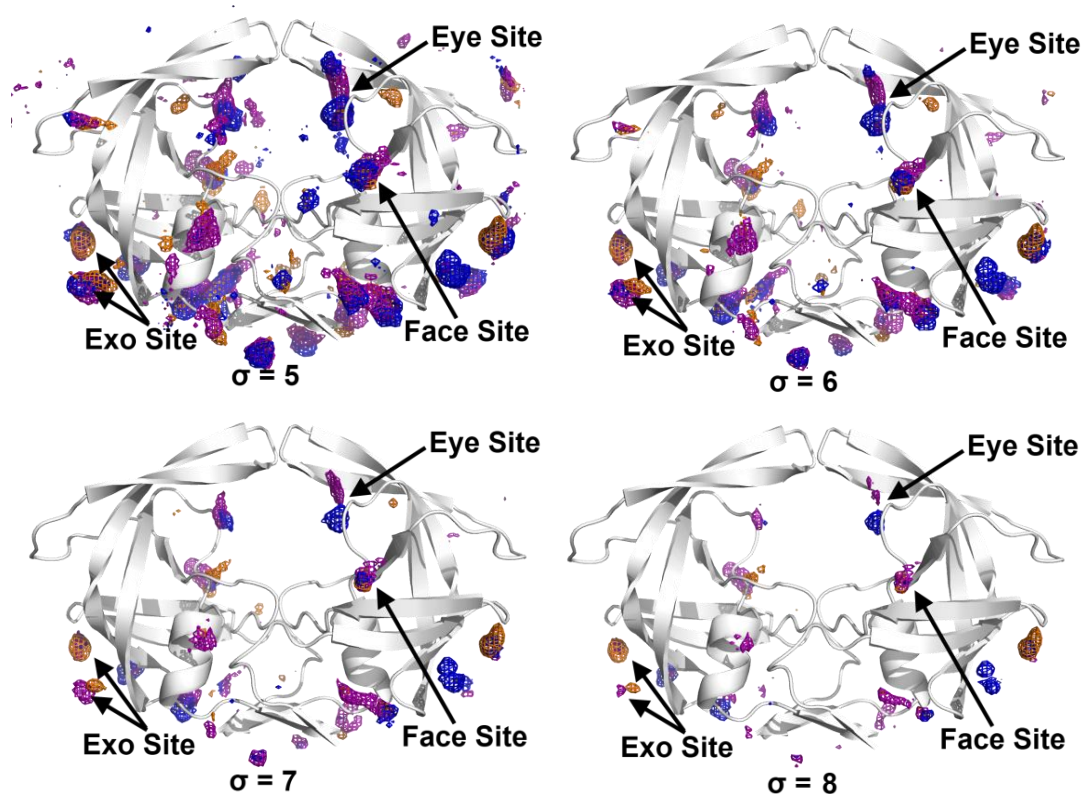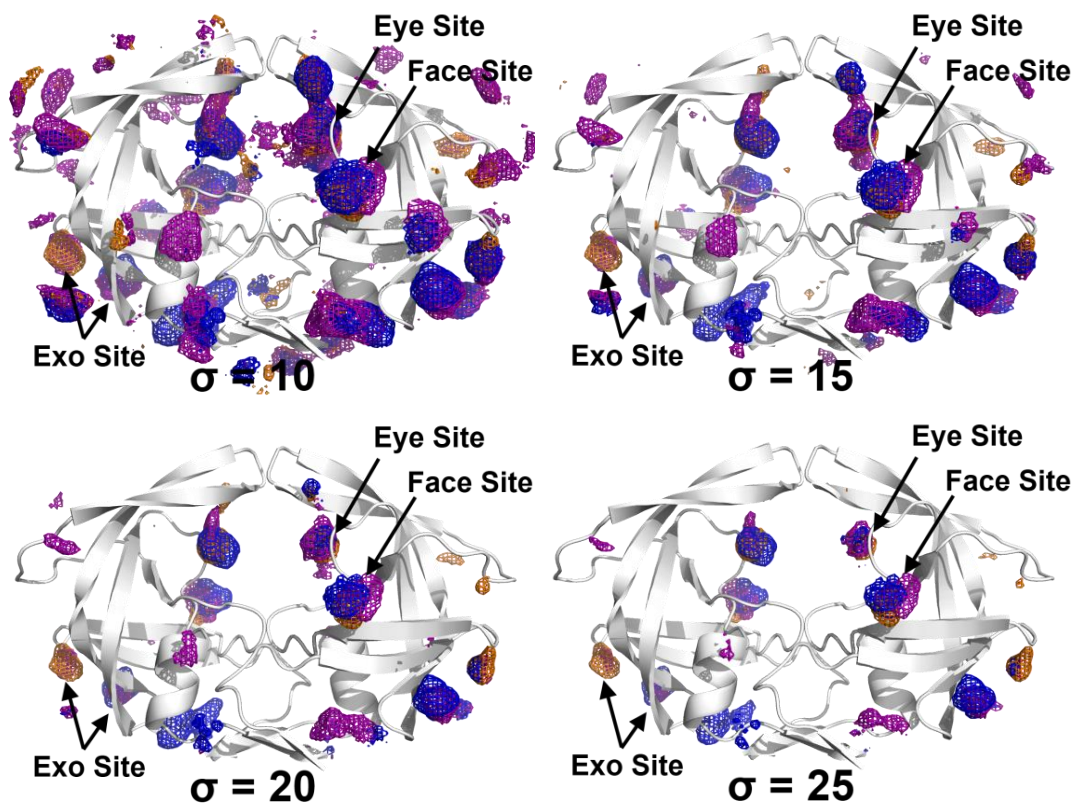
Figure 4.4. Probe occupancies of 50% w/w organic solution in MixMD of HIV-1p. ACN, IPA, and 1P3 probe occupancies are colored in orange, blue, and purple, respectively. By normalizing the probe occupancy of the MixMD, the intensity of the occupancy can be quantified by the number of standard deviation value (σ) above the basal level of occupancy. This σ value is represented as the occupancy contour level. Increasing the contour level from $\sigma = 5$ to 8 eliminates weakly occupied spaces. All three organic probes have consistent overlapping occupancy in the Exo, Eye, and Face sites.

Three 5% v/v solution boxes, ACN-water, IPA-water and 1P3-water, with a probe-to-water ratio of 1:55.5 (Table 4.4), were created. The probe-water boxes were pre-mixed with heating and equilibrating at 300 K. RDF confirmed that all probe-water solutions were well mixed, i.e., converged to unity. To identify the binding hot spots of HIV-1p, 5% v/v probe-water MixMD simulations were performed, in which HIV-1p in semi-open conformation was immersed in the pre-mixed boxes. We found that a 5% v/v probe-water MixMD demonstrated significant advantages over the 50% w/w MixMD method. The most important finding was the same probe occupancies of 5% v/v probe-

water MixMD of HIV-1p in semi-open conformation identified the same intense binding hot spots as the orthosteric site, Exo site, the Eye site, and the Face site. However, the 5% v/v MixMD and 50% w/w MixMD differ vastly in the intensity of the probe occupancies. For the 50% w/w probe-water MixMD, a contour level of $\sigma = 10$ would effectively eliminate all probe occupancies, and the margin between intense and weak occupancies is only around 3 to 4 σ values. In contrast, probe occupancy results of 5% v/v probe-water MixMD has a wide intensity range, in which the most intense binding hot spots would remain at 30 σ values above the basal level of occupancy. This extended range of occupancy intensity permits the separation of intense and weak probe occupancies with a wider margin than 50% w/w probe-water MixMD can achieve. Hence, 5% v/v MixMD provides a better signal-to-noise ratio and clearer resolution of binding hot spots (Figure 4.5). Typically, most weak occupancies were eliminated at the occupancy contour level $\sigma = 15$. At the contour level $\sigma = 25$, only the most intense binding hot spots remained: the Exo site, the Eye site, and the Face site. Hence, the 5% v/v MixMD method should be applied in the majority of MixMD studies, and an occupancy contour level of $\sigma \sim 25$ should be applied as a standard cutoff.

Table 4.4. Ratio of organic solvent and water molecule in 5% v/v organic-water solutions

	ACN : water[2]	IPA : water	1P3[4] : water
Volume : Volume	5 : 95	5 : 95	5 : 95
Weight : Weight	3.10 : 95	5.08 : 95	3.93 : 95
Mol : mol	1.0 : 55.5	1.0 : 83.2	1.0 : 80.7
Molecule : Molecule	276 : 15,315	184 : 15,308	190 : 15,333

[1] ACN: 41.05 g/mol; [2] Water: 18.02 g/mol
[3] IPA: 60.10 g/mol; [4] 1P3: 80.09 g/mol

Application of Layered-Solvent Method

5% v/v probe-water MixMD has demonstrated significantly improved signal-to-noise ratio and resolution of probe occupancy over the 50% w/w MixMD method. However, the probe molecules were very diluted in the MixMD system, i.e., ~ 200 probe molecules were used in the ~ 15,000 water molecules MixMD and, thus, may have limited sampling during the MixMD simulations. The probe molecules may spend too much time in the vast volume of the bulk solvent and not enough time on the protein surface during MixMD. To circumvent this issue, we altered the solvation strategy by solvating the protein in a **layered** format: a shell of organic probe molecules was first added to the protein, then a large box of pure water molecules was added to complete the protein-solution system. In this case, the water molecules would compete off the weakly bound probe molecules on the protein surface that have low affinity to the probes. This leaves only the probes that remain in the binding hot spots to stay interacting with the protein for longer period of time to identify the positions of binding hot spots as regions of the surface that have strong probe occupancy.

To confirm that performing MixMD with layered solvation method did not introduce artifacts into the sampling of protein motion and probe distribution, we performed 5 independent runs of both pre-mixed and layered MixMD of HIV-1p with 5% v/v ACN-water solution for assessment. RDFs of both pre-mixed and layered MixMD simulation converged to unity by the completion of equilibrations and remained well mixed for the duration of all production runs. Thus, the layered solvation method does not alter general probe distribution and is suitable for use in MixMD simulation. We also performed ED analysis on the protein from the pre-mixed and the layered MixMD

Figure 4.5. Probe occupancies of 5% v/v organic solution in MixMD of HIV-1p. ACN, IPA, and 1P3 probe occupancies are colored in orange, blue, and purple, respectively. By normalizing the probe occupancy of the MixMD, the intensity of the occupancy can be quantified by the number of standard deviation value (σ) above the basal level of occupancy. Increasing the contour level from $\sigma = 10$ to 25 eliminates weakly occupied spaces. All three organic probes have consistent overlapping occupancy in the Exo, Eye, and Face sites.

simulations and found that the two sets of MixMD are highly correlated. The global similarity factor of the 1st eigenvector of the pre-mixed and layered MixMD has a high value of 0.971 (Table 4.5), demonstrating that the protein dynamics were nearly identical in both pre-mixed and the layered MixMD (Figure 4.6 A).

The probe occupancies of 5% v/v ACN-water MixMD are basically the same for pre-mixed and layered solvation methods. Both methods have $\sigma = 25$ contours at the same locations with roughly the same volumes. Both identify the Exo, Eye, and Face sites as the most occupied sites and have a significant reduction of weak occupancies

when compared to the 50% w/w MixMD (Figure 4.6 B). Furthermore, the intensity of the probe occupancy of 5% v/v MixMD is usually very concentrated on a few binding hot spots. This allows the isolation of the most significant hot spots and generates a much "cleaner" probe occupancy map (Figure 4.6 B).

Table 4.5. Top five ED global similarity factors of 5% v/v ACN MixMD simulations of HIV-1p

	Eigenvector – Eigenvector				
mixed– apo	1 – 1	1 – 2	2 – 3	3 – 2	4 – 4
	0.721	0.702	0.744	0.726	0.744
layered – apo	1 – 1	1 – 2	2 – 1	2 – 4	3 – 5
	0.789	0.748	0.633	0.681	0.704
layered – mixed	1 – 1	1 – 3	3 – 2	4 – 3	5 – 2
	0.971	0.683	0.710	0.853	0.619

(A)   **(B)**   σ = 25
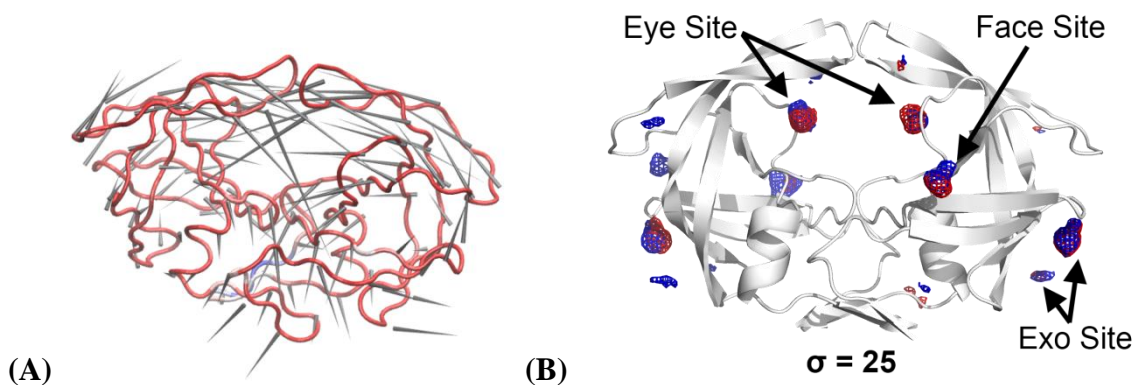
Figure 4.6. Comparison of pre-mixed and layered MixMD of HIV-1p in 5% v/v ACN-water solution. **(A)** Porcupine plot of the dot-product of the 1st eigenvectors of pre-mixed and layered MixMD. The arrow indicates the sum of the vector of movement on the residue. The degree of correlation of the two eigenvectors on the residue is visualized with color, where red color indicates the vectors are correlated and blue color indicates anti-correlated while white color indicates the vectors have no correlation. **(B)** When ACN probe occupancies (mesh) are shown at a high occupancy contour level (σ = 25), only the Exo, Eye, and Face sites are identified. Pre-mixed MixMD probe occupancy is colored in blue and that of the layered MixMD is colored in red.

HIV-1p Binding Hot Spots

It was important to compare the results of the layered simulations across all 3 probe solvents. For each of the 5% v/v ACN-water, IPA-water, and 1P3-water solution systems, five independent sets of 20-ns layered MixMD simulations were performed. The last 10 ns of each of the trajectories was used in the probe occupancy analysis. Using a high contour threshold ($\sigma > 20$), the Exo, Eye, and Face sites were consistently identified as the binding hot spots in all three 5% v/v probe-water solution systems. In addition, the probe occupancies of layered MixMD with 5% v/v ACN, IPA, or 1P3 were very comparable to the corresponding occupancy results from 50% w/w MixMD, but with fewer spurious sites dotted along the protein surface (Figure 4.7).

One of the binding hot spots consistently identified by the ACN, IPA, and 1P3 probes is the Eye site, the flap-tip recognition pocket. When HIV-1p is complexed with orthosteric site inhibitors and adopts a closed conformation, the flaps are closed down and occupy the Eye site, rendering this site unavailable for ligand binding. In contrast, the flaps of HIV-1p in semi-open conformation are open and have the Eye site available for

$\sigma = 25$ $\sigma = 25$ $\sigma = 25$
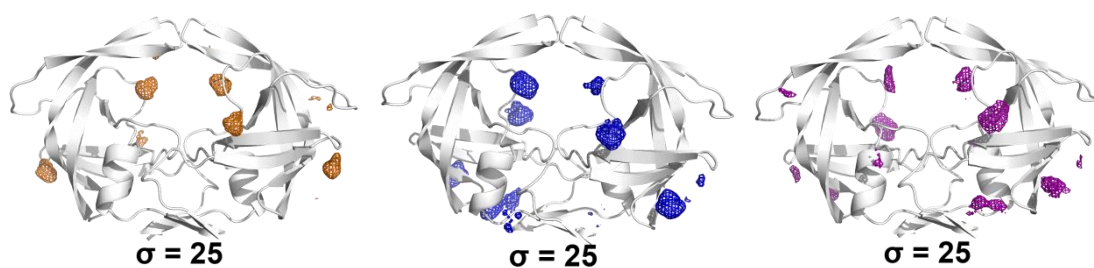
Figure 4.7. Probe occupancy maps of 5% v/v organic solution HIV-1p MixMD with layered solvation method. ACN, IPA, and 1P3 probe occupancies are colored in orange, blue, and purple, respectively. At a high probe occupancy contour level ($\sigma = 25$) all three organic probes have intense occupancies in the Exo, Eye, and Face sites. Very few small, non-specific occupancies are observed compared to the occupancy map of 50% w/w organic solution MixMD.

solvent probes. The Eye site is identified by MixMD, and the probe occupancies overlap with the crystallographic position of 5-nitroindole (Figure 4.8 A).[40]

ACN, IPA, and 1P3 in all MixMD simulations consistently map the Exo site, a long shallow cleft formed by the elbow/cantilever/fulcrum components of the protease and situated far away from the substrate binding site and the Eye site. In fact, MixMD produced two discrete hot spots within the Exo site separated by roughly 8.0 Å. Accordingly, the Exo site can be further divided into two subsites, Exo site 1 (near residues Gly16, Leu63 and Ala71) and Exo site 2 (near residues Ile64, Glu65 and Lys70). The existence of these two subsites is strongly supported by crystallographic evidences in crystal structures such as 3KFP or 3KFN.[40] Buffer additives (e.g., acetate and DMSO) are resolved and found to overlap with the positions of the proposed Exo subsites (Figure 4.8 B). In these crystals, the additives are usually found to occupy the Exo site as a pair. Unlike the Eye site, currently the Exo site has no known function; Perryman et al.[41] have proposed that it may function as an allosteric site that modulates the open-close conformational change of the flap.

Based on the HIV-1p co-crystals of small fragments such as PDB 3KFR and 3KFS, Perryman et al.[40] also proposed a ligand binding site on top/outside of the flap (herein called Flap site), which consists of residues Trp42, Lys43, Pro44, Lys45, Met46, Lys55, Val56, and Arg57. Our probe occupancy results of MixMD with 5% v/v and 50% w/w organic solutions indicate that ACN, IPA, and 1P3 have insignificant probe occupancies at the Flap site that Perryman et al. described. The probe occupancies at this site are very weak in both 5% v/v and 50% w/w MixMD. For 50% w/w MixMD, a low occupancy contour of $\sigma = 5$ would effectively eliminate the probe occupancies at the Flap

site. For 5% v/v MixMD, a contour above σ = 10 is sufficient to remove all probe occupancies observed at the Flap site along with some weak or spurious occupancies on other patches of protein surface (Figure 4.9). This observation suggests that, although MixMD can weakly map the Flap site, this site does not have the optimal physical and chemical properties to attract the binding of hydrophobic organic molecules at the same level as the binding hot spots in the orthosteric binding site and the Exo, Eye, and Face sites. Based on our MixMD result, the Flap site should be given a low priority as a potential binding site for drug-like molecules.
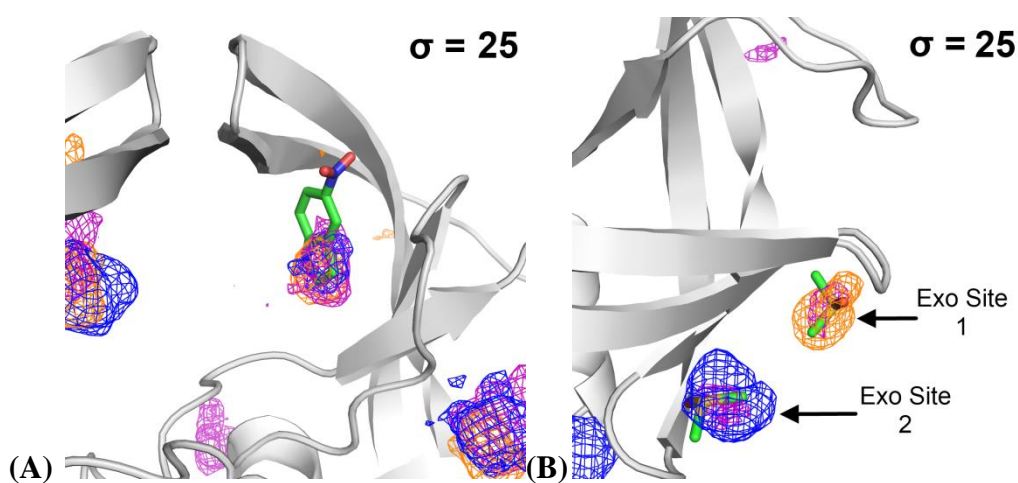
Figure 4.8. Crystallographic additives overlap with the Exo and Eye site probe occupancies. **(A)** Position of fragment 5-nitroindole overlaps with the Eye site mapped by various organic probes. **(B)** Resolved crystallographic buffer additives (PDB 3KFP; two DMSO shown) are found to perfectly overlap with the Exo site probe occupancies predicted by MixMD. Probe occupancy of 5% v/v ACN-water, IPA-water, and 1P3-water systems are in orange, blue, and purple meshes, respectively. All probe occupancies are shown at contour level σ = 25.
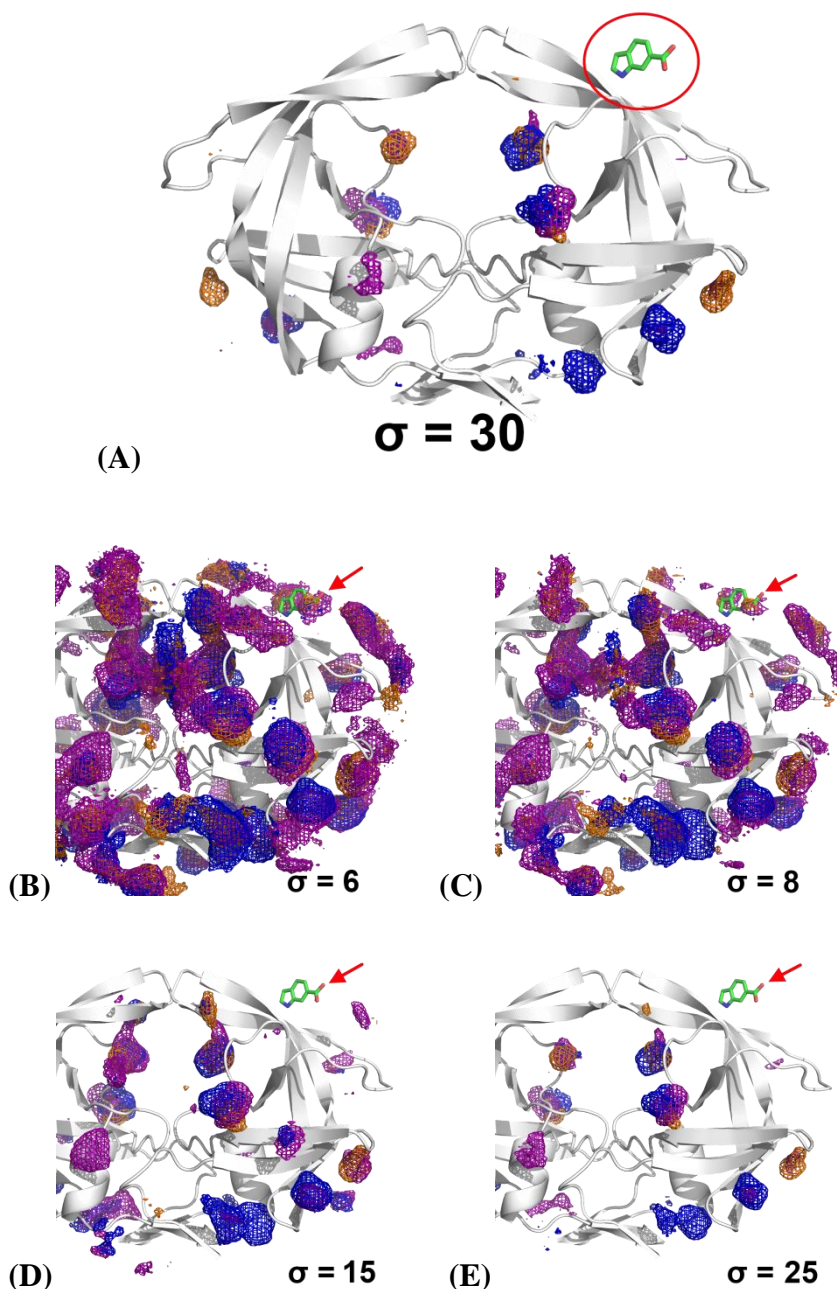
Figure 4.9. Probe occupancies near the Flap site. **(A)** Probe occupancy maps by 5% v/v MixMD with HIV-1p in semi-open conformation. At very high occupancy contour level ($\sigma = 30$), no observable probe occupancy at the Flap site by any of the probe when the occupancy contour levels were tuned to show the most intense occupancies. **(B-E)** The occupancy contour level was increased gradually ($\sigma = 6$, 8, 15 and 25) to eliminate the weak occupancies on the protein surface. The probe occupancies at the Flap site are weak, and as the contour level increases, the occupancies at the site disappear rapidly, indicating they are weak occupancies and should be disregarded. The position of the fragment indole-6-carboxylate was taken from PDB: 3KFR, as indicated by the red arrow. Probe occupancy of ACN-water, IPA-water, and 1P3-water systems are in orange, blue, and purple meshes, respectively.

165

HIV-1p in Closed Conformation

We observed that the substrate binding pockets of HIV-1p in the semi-open conformation were mapped in both 5% v/v and 50% w/w MixMD with ACN, IPA, and 1P3. However, the probe occupancies at these pockets are weak and are shifted in position. They also disappear along with other weak occupancies on the protein surface when the occupancy contour threshold is tuned ($\sigma > 15$) to select the most intense occupancies. The weak mapping of the known substrate binding pockets might result from the conformation of the flaps in semi-open conformation. The positions of the residues lining the substrate-binding pocket shift away when compared to the closed conformation. This creates a less compact and less enclosed binding site. This change in the compactness of the substrate binding pockets makes the pockets of HIV-1p in semi-open conformation provide less binding surface and less van der Waals interactions to attract the small molecule probes. To confirm that the weak probe occupancies in the substrate binding pocket were due to the semi-open flap conformations, five sets of 20-ns, layered, 5% v/v probe-water MixMD were performed for HIV-1p in the closed conformation. The 1PRO structure was used, with the inhibitor removed for the MixMD. The inhibitor amprenavir, found in the crystal structure 1HPV, was used as reference for the positions of the substrate binding pockets;[62] the center-of-mass of the 4 moieties of amprenavir (phenyl, 4-aminobenzenesulfonyl, isobutyl and tetrahydrofuran-3-yl) were marked as the center of the S1, S2, S1', and S2' binding pockets, respectively.

From the analysis of MixMD with HIV-1p in closed conformation, intense occupancies were observed for ACN, IPA, and 1P3 in the Exo and Face sites at a contour of $\sigma > 25$. No significant occupancy was observed at the Eye site. This is expected since

the Eye site is occupied by the flap tip of the opposite HIV-1p monomer in closed conformation, and thus, it is not open for small probes to bind. In contrast, significant probe occupancies can be observed in the S1/S1' and S2/S2' orthosteric pockets (Figure 4.10). It is interesting that the occupancies at these pockets are slightly weaker than those of the Exo and the Face sites. The orthosteric site can only be observed definitively at contour thresholds below $\sigma = 20$. This implies that the individual sub-pockets in the active site would not yield significant affinity for our probes, at least not as good as the Exo and Face sites. However, we know that the active site can strongly bind substrates and inhibitors. Therefore, MixMD does not inherently encode the potential synergism of binding multiple, nearby hot spots.

We checked for the presence of probes at the Flap site in our MixMD with closed HIV-1p. Perryman *et al.*[40] indentified the Flap site in crystal structures of HIV-1p in closed conformation. Similar to the results from MixMD with semi-open HIV-1p, only very weak probe occupancies were observed at the Flap site. These occupancies were less
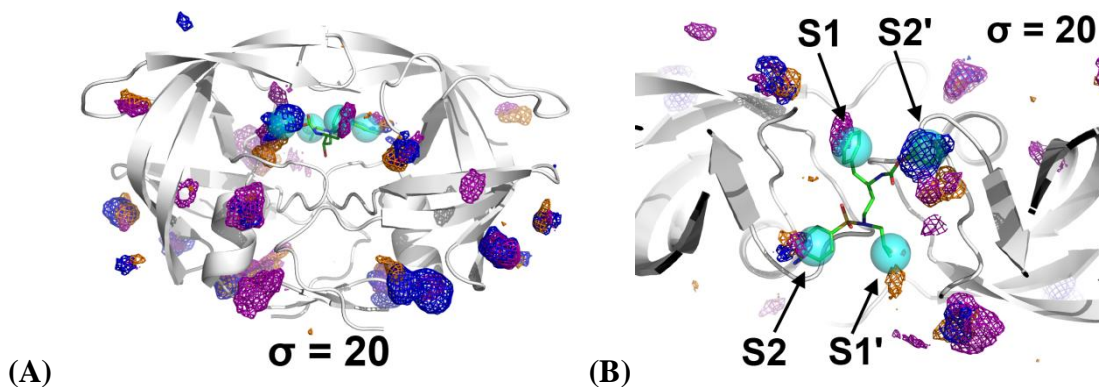
Figure 4.10. Probe occupancies at the substrate binding pockets of HIV-1 protease in closed conformation. **(A)** The overall probe occupancy maps by 5% v/v MixMD with the inhibitor amprenavir in the catalytic site (PDB: 1HPV). **(B)** The probe occupancies at the S1/S2 and S1'/S2' substrate binding pockets. The cyan spheres represent the positions of the hydrophobic pockets. Probe occupancy of ACN-water, IPA-water, and 1P3-water systems are in orange, blue, and purple meshes, respectively.

than 10 σ in these layered, 5% v/v simulations, indicating these occupancies are less significant when compared to the probe occupancies at the Exo, Face, and the active site.

Shallow Protein Surfaces and the Influence of Crystal-Packing Interfaces

In addition to the Eye sites and the active site occupancies observed in the MixMD of HIV-1p in semi-open and closed conformations, respectively, there are substantially strong probe occupancies on several patches of shallow protein surface. These are consistently mapped by ACN, IPA, and 1P3 probes used in the 50% w/w and 5% v/v MixMD simulations of HIV-1p. The most notable unknown hot spots are the Face site (a depression near the catalytic site of HIV-1p composed of residues Lys20, Glu21, Ala22, Leu23, Ile24, Thr80, Pro81, Ala82, Asn83, and Ile84) and the Exo site (a shallow cleft formed by the elbow/cantilever/fulcrum components of the protease, which consists of residues Gly16, Leu63, Ile64, Glu65, Lys70, and Ala71). Additionally, a Flap site (residues Trp42, Lys43, Pro44, Lys45, Met46, Lys55, Val56, and Arg57) is also proposed based on crystallographic evidence.[40]

These aforementioned sites are currently not known to have significance in the activity of HIV-1p, and they might first be regarded as spurious hot spots. We hypothesize these patches of protein surface might be sites that are able to form specific crystal contacts. MixMD basically identifies patches of protein surface that are easy to desolvate, easy for organic moieties to displace water. Hot spots for small molecules fit this definition, but so do crystal packing interfaces. That association inherently requires favorable protein-protein contacts. Such binding sites have been observed and explained by Liepinsh and Otting[14] when they compared the NMR result of HEWL in various

organic-aqueous solutions to the MSCS results by Mattos *et al.*,[9] where many crystallographically determined binding sites other than the substrate binding site, usually near the crystal contact regions, are absent in the NMR result. A similar observation has also been described by Mattos *et al.*[12] where MSCS of elastase resolved solvent probes outside the known binding pockets. They were mostly clustered at or near the protein surfaces that form crystal contacts. In order to understand the observation and the relevance of these surface sites that are consistently and strongly mapped by various organic probes during MixMD, we compared the crystal structures of HIV-1p available in the public domain to the occupancy maps.

From BindingMOAD[63] we obtained 347 crystal structures of ligand-bound HIVp and related proteases (SIV, FIV) that have a diffraction resolution better than 2.5 Å. Together with the unliganded apo structure used in this study, these structures were grouped into 17 crystal symmetry groups, and the symmetry units were generated and visualized through PyMOL. We found that several crystal symmetry groups have crystal contacts at or near the Exo, Face, and Flap sites (Figure 4.11). Such clustering of solvent molecules at or near the crystal-contact regions has been observed in the MCSC of HEWL[9] and elastase.[12] These experimental observations strongly support the legitimacy of our observation of ACN, IPA, and 1P3 occupancies at the Exo and the Face sites, both shallow protein surfaces that can form protein crystal contacts. This supports the accuracy of the novel MixMD method, though it points to the fact that some patches that are easy to desolvate may not have a regulatory influence in some systems. Fortunately, these sites are often relatively flat, not concave like true binding sites.
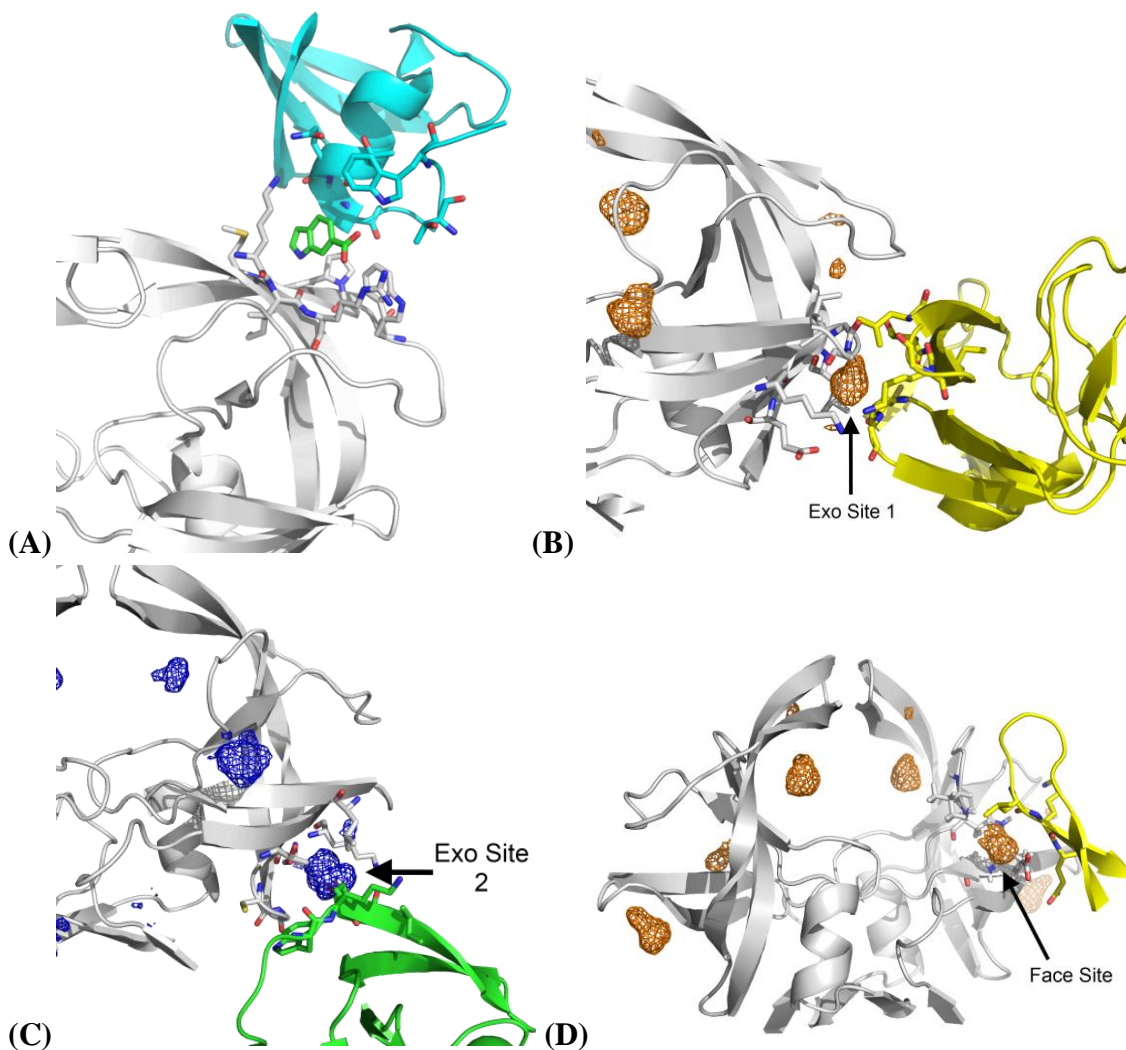
Figure 4.11. Probe occupancies near the Exo and the Face sites as potential sites for non-specific crystal contacts. **(A)** In 3KFR, a pseudo binding pocket forms between the crystal contacts of the Flap site and the symmetry unit (cyan) to accommodate the fragment indole-6-carboxylate. No intense MixMD probe occupancy was observed at this site. **(B)** In 2AZC, Exo site 1 forms crystal contact with the Exo site of its symmetry unit (yellow). Only ACN occupancy is shown. **(C)** In 1DAZ, Exo site 2 forms crystal contact with its symmetry unit (green). Probe occupancy can be found between the two units. For clarity, only IPA occupancy is shown. **(D)** In 2AZC, a small crystal contact formed between the Face site and its symmetry unit (yellow) is shown. Significant probe occupancy is seen at this site (only ACN occupancy is shown).

The Flap site is proposed based on the co-crystal structures of HIV-1p, 3KFR and

3KFS. MixMD was only able to weakly map this site, which may imply that the presence

of resolved fragments in the co-crystals may result from the formation of a pseudo-

binding pocket at the interface of the crystal contacts, which would otherwise not be present under normal conditions. In the crystal structures 3KFR and 3KFS, the Flap site of the asymmetric unit forms crystal contacts with an area near the dimer interface of its symmetry unit (residues Thr4, Trp6, Thr91, Gln92, Ile93 and Gly94), which explains the formation of a pseudo-binding pocket for the binding of fragments (Figure 4.11 A). These results strongly suggest that crystal structures should be studied carefully, especially the presence of small molecules at a shallow cleft on protein surface to avoid misinterpretation of the binding of small molecules to a false site.

The Exo site appears to be a favored interface for non-specific crystal contacts, as we found that crystal structures of HIV-1p in symmetry groups I-222 (e.g. 1ZBG), I-$4_1$22 (e.g. 2AZC), P-112_1 (e.g. 1A8K), P-$2_12_12_1$ (e.g. 1DAZ), P-6_1 (e.g. 1A8G), and P-$6_1$22 (e.g. 1UPJ) have crystal contacts at or near the Exo site. Specifically, we observed that in symmetry groups I-222, I-$4_1$22, P-112_1, P-$2_12_12_1$, P-6_1, and P-$6_1$22 one of the symmetry units is in close proximity to the Exo site 1 of the asymmetric unit[64] (Figure 4.11 B). Probe occupancies from various maps can be found clustered in the crystal contacts between the two units, similar to the observation described by Mattos *et al*.[12] On the other hand, symmetry groups P-112_1, P-$2_12_12_1$, and P-$6_1$22 have symmetry units in close proximity to the Exo site 2. In 1DAZ,[65] crystal contacts are formed between the Exo site 2 of the asymmetric unit and the flap of its symmetry unit (residue 44 to 54) (Figure 4.11 C). Probe occupancies from various maps can be found populated between the two contacting units. Similarly, we found that 2AZC, a crystal in the I-$4_1$22 symmetry group, has crystal contacts that overlap with the ACN, IPA, and 1P3 probe occupancies at the Face site (Figure 4.11 D): crystal contacts formed between the Face site of the

asymmetric unit and a beta-sheet structure (residues Gln18, Leu 19, Lys20, Glu21) of its symmetry unit.

Interestingly, the Face Site residues of HIV-1p are relatively conserved among similar aspartic proteases from other retroviruses. Using the sequence alignment tool CLUSTAL W v1.81[66] available through the SDSC Biology Workbench 3.2 online server[67] (**http://workbench.sdsc.edu**), the sequence similarity of the retroviral proteases was examined: the protease of the simian immunodeficiency virus (SIV, 50% identity to HIV-1p), HIV type-2 (48% identity to HIV-1p), equine infectious anemia virus (EIAV, 32% identity to HIV-1p) and feline immunodeficiency virus (FIV, 21% identity to HIV-1p). Several residues of the Face site, most of them hydrophobic, are relatively conserved among the examined aspartic proteases (Table 4.6). The relatively conserved constitution of hydrophobic residues at the Face Site and the tendency of this site to bind hydrophobic probe molecules indicates that this patch of protein surface may have an intrinsic functionality as a docking site for the client polyproteins. Association to this surface may provide some enthalpic compensation through van der Waals interactions and entropic assistance with the displacement of the loosely bound water molecules on this hydrophobic protein surface.

Table 4.6. Residue similarity at Face site of several viral aspartic proteases

	Identity (HIV1p)	Residues (HIV-1 Protease)									
		20	21	22	23	24	80	81	82	83	84
HIV-1	--	K	E	A	L	L	T	P	V	N	I
SIV	50%	V	E	V	L	L	T	P	I	N	I
HIV-2	48%	V	E	V	L	L	T	P	I	N	I
EIAV	32%	L	N	V	L	L	I	P	V	T	I
FIV	21%	I	K	F	L	L	V	C	V	L	L

4.5 Conclusion

In order to identify ligand binding pockets or protein-protein interfaces for rational drug discovery and design, computational techniques have been developed to map out the binding hot spots on protein surfaces. Many of them rely on a static crystal structure as a template and do not take into account of the effect of probe-water competition on protein surface. They usually suffer from the same shortcoming of many false hot spots identified on the protein surface, along with the true binding sites. To clearly find true binding hot spots and eliminate false positives, we have further refined the MixMD method. MixMD uses small, soluble organic probes with reasonable physical properties. This allows us to avoid the introduction of unphysical parameters to prevent the undesirable aggregation problem of very hydrophobic molecules, such as benzene and propane. We demonstrated that MixMD positively identifies the allosteric, hydrophobic Eye site that we previously proposed in semi-open HIV-1 protease, a site that is supported by crystallographic evidence with 5-nitroindole bound in HIV-1 protease. Interestingly, MixMD reproduces some interesting crystallographic results by identifying binding hot spots on the shallow protein surface that are known to form crystal contacts, which explains the observation of strong probe occupancy at locations not known for ligand binding.

4.6 Acknowledgements

4.7 References

1. Marrone, T. J.; Briggs, J. M.; McCammon, J. A. Structure-Based Drug Design: Computational Advances. *Annu. Rev. Pharmacol. Toxicol.* **1997**, *37*, 71-90.

2. Davis, A. M.; Teague, S. J.; Kleywegt, G. J. Application and Limitations of X-Ray Crystallography Data in Structure-Based Ligand and Drug Design. *Angew. Chem. Int. Ed.* **2003**, *42*, 2718-2736.

3. Hubbard, R. E. Structure-Based Drug Discovery: An Overview. Great Britain: *Royal Soc. of Chem.* **2006**.

4. Tsai, C.-J.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. Studies of Protein-Protein Interfaces: A Statistical Analysis of the Hydrophobic Effect. *Prot. Sci.* **1997**, *6*, 53-64.

5. Ringe, D.; Mattos, C. Analysis of the Binding Surfaces of Proteins. *Med. Res. Rev.* **1999**, *19*, 321-331.

6. Soga, S.; Shirai, H.; Kobori, M.; Hirayama, N. Use of Amino Acid Composition to Predict Ligand-Binding Sites. *J. Chem. Inf. Model.* **2007**, *47*, 400-406.

7. Soga, S.; Shirai, H.; Kobori, M.; Hirayama, N. Identification of the Druggable Concavity in Homology Models using the PLB Index. *J. Chem. Inf. Model.* **2007**, *47*, 2287-2292.

8. Allen, K. N.; Bellamacina, C. R.; Ding, X.; Jeffery, C. J.; Mattos, C.; Petsko, G. A.; Ringe, D. An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins. *J. Phys. Chem.* **1996**, *100*, 2605-2611.

9. Mattos, C.; Ringe, D. Locating and Characterizing Binding Sites on Proteins. *Nat. Biotech.* **1996**, *14*, 595-599.

10. Wang, Z.; Zhu, G.; Huang, Q.; Qian, M.; Shao, M.; Jia, Y.; Tang, Y. X-Ray studies on Cross-Linked Lysozyme Crystals in Acetonitrile-Water Mixture. *Biochim. Biophy. Acta* **1998**, *1384*, 335-344.

11. English, A. C.; Groom, C. R.; Hubbard, R. E. Experimental and Computational Mapping of the Binding Surface of a Crystalline Protein. *Prot. Eng.* **2001**, *14*, 47-59.

12. Mattos, C.; Bellamacina, C. R.; Peisach, E.; Pereira, A.; Vitkup, D.; Petsko, G. A.; Ringe, D. Multiple Solvent Crystal Structures: Probing Binding Sites, Plasticity and Hydration. *J. Mol. Biol.* **2006**, *357*, 1471-1482.

13. Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering High-Affinity Ligands for Protein: SAR by NMR. *Science* **1996**, *274*, 1531-1534.

14. Liepinsh, E.; Otting, G. Organic Solvents Identify Specific Ligand Binding Sites on Protein Surfaces. *Nat. Biotech.* **1997**, *15*, 264-268.

15. Damm, K. L.; Carlson, H. A. Exploring Experimental Sources of Multiple Protein Conformations in Structure-Based Drug Design. *J. Am. Chem. Soc.* **2007**, *129*, 8225-8235.

16. Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macro Molecules. *J. Med. Chem.* **1985**, *28*, 849-857.

17. Miranker, A.; Karplus, M. Functionality Maps of Binding Sites: a Multiple Copy Simultaneous Search Method (MCSS). *Proteins* **1991**, *11*, 29-34.

18. Dennis, S.; Kortvelyesi, T.; Vajda, S. Computational Mapping Identifies the Binding Sites of Organic Solvents on Proteins. *Proc. Nat. Am. Sci. U.S.A.* **2002**, *99*, 4290-4295.

19. Landon, M. R.; Amaro, R. E.; Baron, R.; Ngan, C. H.; Ozonoff, D.; McCammon, J. A.; Vajda, S. Novel Druggable Hot Spots in Avian Influenza Neuraminidase H5N1 Revealed by Computational Solvent Mapping of a Reduced and Representative Receptor Ensemble. *Chem. Biol. Drug Des.* **2008**, *71*, 106-116.

20. Brenke, R.; Kozakov, D.; Chuang, G.-Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. Fragment-Based Identification of Druggable 'Hot Spots' of Proteins using Fourier Domain Correlation Techniques. *Bioinformatics* **2009**, *25*, 621-627.

21. Seco, J.; Luque, F. J.; Barril, X. Binding Site Detection and Druggability Index from First Principles. *J. Med. Chem.* **2009,** *52*, 2363-2371.

22. Guvench, O.; MacKerell, A. D. Computational Fragment-Based Binding Site Identification by Ligand Competitive Saturation. *PLoS Comput. Biol.* **2009**, *5*, e1000435.

23. Raman, E. P.; Yu, W.; Guvench, O.; Mackerell, A. D. Reproducing Crystal Binding Modes of Ligand Functional Groups Using Site-Identification by Ligand Competitive Saturation (SILCS) Simulations. *J. Chem. Inf. Model.* **2011**, *51*, 877-896.

24. Foster, T. J.; Mackerell, A. D.; Guvench, O. Balancing Target Flexibility and Target Denaturation in Computational Fragment-Based Inhibitor Discovery. *J. Comp. Chem.* **2012**, *33*, 1880-1891.

25. Bakan, A.; Nevins, N.; Lakdawala, A. S.; Bahar, I. Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules. *J. Chem. Theory Comp.* **2012**, *8*, 2435-2447.

26. Raman, E. P.; Vanommeslaeghe, K.; MacKerell, A. D. Site-Specific Fragment Identification Guided by Single-Step Free Energy Perturbation. *J. Chem. Theory Comput.* **2012**, *8*, 3513-3525.

27. Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveum, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: A Comprehensive Resource for 'omics' Research on Drugs. *Nucleic Acids Res.* **2011**, *39*, D1035-D1041.

28. Lexa, K. W.; Carlson, H. A. Full Protein Flexibility is Essential for Proper Hot-Spot Mapping. *J. Am. Chem. Soc.* **2011**, *133*, 200-202.

29. Sevier, C. S.; Kaiser, C. A. Formation and Transfer of Disulphide Bonds in Living Cells. *Nat. Rev. Mol. Cellu. Biol.* **2002**, *3*, 836-847.

30. Kohl, N. E.; Emini, E. A.; Schleif, W. A.; Davis, L. J.; Heimbach, J. C.; Dixon, R. A.; Scolnick, E. M.; Sigal, I. S. Active Human Immunodeficiency Virus Protease is Required for Viral Infectivity. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 4686-4690.

31. Kräusslich, H.-G. Human Immunodeficiency Virus Proteinase Dimer as Component of the Viral Polyprotein Prevents Particle Assembly and Viral Infectivity. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 3213-3217.

32. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins* **2006,** *65*, 712-725.

33. Hornak, V.; Simmerling, C. Targeting Structural Flexibility in HIV-1 Protease Inhibitor Binding. *Drug Discov. Today* **2007**, *12*, 132-138.

34. Ishima, R.; Louis, J. M. A Diverse View of Protein Dynamics from NMR Studies of HIV-1 Protease Flaps. *Proteins* **2008**, *70*, 1408-1415.

35. Galiano, L.; Ding, F.; Veloro, A. M.; Blackburn, M. E.; Simmerling, C.; Fanucci, G. E. Drug Pressure Selected Mutations in HIV-1 Protease Alter Flap Conformations. *J. Am. Chem. Soc.* **2009**, *131*, 430-431.

36. Sadiq, S. K.; De Fabritiis, G. Explicit Solvent Dynamics and Energetics of HIV-1 Protease Flap Opening and Closing. *Proteins* **2010**, *78*, 2873-2885.

37. Spinelli, S.; Liu, Q. Z.; Alzari, P. M.; Hirel, P. H.; Poljak, R. J. The Three-Dimensional Structure of the Aspartyl Protease from the HIV-1 Isolate BRU. *Biochimie* **1991**, *73*, 1391-1396.

38. Heaslet, H.; Rosenfeld, R.; Giffin, M.; Lin, Y.-C.; Tam, K.; Torbett, B. E.; Elder, J. H.; McRee, D. E.; Stout, C. D. Conformational Flexibility in the Flap Domains of Ligand-Free HIV Protease. *Acta Cryst.* **2007**, *D63*, 866-875.

39. Damm, K. L.; Ung, P. M.-U.; Quintero, J. J.; Gestwicki, J. E.; Carlson, H. A. A Poke in the Eye: Inhibiting HIV-1 Protease through its Flap-Recognition Pocket. *Biopolymers* **2008**, *89*, 643-652.

40. Perryman, A. L.; Zhang, Q.; Soutter, H. H.; Rosenfeld, R.; McRee, D. E.; Olsen, A. J.; Elder, J. E.; Stout, C. D. Fragment-Based Screen Against HIV Protease. *Chem. Biol. Drug Des.* **2010**, *75*, 257-268.

41. Perryman, A. L.; Lin, J.-H.; McCammon, J. A. Restrained Molecular Dynamics Simulations of HIV-1 Protease: the First Step in Validating a New Target for Drug Design. *Biopolymers* **2006**, *82*, 272-284.

42. Tozzini, V.; Trylska, J.; Chang, C,-e; McCammon, J. A. Flap Opening Dynamics in HIV-1 Protease Explored with a Coarse-Grained Model. *J. Struct. Biol.* **2007**, *157*, 606-615.

43. Mittal, S.; Cai, Y.; Nalam, M. N. L.; Bolon, D. N. A.; Schiffer, C. A. Hydrophobic Core Flexibility Modulates Enzyme Activity in HIV-1 Protease. *J. Am. Chem. Soc.* **2012**, *134*, 4163-4168.

44. Sham, H. L.; Zhao, C.; Stewart, K. D.; Betebenner, D. A.; Lin, S.; Park, C. H.; Kong, X.-P.; Rosenbrook, W.; Herrin, T.; Madigan, D.; Vasavanonda, S.; Lyons, N.; Molla, A.; Saldivar, A.; Marsh, K. C.; McDonald, E.; Wideburg, N. E.; Denissen, J. F.; Robins, T.; Kempf, D. J.; Plattner, J. J.; Norbeck, D. W. A Novel, Picomolar Inhibitor of Human Immunodeficiency Virus Type 1 Protease. *J. Med. Chem.* **1996**, *39*, 392-397.

45. Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER 11*, University of California, San Francisco.: **2010**.

46. Hornak, V.; Okur, A.; Rizzo, R. C.; Simmerling, C. HIV-1 Protease Flaps Spontaneously Open and Reclose in Molecular Dynamics Simulations. *Proc. Nat. Am. Sci. U.S.A.* **2006**, *103*, 915-920.

47. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of *n*-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327-341.

48. Darden, T. A.; York, D. M.; Pedersen, L. G. Particle Mesh Ewald. An N.log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089-10092.

49. Andrea, T. A.; Swope, W. C.; Andersen, H. C. The Role of Long Ranged Forces in Determining the Structure and Properties of Liquid Water. *J. Chem. Phys.* **1983**, *79*, 4576-4584.

50. Grabuleda, X.; Jaime, C.; Kollman, P. A. Molecular Dynamics Simulation Studies of Liquid Acetonitrile: New Site Model. *J. Comput. Chem.* **2000**, *21*, 901–908.

51. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996,** *118*, 11225-11236.

52. Jorgensen, W. L.; McDonald, N. A. Development of an All-Atom Force Field for Hetercycles. Properties of Liquid Pyridine and Diazenes. *J. Mol. Struct. (Theochem)* **1998**, *424*, 145-155.

53. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926-935.

54. Goetz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Grand, S. L.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER – Part I: Generalized Born. *J. Chem. Theory Comput.* **2012**, *8*, 1542-1555.

55. García, A. E. Large-Amplitude Nonlinear Motions in Proteins. *Phys. Rev. Lett.* **1992**, *68*, 2696-2699.

56. Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential Dynamics of Proteins. *Proteins* **1993**, *17*, 412-425.

57. Mongan, J. Interactive Essential Dynamics. *J. Comp.-aided Mol. Des.* **2004**, *18*, 433-436.

58. Humphrey, W.; Dalke, A.; Schulten, K. VMD – Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33-38.

59. Barrett, C. P.; Hall, B. A.; Noble, M. E. M. Dynamite: A Simple Way to Gain Insight into Protein Motions. *Acta. Cryst. D* **2004**, *60*, 2280-2287.

60. PyMOL Molecular Graphics System, Version 1.5.0.1 Schrödinger, LLC.

61. Jordan, S. P.; Zugay, J.; Darke, P. L.; Kuo, L. C. Activity and Dimerization of Human Immunodeficiency Virus Protease as a Function of Solvent Composition and Enzyme Concentration. *J. Biol. Chem.* **1992**, *267*, 20028-20032.

62. Kim, E. E.; Baker, C. T.; Dwyer, M. D.; Murcko, M. A.; Rao, B. G.; Tung, R. D.; Navia, M. A. Crystal Structure of HIV-1 Protease in Complex with VX-478, a Potent and Orally Bioavailable Inhibitor of the Enzyme. *J. Am. Chem. Soc.* **1995**, *117*, 1181-1182.

63. Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother of All Databases). *Proteins* **2005**, *60*, 333-340.

64. Heaslet, H.; Kutilek, V.; Morris, G. M.; Lin, Y. C.; Elder, J. H.; Torbett, B. E.; Stout, C. D. Structural Insights into the Mechanism of Drug Resistance in HIV-1 Protease NL4-3. *J. Mol. Biol.* **2006**, *356*, 967-971.

65. Mahalingam, B.; Louis, J. M.; Reed, C. C.; Aomat, J. M.; Krouse, J.; Wang, Y. F.; Harrison, R. W.; Weber, I. T. Structural and Kinetic Analysis of Drug Resistant Mutants of HIV-1 Protease. *Eur. J. Biochem.* **1999**, *263*, 238-245.

66. Thompson, J. D.; Higgins, D. G.; Gibson, T. J. CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Res.* **1994**, *22*, 4673-4680.

67. Subramaniam, S. The Biology Workbench – A Seamless Database and Analysis Environment for the Biologist. *Proteins* **1998**, *32*, 1-2.

CHAPTER 5

Identification of Key Hinge Residues Important for Nucleotide-Dependent Allostery

in *E. coli* Hsp70/DnaK

5.1 Abstract

DnaK is a molecular chaperone that has important roles in protein folding. The hydrolysis of ATP is essential to this activity, and the effects of nucleotides on the structure and function of DnaK have been extensively studied. However, the key residues that govern the conformational motions that define the apo, ATP-bound, and ADP-bound states are not entirely clear. Here, we used molecular dynamics simulation, mutagenesis and enzymatic assays to explore the molecular basis of this process. Simulations of DnaK's nucleotide-binding domain (NBD) in the apo, ATP-bound, and ADP/P$_i$-bound states suggested that each state has a distinct conformation, consistent with available biochemical and structural information. The simulations further suggested that large shearing motions between subdomains I-A and II-A dominated the conversion between these conformations. We found that several evolutionarily conserved residues, especially G228 and G229, appeared to function as a hinge for these motions, because they

predominantly populated two distinct states depending on whether ATP or ADP/P$_i$ was bound. Consistent with the importance of these "hinge" residues, alanine point mutations caused DnaK to have reduced chaperone activities *in vitro*. Together, these results clarify how sub-domain motions communicate allostery in DnaK.

5.2 Introduction

Escherichia coli DnaK is a member of the heat shock protein 70 (Hsp70) family of molecule chaperones that assists in protein folding and minimizes protein aggregation.[1-4] Because of its central role in the proteostasis network, DnaK has been suggested as a promising new anti-bacterial target, and its human ortholog, Hsp70, is a drug target for the treatment of cancer[5,6] and neurodegenerative disorders.[7,8] These observations have led to an increased interest in understanding the structure and function of Hsp70/DnaK.

DnaK, like other Hsp70s, consists of a nucleotide-binding domain (NBD) and a substrate-binding domain (SBD) tethered by a flexible linker. The NBD is composed of four subdomains, I-A, II-A, I-B and II-B, arranged to form a nucleotide-binding cleft (Figure 5.1).[9,10] The NBD binds and hydrolyzes ATP, and this activity is important for chaperone functions. Specifically, nucleotide turnover in the NBD regulates binding of misfolded proteins in the SBD through an inter-domain allosteric network.[8,11-13]

Crystallography and NMR have been used to provide important insights into the effects of nucleotides on the structure and function of DnaK. For example, a comparison of the crystal structures of the NBD in the apo form (1DKG[14]) and the ADP-bound form

(1BUP[15], 1KAZ[16]) suggests a substantial, nucleotide-dependent movement in subdomain II-B (Figure 5.1). This motion appears to involve rotations of subdomain II-B in relation to subdomain II-A, which is mediated by a sheet-coil-helix element (residues 222-234). NMR studies have further suggested that this region may act as a hinge for subdomain motions[17,18] and more recent structural studies on an ATP-bound form of DnaK[19,20] further support this idea. In order to understand the specific role of the sheet-coil-helix element in this hinge motion, Chang *et al.*[21] mutated a series of residues between subdomains II-A and II-B. This work identified a number of mutations that disrupted allostery and chaperone function *in vitro* and *in vivo*. However, the underlining molecular mechanisms of these mutations are not yet clear.

Computational simulations have further advanced our knowledge of allostery in Hsp70/DnaK. For example, coarse-grained molecular dynamics of full-length Hsp70 demonstrated the collective motions that are essential in the allosteric communications between the NBD and SBD.[22,23] Likewise, all-atom simulations identified residues essential in the binding of DnaK to nucleotide-exchange factors (NEFs).[24] Together, these studies have provided insight into the mechanism of DnaK and, more broadly, this system has contributed to our understanding of dynamics and allostery in biology.

Despite these insights, the molecular mechanisms of hinge motions in DnaK's NBD are still not clear. Specifically, it is not known which residues are essential to the hinge motions and how these molecular motions are affected by nucleotides. Here, we have used dynamics simulations and mutagenesis to examine the detailed communication between the protein and the nucleotide. Conformational behavior during the simulations pointed to the allosteric importance of certain residues. These residues were then

compared to their evolutionary conservation, which further highlighted the importance of the identified residues. We found that hinge residues, including G228 and its neighbors, are key regulators of this transition and were nearly invariant between prokaryotic DnaK and mammalian Hsp70, highlighting the importance of these residues for chaperone function. These results provide a detailed molecular mechanism linking ATPase activity to structural transitions in DnaK/Hsp70.
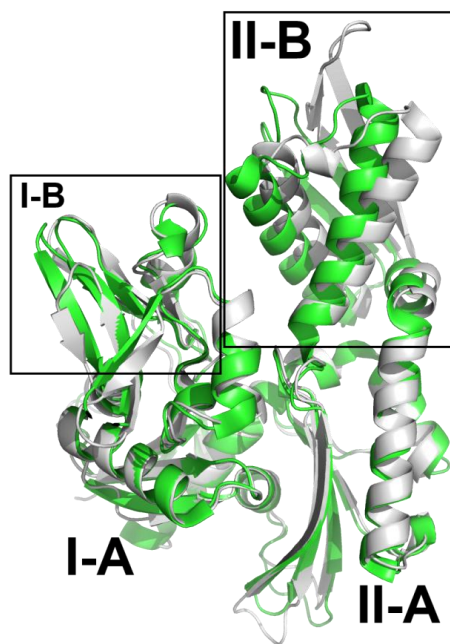
Figure 5.1. Comparison of the open and closed conformations of Hsp70/DnaK nucleotide-binding domain. Grey cartoon represents Hsc70's NBD (PDB: 1BUP) in the "closed" conformation and the green cartoon is DnaK's NBD (PDB: 1DKG) in the "open" conformation. Most of the conformational difference stems from the position of subdomain II-B relative to I-B.

5.3 Materials and Methods

Molecular Dynamics Simulations

Coordinates of *E. coli* DnaK NBD in complex with the nucleotide-exchange factor, GrpE (PDB: 1DKG[14]) were obtained from the PDB, and GrpE was discarded. The

missing side chains and short loops of NBD were introduced using Molecular Operating Environment[25] version 2005.06. AMBER 10[26] was then used to perform unrestrained all-atom LD simulations. Models of DnaK NBD in different states (apo-, ATP-, and ADP/P_i-bound) were built for these simulations. The structure 1DKG was crystallized without bound cofactors that are essential to ATPase function,[27] so the essential ions, ATP, and ADP/P_i were introduced by transferring the coordinates from the crystal structures of the closely related homolog, bovine Hsc70. The cofactors ADP, PO_4^{3-}, Mg^{2+}, and K^+ were obtained from the 1BUP,[15] whereas ATP, Mg^{2+}, and K^+ ions were obtained from 1KAZ.[16]

The protein was modeled using the FF99SB force field.[28] Parameters for P_i and Mg^{2+} were generated using the ANTECHAMBER module of AMBER; GAFF[29,30] and AM1-BCC charges were applied.[31,32] Nucleotide parameters developed by Meagher *et al.*[33] were used for ATP and ADP. The SHAKE algorithm[34] was used to restrain hydrogen atoms. Model II of a modified Generalized Born approach[35] was used to implicitly model aqueous solvation. Collision frequency of 1 ps^{-1} was used. A 999 Å nonbonded interaction cutoff was used. Default dielectric values were used: interior = 1 and exterior = 78.5.

Five independent LD simulations for each of the NBD states (apo, ADP/P_i-bound, and ATP-bound) were initiated with different random-number seeds. Hydrogen atoms were first minimized, followed by residue side chains, and finally an all-atom minimization. Heating and restrained equilibrations were followed, in which the system was heated gradually from 100 to 300 K during the first two equilibration phases, and the temperature remained at 300 K for the remaining equilibration and production phases. Restraints were placed on all heavy atoms and gradually relaxed over the first four

equilibration phases. using force constants from 2.0 to 0.1 kcal/mol·Å^2. Restraints were maintained on backbone atoms in the fifth equilibration phase, using a force constant of 0.1 kcal/mol·Å^2. All restraints were removed in the sixth phase. The first three phases were performed for 20 ps each, followed by 50 ps for the fourth and fifth. The sixth phase, unrestrained equilibration, was run for 740 ps, and the production phase was run for 5 ns. A time step of 1 fs was used, and snapshots were collected every 1 ps. Results of the LD simulations were analyzed by calculating the variance-covariance matrix of the trajectory using the ptraj module of AMBER. In-house scripts were used to parse subsets of data and to plot the correlation results for visualization and analysis.

Materials

Reagents were obtained from the following sources: Platinum *Pfx* DNA polymerase (Invitrogen, Carlsbad, CA); pMCSG7 plasmid (Midwest Center for Structural Genomics, Bethesda); QuikChange site-directed mutagenesis kit (Stratagene, La Jolla, CA); DNA purification Wizard Plus SV miniprep kit (Promega); His-tagged Tobacco Etch Virus (TEV) protease (Invitrogen); Z Competent *E. coli* transformation cells (Zymo Research); ATP-agarose column (Sigma); nickel-nitrilotriacetic (Ni-NTA) His-Bind resin (Novagen, Darmstadt, Germany); bicinchoninic acid (BCA) protein assay (Thermo Scientific); *dnaK E. coli* cells (a generous gift from Dr. Ursula Jacob); Malachite green, ammonium heptamolybdate tetrahydrate, polyvinyl alcohol (Sigma); NRLLLTG peptide (University of Michigan Peptide Core);[36] luciferase and SteadyGlo Reagent (Promega).

Plasmids Construction and Protein Purification

The *E. coli dnaK* genes were amplified by PCR using Platinum *Pfx* DNA polymerase and inserted into the pMCSG7 plasmid as NdeI-HindIII fragment through ligation-independent cloning.[37] The optional N-terminal $(His)_6$-tag with a TEV cleavage site was expressed along with the inserted *dnaK* gene. The primers for site-directed mutagenesis (see the Supplemental Information) were designed based on the report of Zheng *et al.*,[38] and QuikChange mutagenesis was carried out following the manufacturer protocol. Primers (Invitrogen) used for *dnaK* mutagenesis are listed in the Supplementary Data section. Plasmid DNA purification was carried out using Wizard Plus SV miniprep kit.

Based on the simulations, the following *dnaK* mutants were made: I202A, S203A, G223A, L227A, and G228A. Each mutant and the wild-type (WT) DnaK was expressed as an N-terminal His-tagged protein and purified from BL21(DE3) cells as previously described.[21,39] Briefly, with all purification steps carried out at 4 °C, cell pellets were suspended in buffer A (25 mM Tris, 10 mM KCl, 5 mM $MgCl_2$, pH 7.5) containing 0.01 mM PMSF and 1 mM DTT and disrupted using a microfluidizer (Microfluidics). Purification began with the removal of His-tagged DnaK from the lysate by batch purification using Ni-NTA His-Bind resin. The His-tag of the partially purified DnaK was then cleaved by His-tagged TEV protease (1 mM DTT, 4 °C, overnight incubation).[37] After adjusting the $MgCl_2$ and KCl concentrations to 10mM, the sample was further purified by ATP-agarose column. The column was then washed extensively with buffer A and buffer A containing 1 M KCl. The bound protein was eluted with buffer A containing 3 mM ATP. The eluate would run through the Ni-NTA column to

capture the remaining cleaved His-tag and His-tagged proteins. Finally, the eluted protein was exchanged into buffer A and concentrated.

Purity of the proteins was greater than 90%, as judged by Coomassie staining. Circular dichroism spectra were consistent with the structure of folded NBD. Proteins in buffer A (DnaJ in modified buffer indicated above) were frozen on liquid nitrogen and stored at -80 $^{\circ}$C until use.

Circular Dichroism Spectroscopy

WT DnaK and mutants were prepared in 10 mM sodium phosphate buffer with 100 mM sodium fluoride (pH7.4) and circular dichroism (CD) spectra were collected at 0.1 mg/ml in a 0.1-cm cuvette at room temperature. CD spectra were recorded on a Jasco J-715 spectropolarimeter (Jasco, Easton, MD) at 1-nm intervals from 190 to 240 nm at a scanning speed of 50 nm/min and a 5.0 nm bandwidth. Each spectrum reported is the average of 15 scans after the subtraction of the base-line spectrum (phosphate buffer with sodium fluoride) and normalization.

Thermal Stability Assay

ThermoFluor is a fluorescence-based thermal shift assay system that assesses protein stability.[40,42] To test the relative stability of DnaK and its mutants, each protein was diluted to 0.2 mg/mL in either buffer A (100 mM sodium phosphate, 5 mM $MgCl_2$, 20 mM KCl, pH 7.2) or buffer B (100 mM HEPES, 5 mM $MgCl_2$, 20 mM KCl, pH 7.2), as indicated. To these solutions was added ADP or ATP (1 mM) and 1,8-ANS (100 µM). These samples were placed in 384 well plates and covered with silicone oil to minimize

evaporation. The fluorescence of ANS was monitored with excitation/emission at 375/480 nm. Measurements were performed in continuous mode, using a temperature range of 30 to 85 $^\circ$C and increments of 1 $^\circ$C. The assay was conducted with the thermospectrometer TF-1536 (Johnson & Johnson).

ATPase Assays

ATPase activity was determined by detecting the amount of phosphate accumulated in the reaction. Detection was achieved through the colorimetric determination of the colored malachite green-phosphomolybdate complex formed by the coordination of phosphate with ammonium molybdate under acidic condition.[43,44] The assay procedure followed previous reports[45] with modifications. Stock solutions of malachite green (0.081% w/v), polyvinyl alcohol (2.3% w/v), and ammonium heptamolybdate tetrahydrate (5.7% w/v in 6 M HCl) were prepared and mixed with water in the ratio of 2:1:1:2, respectively, to generate the malachite green reagent. This reagent was stable for use at 4 $^\circ$C for at least 1 week. All absorbance and luminescence measurements were performed using a SpectraMax M5 (Molecular Devices, Sunnyvale, CA).

Chaperone system was prepared with the addition of DnaK, DnaJ, and/or GrpE to a total volume of 15 µL in each well.[39] 10 µL of 2.5 mM ATP was added to start the reaction. The final concentrations were as follows: ATP (1 mM), DnaK (1 µg) unless otherwise noted. DnaJ varied from 0 to 3 µg and GrpE from 0 to 0.2 µg. Intrinsic ATPase rate was measured with DnaK in the absence of co-chaperones or substrate.

For steady state conditions, samples were incubated at 37 °C for 1–3 h, and then 80 µL of malachite green reagent was added to each well, immediately followed by 10 µL of 32% (w/v) sodium citrate to quench ATPase activity and to develop the colored complex. Samples were incubated at 37 °C for 15 min. before absorbance was measured at wavelength 620 nm. All experiments were performed in triplicate, and the signal from nonspecific ATP hydrolysis in controls lacking DnaK was subtracted. A phosphate standard curve (potassium dibasic phosphate) was used to convert the units to µM of P_i/µg of DnaK/hr.

Luciferase Refolding Assay

The luciferase refolding activity of DnaK and mutants was evaluated through the refolding of unfolded luciferase and the subsequent luminescent reaction it catalyzes.[46] Denatured firefly luciferase was prepared with a concentrated stock (8.2 µM) of luciferase with 6 M guanidine hydrochloride in 25 mM HEPES buffer (50 mM potassium acetate, 5 mM DTT, pH 7.2). This stock was incubated at room temperature for 1 h and then diluted to 0.2 µM with the same HEPES buffer without guanidine hydrochloride. This dilution preparation was used as the stock solution for final sample preparation. 10 µL of enzyme mix containing DnaK, DnaJ, GrpE, and denatured firefly luciferase in 39 mM HEPES (170 mM potassium acetate, 1.7 mM magnesium acetate, 3 mM DTT, 12 mM creatine phosphate, 50 units/ml creatine kinase, pH 7.6) was first added into each well, followed by the addition of 4 µL of 3.5 mM ATP dissolved in water to initiate the DnaK refolding reaction. The final concentration of DnaK was 1 µg; denatured luciferase was 8 nM, and ATP was 1 mM. Reaction equilibrium was reached after 1 h of incubation

at 37 °C. 14 μL of 0.5 or 2% (v/v) SteadyGlo reagent in 50 mM glycine buffer (30 mM MgSO$_4$, 10 mM ATP, and 4 mM DTT, pH 7.8) was added into each well, and the luminescence was measured. For each experiment, the signal from a negative control containing all reaction components but DnaK was subtracted.

Induced Fit Docking

Docking poses of the ligand 115-7C in the DnaK NBD was carried out with the induced-fit docking (IFD),[47] a docking protocol developed by Schrödinger.[48] The IFD protocol uses the Glide docking module and the Prime modeling modules of the Schrödinger Suite and involves five steps: (1) generating alternative receptor loop conformations, (2) docking the ligands into the defined binding site with softened-potential, (3) refining residue side chain conformations near each ligand pose from previous docking step, (4) redocking ligand into the induced-fit optimized receptor, and (5) scoring the ligand-receptor complex.

Ligand and protein receptor were prepared in Maestro (version 8.5) and default OPLS force field was applied. IFD performed a constrained minimization of the receptor, followed by the loop prediction function of Prime to generate a maximum of 5 alternative loop (D208 to T215) conformations for IFD. Residue side chains within 5.0 Å of the defined loop segment was also refined. Predicted alternative structures with energy 30 kcal/mol above the lowest-energy structure were rejected. In the first stage of docking, energy potential grid was centered at residue E206 (based on the NMR results) and the size of the grid box was determined automatically. Glide docked the ligand into each of the five alternative receptor structures with a softened van der Waals potential and the top ten scoring ligand poses were retained. To "induced-fit" the receptor side chains to the

docked ligand, side chain of residues within 5.0 Å of the docked ligand was refined with Prime, allowing the side chains to adopt optimal conformation to interact with the ligand. The ligand was then re-docked into the induced-fit receptor and the final ligand poses were scored using a combination of Glide SP score and the receptor energetic terms derived from the Prime. Ligand poses with ligand–receptor interactions in agreement with the observed NMR chemical shift data were retained for analysis.

5.4 Results and Discussion

Dynamics Simulations of DnaK's NBD

To initiate these studies, models of DnaK in different nucleotide states (apo, ADP/P_i-bound, and ATP-bound) were constructed based on the crystal structure of the DnaK NBD (PDB: 1DKG). The nucleotide-bound models were constructed from the apo state because structures of these states were not yet available at the time of this study,[19] but these models agree well with recently reported structures (RMSD = 3.2 Å for the large, flexible, multidomain protein). Since 1DKG does not carry nucleotide and metal ions, these cofactors were introduced from structures of bovine Hsc70 (PDB: 1BUP and 1KAZ). These models were solvated with the Generalized Born implicit solvent model to accelerate the conformational sampling using the Langevin dynamics (LD) method. Then, for each model, five independent trajectories were generated and confirmed to be structurally stable (Figure 5.2 and Figure 5.3).
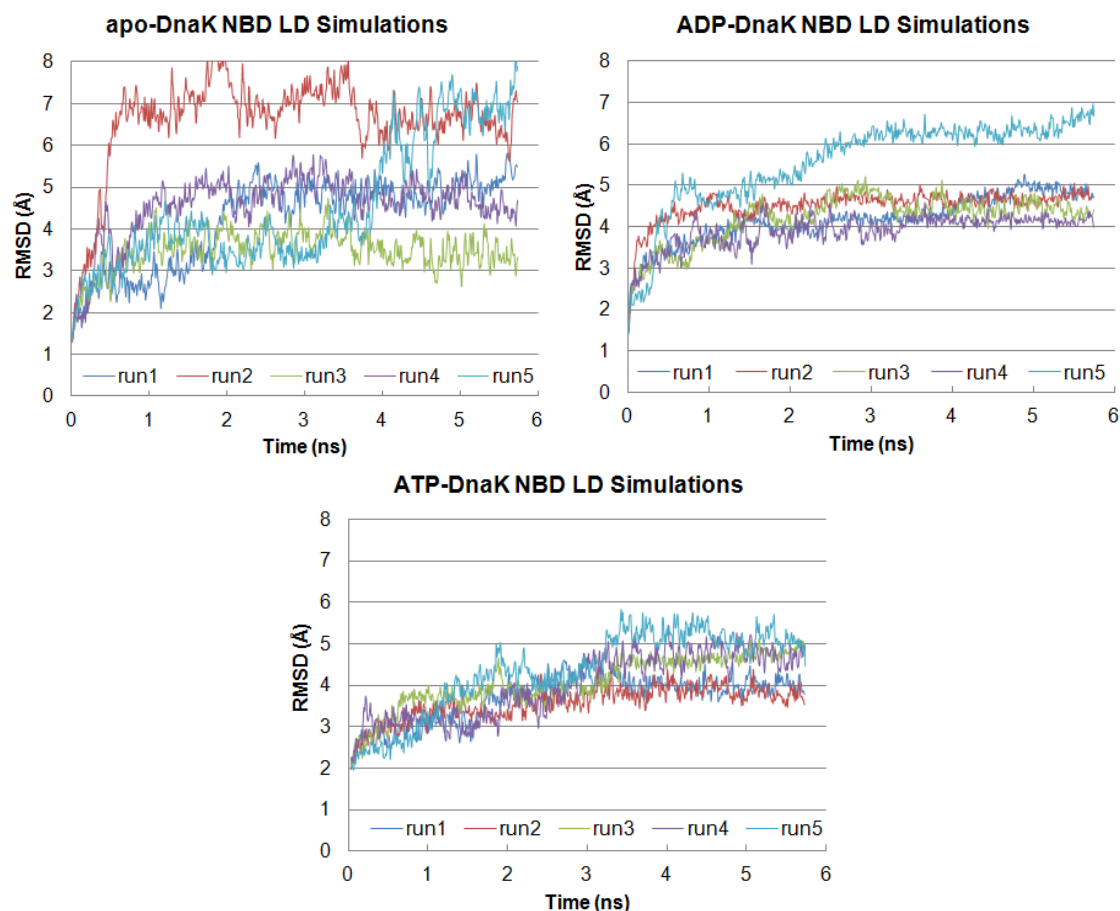
Figure 5.2. Stability of DnaK NBD in LD simulations. Three nucleotide-bound states were modeled. In the production run (after 0.75-ns), apo state simulations remained in the "open" conformation, with large fluctuations in C_α RMSD, ~ 4.9 ± 1.4 Å. The NBD in either ADP- or ATP-bound states converted from an initially "open" conformation to a "closed" conformation, leading to a relatively high Cα RMSD. The mean Cα RMSD of ADP-DnaK and ATP-DnaK complexes were ~ 4.6 ± 0.8 Å and ~ 4.1 ± 0.6 Å, respectively. However, the trajectories became stable once closing occurred.

When all restraints were removed, the NBD bound to either ATP or ADP/P$_i$ *spontaneously converted* from the initial "open" conformation of an apo state to a "closed" conformation in which the subdomains I-B and II-B moved toward each other. Previous dynamics simulations of DnaK and its homologs bound to various nucleotides also show generally rigid motions of IIB relative to IB and IIA.[22-24,49,50] These motions were primarily driven by movement of the entire subdomain II-B, such that the global

conformation of the NBD can be described by the relative position of the intact subdomains. Thus, the distances between the center-of-mass (COM) of II-B and subdomains I-A/B were used to quantify and compare the "openness" of each conformation (Table 5.1). Comparing the apo, ATP-, and ADP/P_i-bound states of NBD, we observed that the apo state adopts the most "open" conformation, with an average inter-subdomain distance of ~32 Å. In contrast, the ATP- and ADP/P_i-bound states adopt closed conformations, with average distances of 27.3 and 26.1 Å, respectively. These observations are consistent with biochemical studies on *E. coli* DnaK[51] in which the binding of DnaK to both ATP- and ADP was found to be similar and tight (K_d ~ 1 nM) and slow nucleotide release rate (0.008 min^{-1} and 0.02 min^{-1}, respectively).

In some of the LD trajectories, the α-helix of subdomain II-B (residues 230-280) was observed to "bend" near residues 262 – 267 (Figure 5.4). This bending did not appear to be dependent upon the nucleotide state because it was observed in the simulations of the apo, ATP-, and ADP/P_i-bound NBDs. Similar helix bending has also been described in actin and a phosphatase with a similar fold.[9] Although the functional importance of this transition is not clear, it may be involved in assisting with the full opening and closing of the NBD.

Table 5.1. Center-of-mass distance between subdomains II-B and I-A/B in Langevin Dynamics.

	Subdomain II-B to I-A/B+II-A COM Distance (Å)		
	apo	ATP	ADP/P_i
Run 1	34.7 ± 1.6	27.0 ± 0.6	26.7 ± 0.4
Run 2	29.7 ± 0.8	27.7 ± 0.5	25.9 ± 0.4
Run 3	34.0 ± 1.0	25.3 ± 0.5	26.2 ± 0.4
Run 4	27.0 ± 0.5	27.9 ± 0.4	27.1 ± 1.6
Run 5	32.4 ± 3.8	27.3 ± 0.6	25.5 ± 0.4
Mean	31.9 ± 3.5	27.3 ± 1.0	26.1 ± 0.9
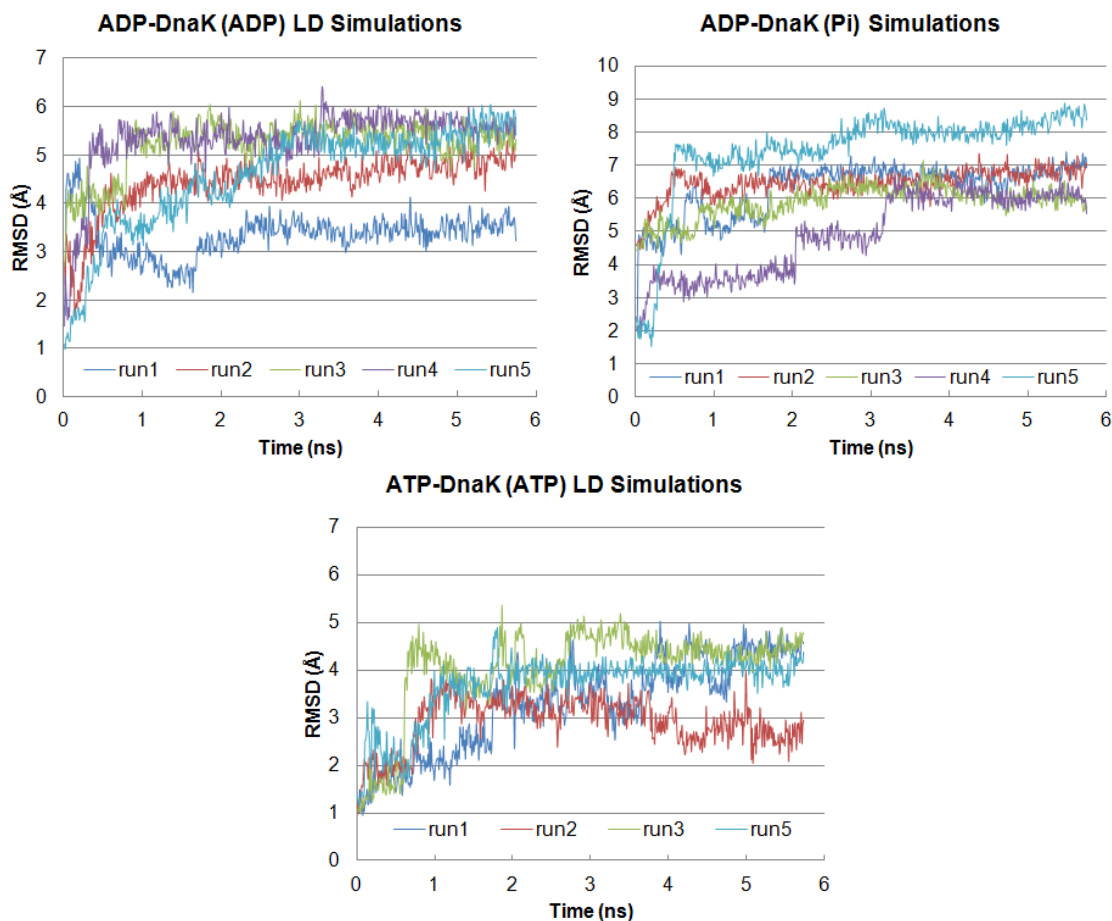
Figure 5.3. Stability of cofactors in the LD simulations. The ADP-bound NBD complex contained both ADP and P_i. In the production run (after 0.75-ns), the mean heavy-atom RMSD of ADP and P_i in the ADP-DnaK NBD complex were 4.7 ± 0.9 Å and 6.4 ± 1.1 Å, respectively. For ATP in the ATP-DnaK NBD complex, the mean heavy-atom RMSD throughout the trajectories was 3.7 ± 0.7 Å.
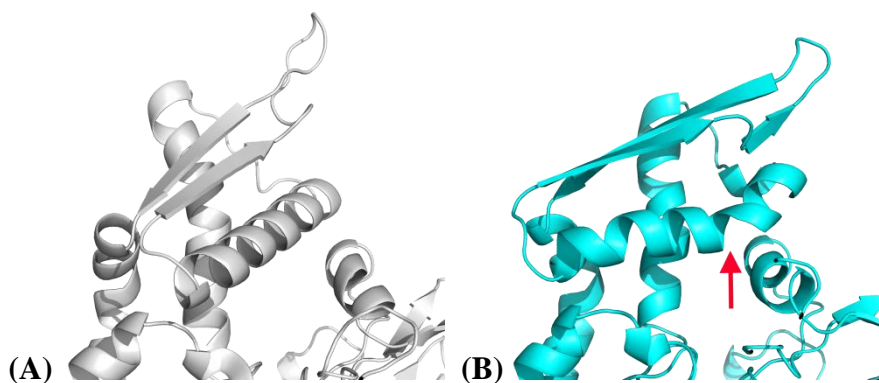
Figure 5.4. Conformational change in the α-helix (residues 257-274) of NBD subdomain II-B. **(A)** The most common conformation of the α-helix of subdomain II-B, as observed in most Hsp70 crystal structures. **(B)** Bending of the α-helix near residue 262 was observed in several LD simulations (red arrow).

194

Essential Dynamics

To study conformational changes in the NBD, essential dynamics (ED) was used to examine the LD trajectories of each model. Previous studies have shown that combinations of several independent dynamics simulations often provide sufficient sampling of conformational space for ED.[52,53] It utilizes principle component analysis to extract the subsets of correlated motions, called principle components or eigenvectors, and sorts them according to their contribution to the overall motion of the ensemble. Previous studies have shown that combinations of several independent dynamics simulations often provide sufficient sampling of conformational space for ED.[54] Thus, five independent LD trajectories were combined for each model and analyzed with ED.

In all cases, the first 5 eigenvectors (1st – 5th) represented ~ 80% of the principle motions of the protein (Figure 5.5). The apo and nucleotide-bound models yielded generally similar global motions, and the lowest mode of the ED was a shear motion between subdomains I and II (Figure 5.6 A). This motion involved two crossed α-helices (171 – 179 and 367 – 377) with inter-helical angles of approximately 70° to 80°. This shear motion occurred in the interface between subdomains I-A and II-A, and it manifested as a large displacement between subdomains I-B and II-B. Actin and glucokinase, which both have similar folds to DnaK, have been shown to undergo a similar shear motion.[55,56] The second eigenvector involved a rotational motion of subdomain II-B, in which II-B moves relatively independent to the other subdomains (Figure 5.6 B). This motion appeared to be a hinged-motion between subdomains II-A and II-B. Together, the 1st and 2nd eigenvectors described motions involved in the
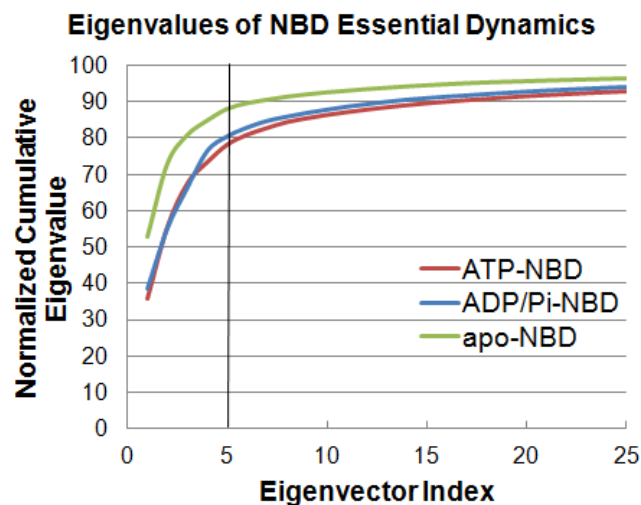
Figure 5.5. Normalized cumulative eigenvalues of NBD essential dynamics. In all cases, 25 eigenvectors describe > 90% of the essential motions in the simulations, while 5 of the lowest modes are sufficient to describe ~ 80% of the essential motions (vertical line).

opening and closing of the nucleotide-binding site, and these motions are consistent with experimental observations from NMR.[18]

Torsion Angle Analysis

The movement of protein domains is often linked to protein function,[57] and these movements typically involve the motion of rigid domains relative to each other.[55,58] These structural changes cluster into two major mechanisms: shear movements in which domains slide along one another while maintaining closely packed interfaces, and hinged movements in which relatively disordered regions connecting the domains undergo significant backbone conformational changes.[55] Hinged movements usually involve a small set of residues that adopt distinct torsion angles of their φ (C'-N-C$^\alpha$-C') and ψ (N-C$^\alpha$-C'-N) bonds, enabling these individual "hinge" amino acids to mediate larger motions between domains or subdomains.[58] Therefore, it is often possible to identify the key
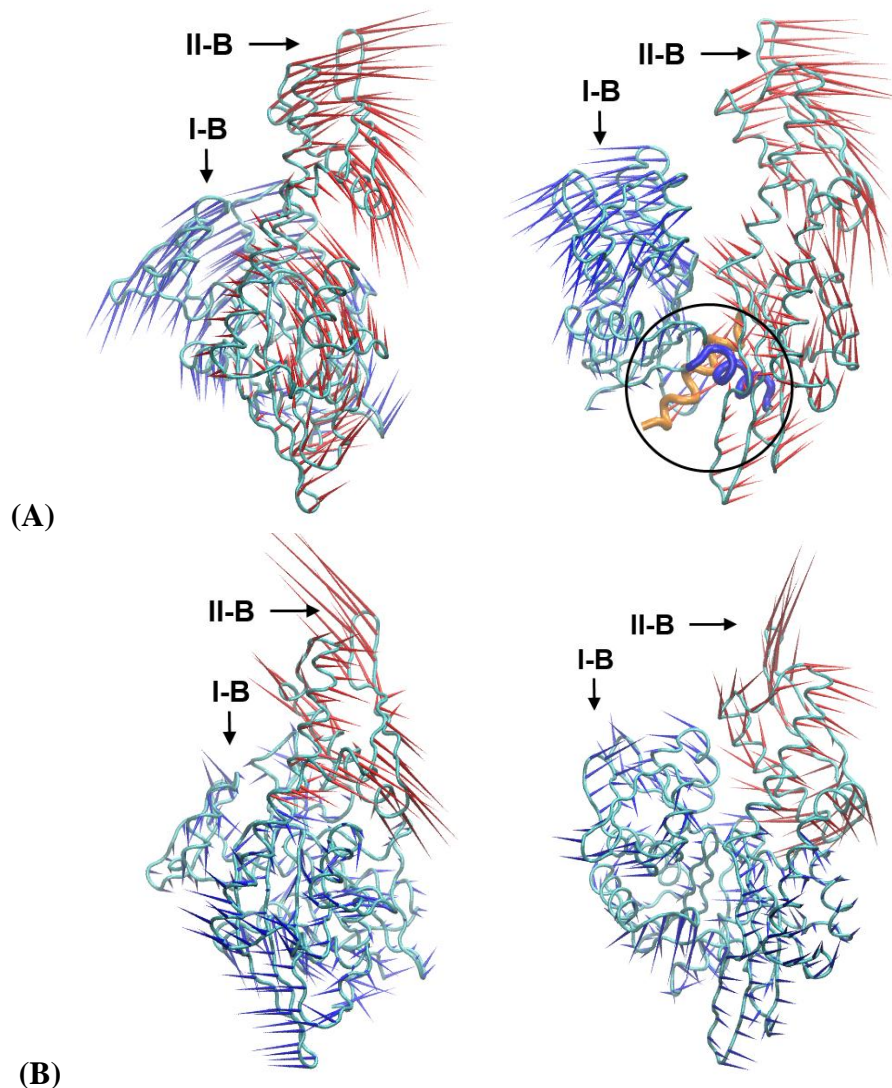
Figure 5.6. **(A)** The 1st eigenvector of NBD essential dynamics (ED) involves a shearing motion of the domains I and II. *Left* – side view of the NBD, in which ED vectors of domains I and II are colored blue and red, respectively. *Right* – front view of the NBD. The shearing motion of domains I and II is manifested through the two helices at the interface of subdomains I-A and II-A, as indicated in the circle. **(B)** 2nd eigenvector of NBD ED involves a rotating motion of primarily subdomain II-B, which may change the distance between subdomains I-B and II-B. Subdomain II-B is colored red, and the other three subdomains are in blue.

hinge residues by examining which amino acids undergo changes in their φ-ψ torsion angles during dynamic simulations.

One of the key questions in the DnaK system is which residues might be involved in the nucleotide-dependent hinge motions. To detect such hinge residues, we performed C_α torsion angle analysis on the NBD trajectories (Figure 5.7). Our Ramachandran results of the amino acids are similar to those described by Hovmöller et al.,[59] in which residues in the helix, sheet, and random coil region have different φ-ψ distributions. As expected, we found that the great majority of residues, such as A191 and K245, do not have the properties of hinged movement, such that the Ramachandran plot of these residues show a single φ-ψ torsion angle cluster throughout the simulations. In contrast, a small number of residues, such as G74, had more than one conformation evident from multiple φ-ψ clusters, which suggests that they may be involved in hinged movement (Table 5.2). Further, a subset of these residues (N13, N14, P67, A68, I73, R84, D85, T184, T185, S223, L227, G228, and G229) had multiple conformations with distinct φ-ψ clusters. Further, some of these φ-ψ clusters were correlated with changes in nucleotide state.

Next, we examined the location of the residues with multiple φ-ψ clusters on the DnaK NBD structure. These residues are found in all 4 subdomains with a bias for the flexible loops between subdomains I-A and I-B (Figure 5.8 A). Additionally, several clusters were grouped in subdomain II-A, including residues 183-186 and 340-342, which form flexible loops, and residues 195, 198 and 199, which are found in a flexible β-hairpin that coordinates with the phosphates of the nucleotide (Figure 5.8 B).

Table 5.2. Residues with multiple φ-ψ clusters in dynamic simulations

	Residue
I-A (1-36, 111-180)	N13, S14, D20
I-B (37-110)	G51, P67, A68, I73, G74, Q83, R84, D85, E104, V105
II-A (181-226, 311-383)	K183, G184, T185, G186, L195, G198, T199, I202, S203, N222, S223, L227, G228, G229, V340, G341, G342, K363
II-B (230-310)	V281, A288, D289
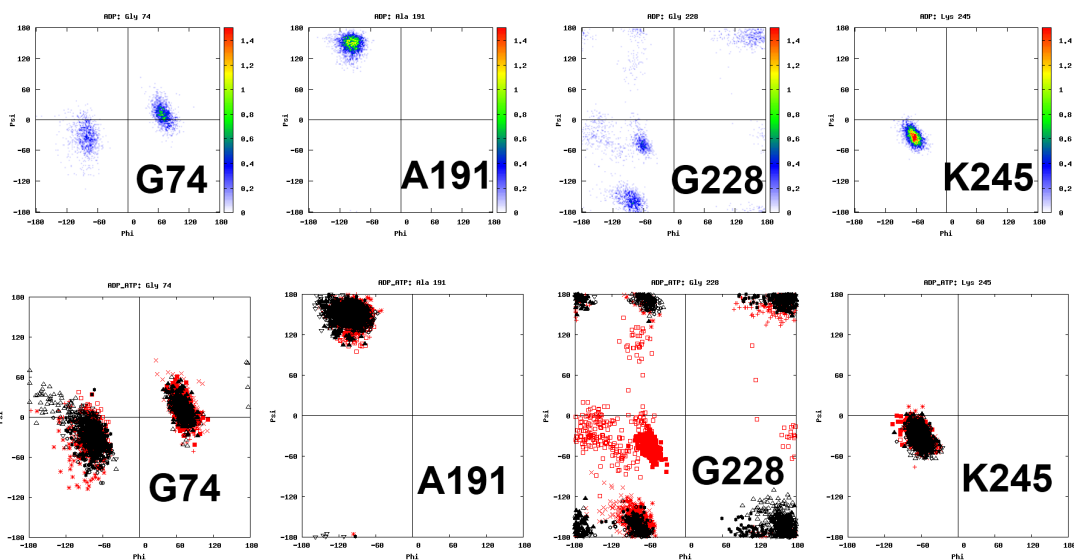
Figure 5.7. Examples of Ramachandran plots of NBD residues in the LD simulations. **(A)** Density maps of the φ-ψ torsion angles that were highly occupied (red). Residues not involved in hinged movement usually had one φ-ψ torsion angle cluster in the plot (A191 and K245). However, other "hinge" residues had two or more φ-ψ clusters (G74 and G228), which indicated substantial conformational changes in the backbone. **(B)** Comparing the ψ-φ clusters generated from the ADP/P_i-bound (red) and ATP-bound (black) trajectories. Residues not involved in hinged movement (A191, K245) have overlapped φ-ψ clusters, while hinge residues have φ-ψ clusters that differ significantly. Some of these hinge residues were affected by nucleotide changes (G228), while others were not (G74).

In DnaK, the lobe between I73/G74 and E104/V105 consists of a helix-loop-sheet structure. We found that this lobe is capped by two sets of residues with multiple φ-ψ clusters (Figure 5.7 and Figure 5.8 A), suggesting that these residues might function as hinges. This movement might be interesting because this lobe is directly adjacent to the nucleotide-interacting residue R71, which plays a critical role in enzymatic activity. Interestingly, sequence alignment (data not shown) of the Gram-(-) *E. coli* DnaK and a Gram-(+) *Geobacillus kaustophilus* DnaK indicates an extensive deletion in Gram-(+) DnaK that corresponds to residues 79-102, the helix-loop-sheet lobe capped by the hinge residues.[60]
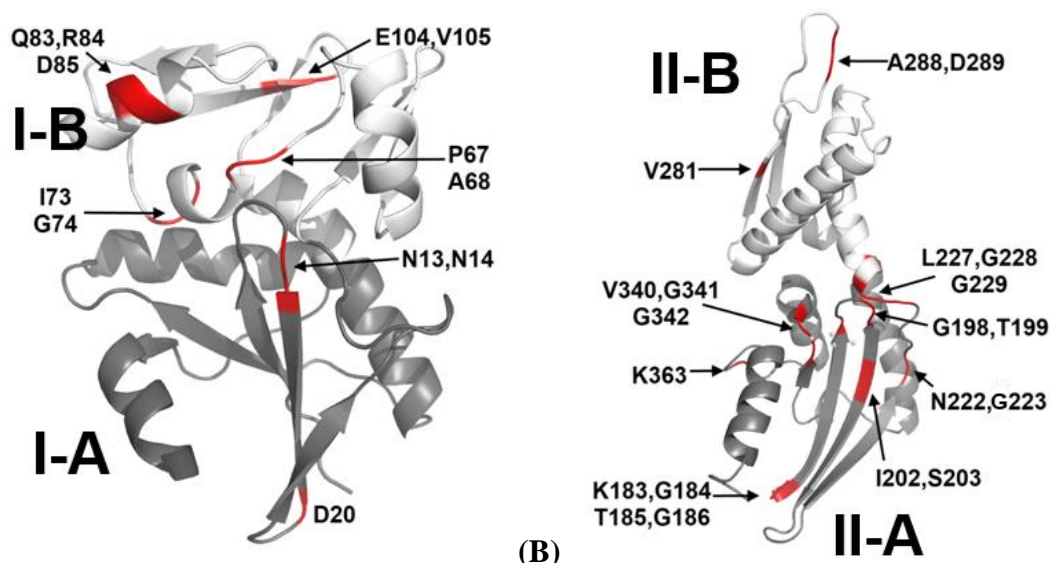
Figure 5.8. Residues with multiple Cα φ-ψ states during dynamics simulations. **(A)** Cartoon representation of Domain I. Subdomain I-A is colored in gray and I-B in white. **(B)** Cartoon representation of Domain II. Subdomain II-A is colored in gray and II-B in white. Residues with multiple φ-ψ states are colored in red. Most colored residues are located in random coils, loops or end of the α-helices or β-sheets.

Finally, another significant cluster of residues with multiple φ-ψ clusters was identified near the junction between the subdomains II-A and II-B. These clusters included residues in β-sheets (I202, S203) and random coils (G223, L227, G228, G229). The β-sheet residues were not sensitive to nucleotide state (Figure 5.9), implying that they may be involved in intrinsic protein dynamics. However, residues G223, L227, G228, and G229 had multiple φ-ψ states, and their states changed in response to nucleotide (Figure 5.10 A-D). Residues with multiple φ-ψ states and change in population in response to nucleotide are shown in Figure 5.10 E. These residues seemed like good candidates for playing a key role in nucleotide-dependent structural transitions and allostery, as suggested by Chiappori et al.[49]
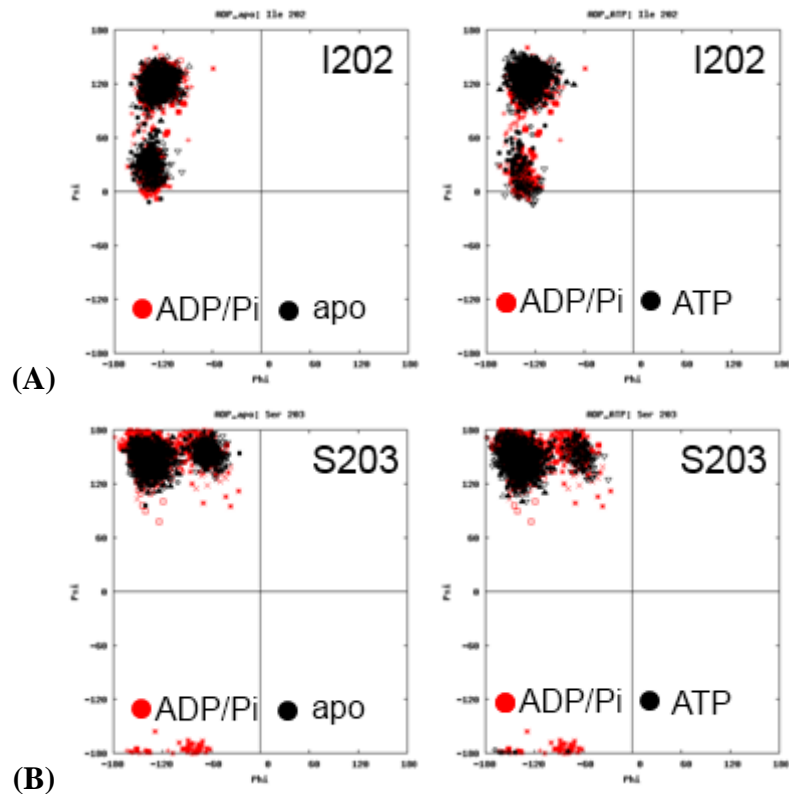
Figure 5.9. Ramachandran plots of NBD residues in different nucleotide-bound states. **(A)** I202 and **(B)** S203. The φ-ψ torsion angle clusters do not appear to respond to change in nucleotide-bound state.

Correlations to Nucleotide State

As a complementary way to understand how nucleotides might affect motions in DnaK, the correlation matrices of each trajectory from the LD simulations were compared (Figure 5.11). Then, residues with motions strongly correlated with ATP or ADP were extracted and ranked according to their degree of absolute correlation (1.0 correlation was just as important as -1.0 anti-correlation). The 60 most-correlated residues were tallied for each LD trajectory. Residues that were highly correlated in at least 4 of the 10 independent trajectories were considered significant (Table 5.3).
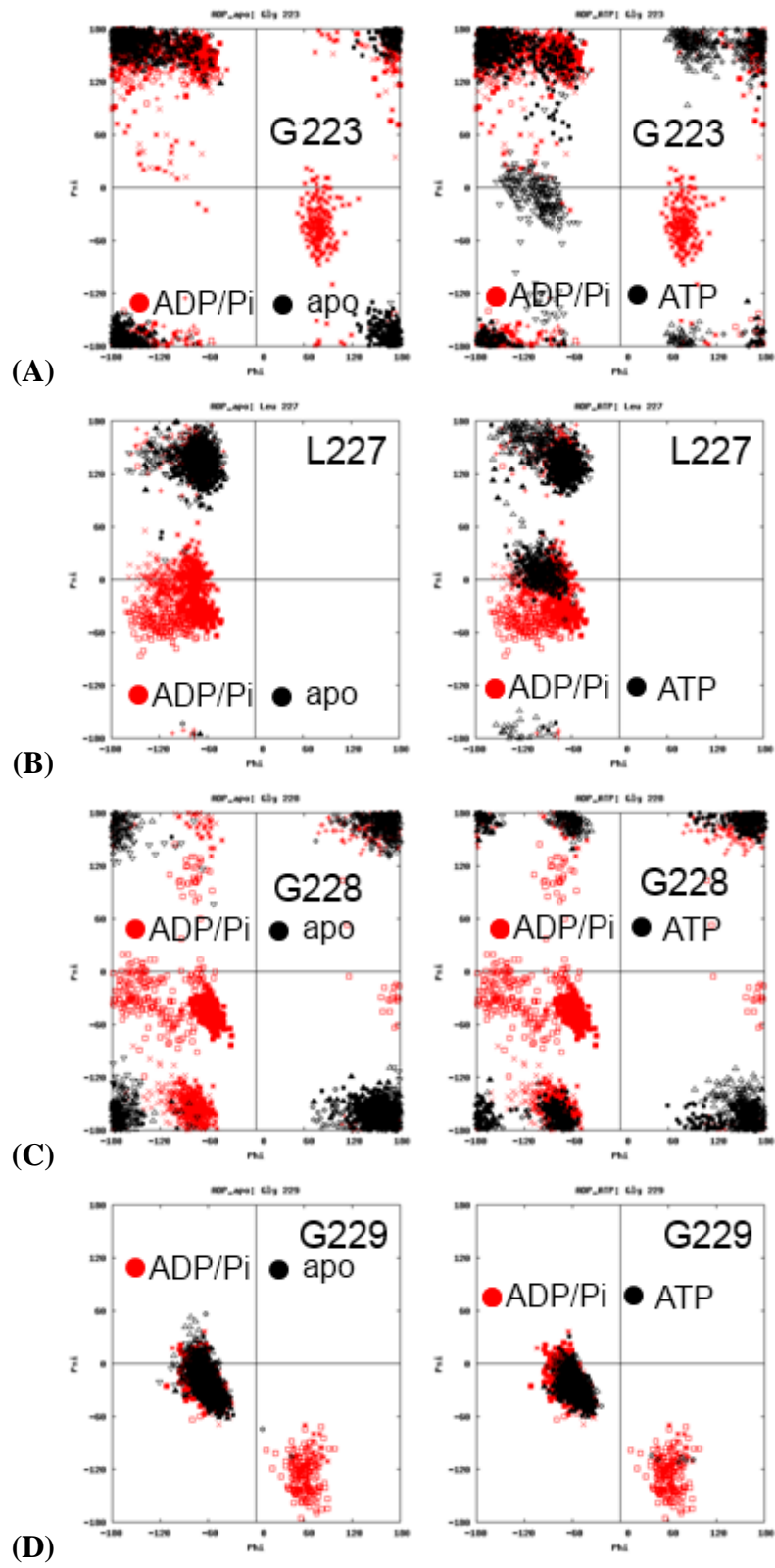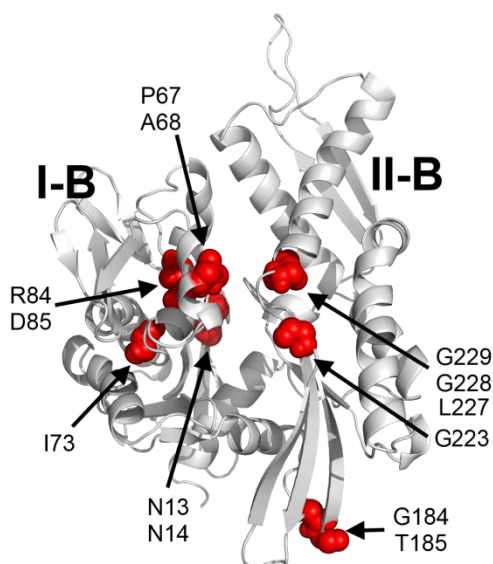
(E)

Figure 5.10. Ramachandran plots of residues **(A)** G223, **(B)** L227, **(C)** G228, and **(D)** G229 in different nucleotide-bound states. These residues had φ-ψ torsion angle clusters that were significantly different in the ADP/P$_i$ state compared to the apo state and ATP-bound state. **(E)** The locations of these nucleotide-sensitive residues are shown in red.

(A)   **(B)**
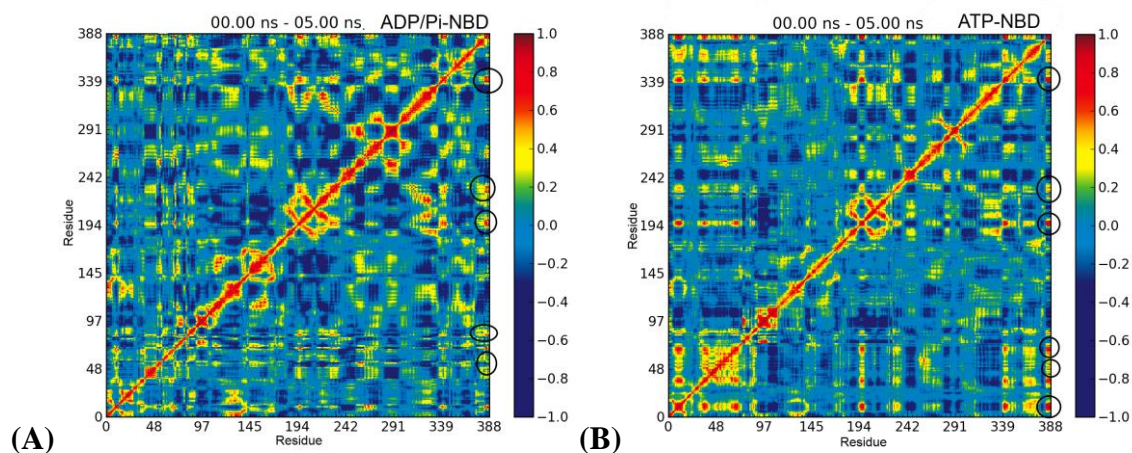
Figure 5.11. Correlation maps of nucleotide-NBD LD simulations. Strong positive correlations are in red and yellow, anti-correlated motion in dark blue. Over the entire simulation, residues that have strongest positive correlations with **(A)** ADP and **(B)** ATP are circled in black (G51, A68, L195, G223, G228, G229 and V340).

Table 5.3. Residues with strong correlation to nucleotide in NBD LD simulations

Subdomain	Residue
I-A	D8, C15, D33, P37, S38, P143, A144, Y145
I-B	I39, G51, L66, A68, L72
II-A	D194, L195, F200, G223, T225, G228, G229, V230, L339, T344, M346, P347, V349
II-B	A265, K268, A269, S274

203

Residues with weak or no correlation to the nucleotides were largely found on the surface of the NBD and at the flexible loops at the distal ends of the subdomains. In contrast, residues with strong correlation were found in the interior of all four subdomains. As expected, most of these residues were in the vicinity of the nucleotide-binding site (< 8 Å from the nucleotide), but a few were further away (Figure 5.12 A). Comparing these residues to those identified in the φ-ψ torsion angle analysis, we found that several residues are present in both lists, including G51, A68, L195, G223, G228, G229, and V340 (Figure 5.12 B).

An examination of these residues suggested that they may have significant effects on the functionality of NBD. For example, A68 is adjacent to K70 and R71, key residues that bind to nucleotides.[61] Likewise, L195 and V340 may interact with I202 and S203 to define the shape of the nucleotide-binding site. Residues G223, G228, and G229, were especially interesting because (as mentioned above) they are located away from the nucleotide-binding site and situate in a random coil that tethers subdomains II-A to II-B.

To further explore the potential importance of the "hinge" residues, we studied which residues had correlated motions with G223, G228 and G229. Dynamics simulations of NBD in the apo, ATP-bound, and ADP/P_i-bound states were examined, and among the top 60 strongly correlated and anti-correlated residues, we found strong enrichment of components of the known proline switch that consists of residues K70, R71, P143, A144, Y145, F146, R151 and E171. The proline switch has previously been found to regulate allostery in Hsp70 chaperones.[61] In simulations of the NBD in all the apo-, ATP- and ADP/P_i-bound states, K70 and R71 were consistently correlated with each of the "hinge" residues (i.e., G223, G228, and G229). In contrast, P143, A144,
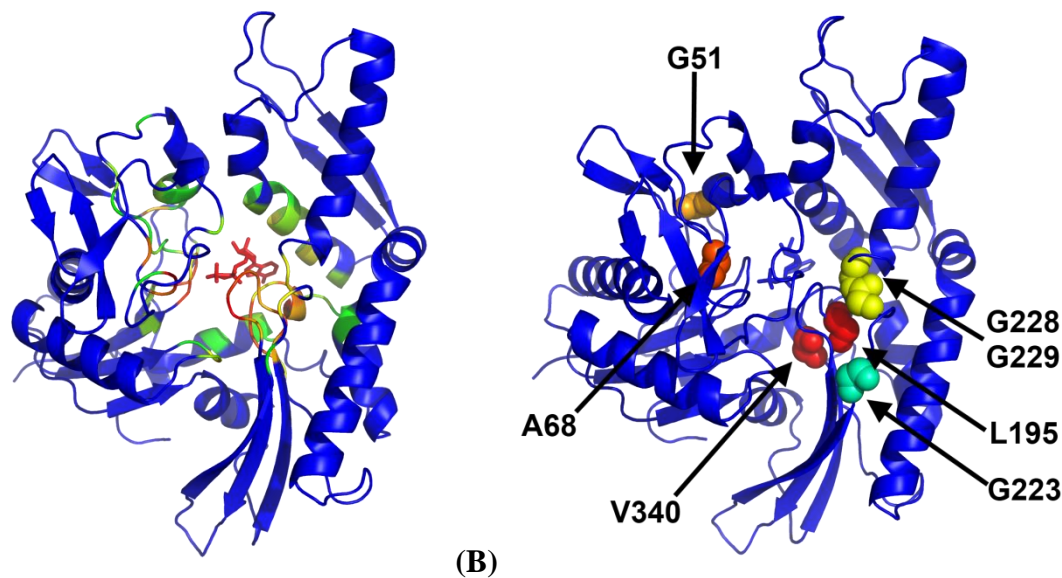
Figure 5.12. (A) Residues correlated with the nucleotides, listed in Table 3, are colored by how frequently the behavior is observed in the LD simulations of ATP- and ADP/P_i-bound NBD: red, 10 out of 10; yellow, 6 out of 10; green, 4 out of 10; blue < 4. **(B)** Most of the correlated residues are in the nucleotide binding site; here, we highlight the residues with multiple φ-ψ torsion angle clusters, strong correlated motion to nucleotides, and > 8 Å away from nucleotide.

Y145, F146, R151, and E171 correlated to G228 more frequently in the ATP-bound and ADP/P_i-bound states than in the apo-state. This result might be expected because NBD in the apo-state simulations had a wider range of movements (Figure 5.2). In all cases, correlations between the hinge residues and the proline-switch residues appeared more frequent in the ATP-bound state than in the ADP/P_i-bound state. These results suggest that interactions among ADP, P_i, the hinge residues, and the proline-switch residues may be looser than when ATP is bound.

Hinge Residues are Highly Conserved

To understand whether the proposed φ-ψ "hinge" residues might be preferentially conserved among Hsp70 family members, BLASTP[62] was used to align the sequences of

DnaK and all other available members of the Hsp70 family (3700 orthologs) from the GenBank database. Several residues in the random coil that tether subdomains II-A and II-B (G223, L227, G228, and G229) were highly conserved among bacteria, with over 90% identity and nearly 100% similarity. Remarkably, *high conservation was also observed across all kingdoms, particularly for L227, G228, and G229* (Table 5.4). This level of conservation was greater than expected for the average residue in these chaperones; the overall conservation between the prokaryotic *E. coli* DnaK and the eukaryotic human Hsp70 is ~ 50%. These findings suggest that the highly conserved residues might be especially important for the function of Hsp70 family members.

Experimental Testing of DnaK Point Mutations – Residue G228 is Critical for Chaperone Functions

Next, we designed mutations of residues in DnaK to understand their potential role in allostery and nucleotide-dependent structural changes. In this effort, we focused on residues that were highly conserved and predicted by the simulations to be involved in hinge motions and placed these mutations into two groups, depending on whether their motions were correlated with nucleotide (Table 5.5). These are the same residues with unique φ-ψ behavior in Figure 5.9 and Figure 5.10.

We expressed and purified WT DnaK and each of the DnaK point mutants.[21] All of the purified proteins were properly folded as determined by thermal stability and circular dichroism measurements (Figure 5.13), suggesting that the alanine mutations did not damage the overall fold or stability of the chaperone.

Table 5.4. Conserved residues in subdomains II-A/II-B hinge region of DnaK NBD

Residue	Identity (*Conserved mutation*)		
	Bacteria (1714 Orthologs)	Animals (1388 Orthologs)	Plants (597 Orthologs)
I202	31% (*64%*)	5% (*90%*)	5% (*83%*)
S203	**90% (*9%*)**	**92% (*7%*)**	86% (*10%*)
E217	77% (*16%*)	87% (*9%*)	79% (*12%*)
A220	55% (*41%*)	48% (*51%*)	59% (*39%*)
T221	83% (*13%*)	**94% (*4%*)**	88% (*5%*)
N222	41% (*41%*)	14% (*11%*)	17% (*22%*)
G223	**98% (*1%*)**	**91% (*1%*)**	83% (*1%*)
D224	83% (*11%*)	**94% (*4%*)**	**91% (*5%*)**
T225	50% (*41%*)	89% (*2%*)	82% (*3%*)
H226	36% (*18%*)	83% (*4%*)	82% (*4%*)
L227	**97% (*2%*)**	**96% (*3%*)**	**92% (*7%*)**
G228	**99% (*1%*)**	**99% (*0%*)**	**98% (*1%*)**
G229	**100% (*0%*)**	**99% (*0%*)**	**99% (*1%*)**
E230	30% (*64%*)	87% (*8%*)	68% (*27%*)

Identity above 90% are highlighted

To assess the effects of the mutations on function, we measured the ATPase activity of the mutants using a malachite-green assay.[36] We also measured the ability of the chaperone to refold denatured firefly luciferase, which is an assay that is commonly used as a measure of chaperone functions *in vitro*.[46] In these experiments, we were interested in whether the mutants might have misregulated ATPase or refolding activities, which we expected would manifest as either an increase or decrease in comparison to WT DnaK. These two biochemical activities are linked to each other, but only indirectly,[21] so changes to either activity is expected to indicate a change in nucleotide-dependent allostery.

These experiments revealed that mutations I202A and L227A decreased ATPase activity by ~30% and refolding function by ~20%, suggesting that these highly conserved residues are important for allostery in DnaK. Mutation G223A had no effect on ATPase activity, but it did have a modest effect on refolding activity (~80% of WT). The S203A

Table 5.5. Activity of DnaK and its mutants

	ATPase Activity (μM P$_i$/μg DnaK/hr)	Fold Change	Luciferase Refolding Activity (%)
WT	11.3 ± 0.2	--	100 ± 12
	Hinge Residues Not Correlated to Nucleotide Motion		
I202A	7.4 ± 0.6	0.7	77 ± 10
S203A	20.8 ± 0.2	1.8	101 ± 12
L227A	8.4 ± 0.9	0.7	80 ± 14
	Hinge Residues Correlated to Nucleotide Motion		
G223A	11.5 ± 0.2	1.0	81 ± 14
G228A	18.5 ± 0.6	1.6	12 ± 3

mutant had normal refolding activity, but it had dramatically accelerated ATPase function (~180% of WT). Chemical modulators of Hsp70/DnaK produce similar effects on ATPase activity and these compounds produce dramatic effects on Hsp70 function in cells,[63] emphasizing the impact of the point mutations.

Finally, G228A had dramatically enhanced ATPase activity (~160% of WT), yet it had only 16% of the WT refolding activity. The results with G228A were particularly striking, supporting the idea that G228 is an important hinge residue. G228 is located between subdomain II-A and II-B, and it likely dictates the relative movement of the two subdomains, as observed in the 2nd mode of the essential dynamics of NBD (see Figure 5.6 B). It is situated between residue L227 and G228, which also show altered φ-ψ behavior based on the nucleotide-bound state in Figure 5.10. These experimental results confirmed the significant residues identified in the LD simulations, and they point to their involvement in DnaK allosteric control.
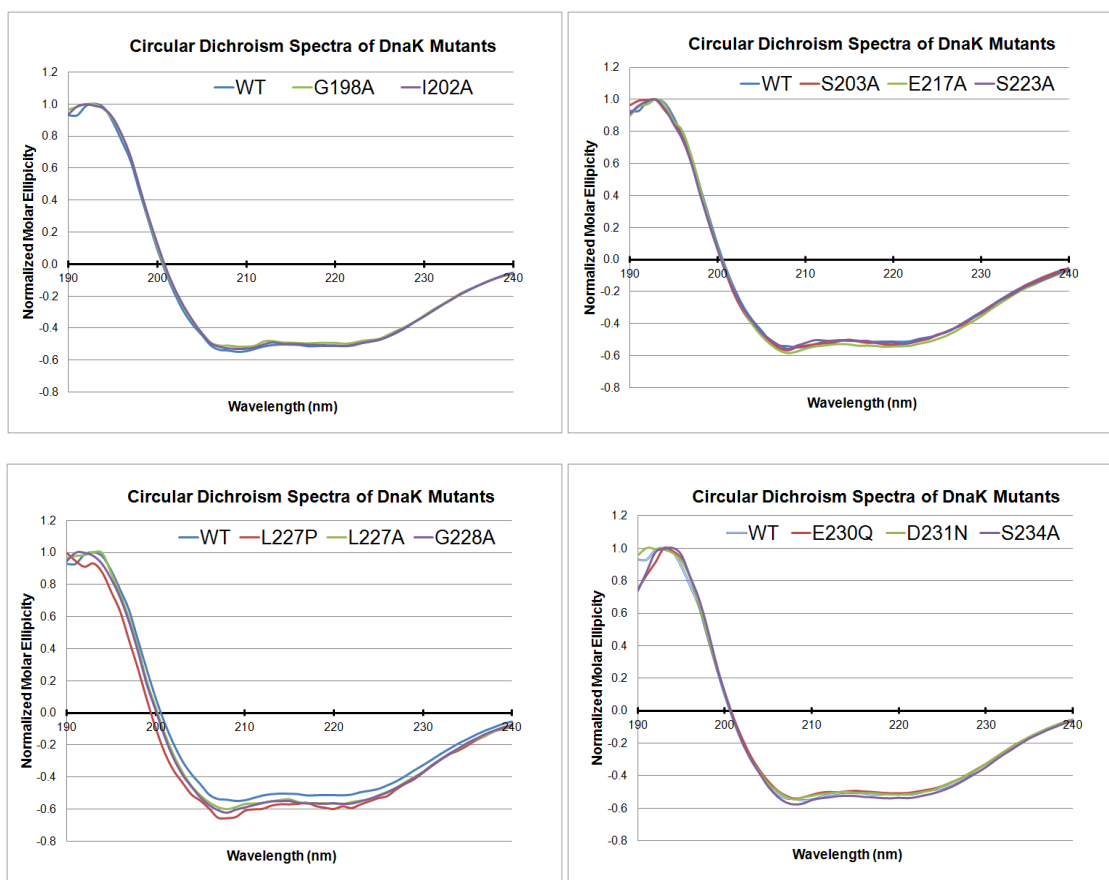
Figure 5.13. Normalized circular dichroism spectra of WT and mutant DnaK. All mutants have very similar fold as the WT DnaK.

5.5 Conclusion

Nucleotide-dependent structural transitions in DnaK are linked to its chaperone functions. While there is much known about the allostery between the NBD and SBD in DnaK,[17,22,64] there is less known about the key residues involved in the subdomain motions of the NBD. Using dynamics simulations of DnaK's NBD in the apo, ATP-, and ADP/P$_i$-bound states, we identified shear motions between subdomains I-A and II-A and a hinge motion that is consistent with recent reports.[18,19] Together, these motions result in a dramatic opening of the nucleotide-binding cleft, which is likely important for

209

nucleotide cycling and for communicating the nucleotide state to the SBD. Using the results of these simulations, we further identified residues, including S223, L227, G228, and G229, as being potentially important to the hinge motions. These residues are very highly conserved through evolution, and they are located at what appeared to be a critical hinge region. These results were supported by torsion angle analyses, which suggested that these residues adopt two distinct backbone conformations. For some of these residues, these clusters correlated with nucleotide state. Finally, mutation of these residues largely supported this model and showed that chaperone functions were misregulated when the hinge residues were mutated to alanine. Together, these studies point to specific hinge residues that are nearly invariant across all kingdoms of life and are important for the nucleotide-dependent motions in DnaK. These findings are important for understanding structure and function in DnaK.

5.6 Allosteric Ligand Binding Modes

Hsp70 is an attractive target for pharmaceutical development as it is linked to numerous disorders. High-throughput screenings have indentified small molecules that modulate Hsp70 catalytic activities.[39,46,65,66] Two compounds, myricetin and 115-7C, have been confirmed to be active modulators of DnaK. Interestingly, they are not nucleotide competitors and do not cross-compete with each other; they act through two novel and different allosteric mechanisms. We intend to indentify the binding modes of

these small molecules and apply that knowledge in future structure-based drug discovery on Hsp70.

Allosteric Site within Subdomain II-A

Wisén *et al.*[65] identified compound 115-7C (Figure 5.14 A) that positively modulated DnaK ATPase and refolding activities similar to the effect of the co-chaperone DnaJ. Interestingly, 115-7C does not compete but act in concert with DnaJ. NMR chemical shift data suggests that 115-7C interacts primarily with subdomain II-A of DnaK NBD (Figure 5.14 B). Simple molecular docking was unable to generate docking poses that match the NMR chemical shift data. All docking poses were placed on the surface of the receptor and did not possess the interactions that match the NMR data.

It is observed that many of the residues with 115-7C-induced chemical shifts are located in three adjunct secondary structural elements of subdomain II-A: β-sheets (R188, I190, V192, I204, I205, E206, I207, E217, L219, D336), an α-helix (E171, A176, A178,) and the random coils (T185 G186, L181, T215). These elements are intrinsically flexible as they have relatively high B-factors in crystal structures, e.g. 1DKG (Figure 5.14 C). Furthermore, the electron density of the random coils is unresolved in the crystal structure, suggesting the flexible nature of this region. Given the intrinsic flexibility of subdomain II-A, it is hypothesized that the gap between the α-helix and β-sheets of subdomain II-A may widen as part of the innate protein breathing motions and open up a cavity for ligand binding. Incidentally, conformations of DnaK NBD with widened α/β-gap (~ 7.0 Å expansion) have been observed in the LD simulations of NBD (Figure 5.14 D). To examine the docking of 115-7C to NBD with this expanded conformation,
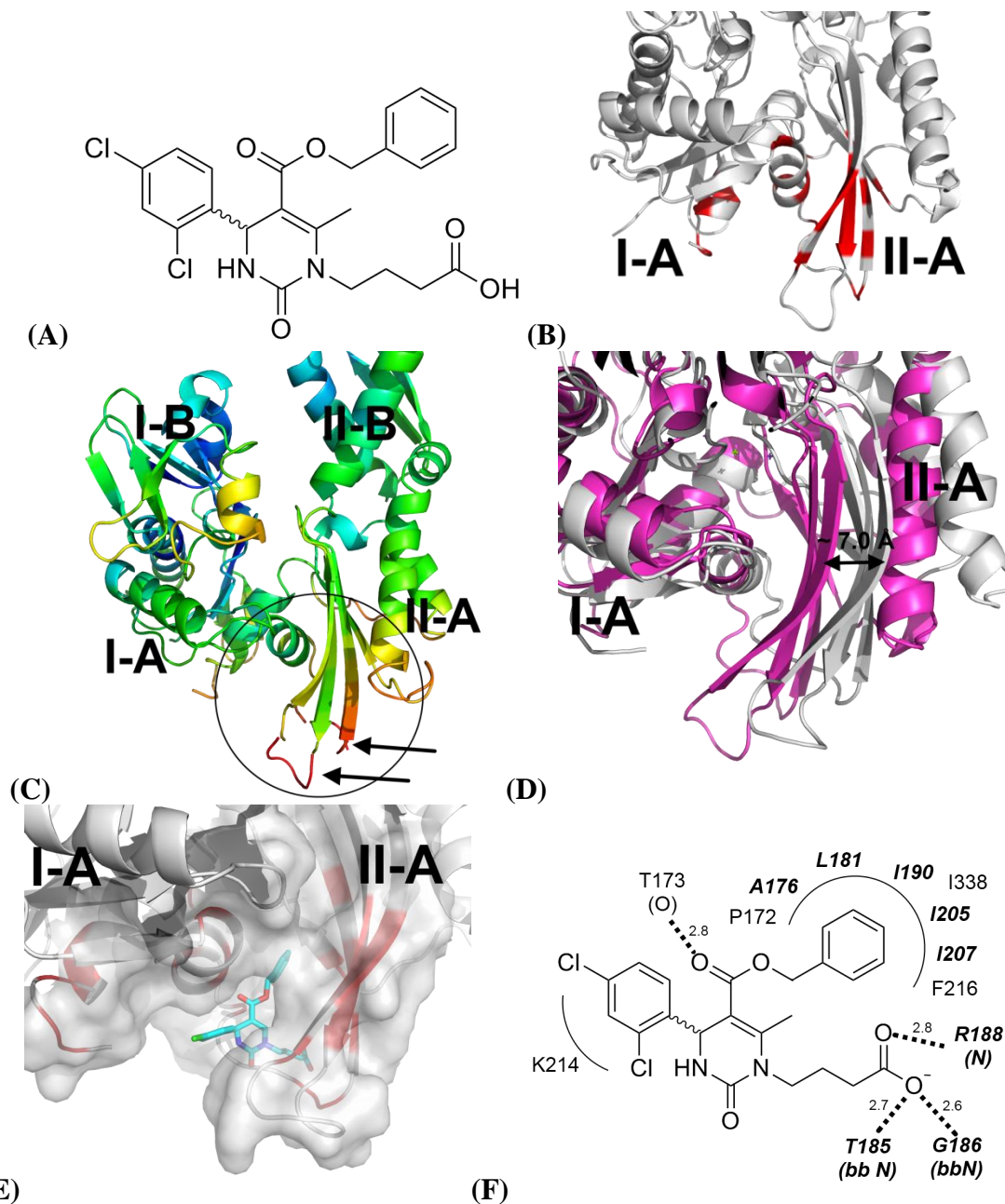
Figure 5.14. **(A)** Chemical structure of 115-7C. **(B)** DnaK residues with 115-7C-induced chemical shifts are colored in red. Majority of the shifts occurs on subdomain II-A. NMR data curtsey of Prof. Erik Zuiderweg. **(C)** B-factors of DnaK NBD (PDB:1DKG). Subdomain II-A (circled) is intrinsically more flexible as it has higher B-factors than other subdomains. Electron density of the random coils (arrows) is unresolved in the crystal structure. **(D)** Expansion of subdomain II-A observed in DnaK NBD simulations. NBD from crystal structure 1DKG and LD simulations are colored in purple and white, respectively. The gap between the α-helix and β-sheets can expand by as much as 7.0 Å. **(E)** Docking pose of 115-7C (R-enantiomer, Cyan) to subdomain II-A generated by Induced-Fit Docking. β-sheets were made flexible to allow expansion of the binding pocket. Residues with 115-7C-induced chemical shifts are colored in red. **(F)** Contacts of docked pose of 115-7C. Residues in bold have 115-7C-induced NMR shifts.

212

induced-fit docking (IFD) protocol available through Schrödinger suite of software was used to model 115-7C in the temporary cavity. Furthermore, conformation of the loop between D208 and T215 was explicitly modeled to allow additional expansion of cavity for docking.

Although simple rigid docking failed to dock 115-7C into the small cavity, IFD successfully docked 115-7C into a cavity within subdomain II-A (Figure 5.14 E). This docked pose has a number of favorable interactions: the ester linker of 115-7C interacts with the hydroxyl group of T173; the carboxylic acid moiety of 115-7C reaches deep into the pocket and forms salt bridges with R188 and the amide backbone of T185 and G186; placement of the phenyl moiety into a deep hydrophobic pocket composed of A176, L181, I190, I205, I207 and I338; dichlorophenyl moiety occupies the opening of the allosteric site and is solvent-exposed (Figure 5.14 F). This docked pose is strongly supported by NMR study as the interactions 115-7C made with DnaK NBD are consistent with the NMR chemical shift data.

It is known that J-domain proteins interact with Hsp70 primarily through the I-A/II-A subdomain interface.[67-69] The proximity of the J-protein interface to the proposed binding site of 115-7C and their synergetic DnaK activation suggest that alternation of 115-7C structure may have different effect on the DnaK-DnaJ interaction. Based on the proposed docked pose of 115-7C, a similar molecule 116-9C, which has a bulky biphenyl moiety instead of a dichlorophenyl, was examined. It was predicted that 116-9C would bind to the II-A allosteric site with similar docking pose; the phenyl and carboxylic acid groups anchor the molecule in the site while the biphenyl moiety would protrude out from the allosteric site. The protruded biphenyl group was predicted to place into the I-

A/II-A interface where J-proteins would bind and therefore interfere with DnaJ binding and stimulation. Experiments confirmed that 116-9C reversed the DnaK activating property of 115-7C and negatively modulated DnaJ-mediated stimulation,[65] confirming the binding pose of 115-7C.

5.7 Acknowledgments

5.8 Supplementary Information

<u>Primers for DnaK mutagenesis.</u>
STOP384 to Ala (TAA to GCA)
```
5' - GAC AAA AAA GCA AAG CTT GCG GCC GCA CTC G-- - 3'
5' - -GC GGC CGC AAG CTT TGC TTT TTT GTC TTT GAC - 3'
```

Gly198 to Ala (GGT to GCT)
```
5' - GGT GGT GCT ACT TTC GAT ATT TCT ATT ATC GAA ATC G-- - 3'
5' - -GA AAT ATC GAA AGT AGC ACC ACC CAG GTC ATA AAC CGC - 3'
```

Ile202 to Ala (ATT to GCT)
```
5' - -CT TTC GAT GCT TCT ATT ATC GAA ATC GAC GAA GTT GAC G-- - 3'
5' - --C GAT TTC GAT AAT AGA AGC ATC GAA AGT ACC ACC ACC CAG GTC - 3'
```

Ser203 to Ala (TCT to GCT)
```
5' - --C GAT ATT GCT ATT ATC GAA ATC GAC GAA GTT GAC GGC G-- - 3'
5' - --C GTC GAT TTC GAT AAT AGC AAT ATC GAA AGT ACC ACC ACC - 3'
```

Glu217 to Ala (GAA to GCA)
```
5' - GGC GAA AAA ACC TTC GCA GTT CTG GCA ACC AAC GG- - 3'
5' - -CC GTT GGT TGC CAG AAC TGC GAA GGT TTT TTC GCC - 3'
```

Gly223 to Ala (GGT to GCT)
```
5' - GTT CTG GCA ACC AAC GCT GAT ACC CAC CTG G-- - 3'
5' - --C CAG GTG GGT ATC AGC GTT GGT TGC CAG AAC - 3'
```

Leu227 to Ala (CTG to GCG)
```
5' - ACC AAC GGT GAT ACC CAC GCG GGT GGT GAA GAC TTC - 3'
5' - GAA GTC TTC ACC ACC CGC GTG GGT ATC ACC GTT GGT - 3'
```

Gly228 to Ala (GGT to GCT)
```
5' - GGT GAT ACC CAC CTG GCT GGT GAA GAC TTC G-- - 3'
5' - --C GAA GTC TTC ACC AGC CAG GTG GGT ATC ACC - 3'
```

Glu230 to Gln (GAA to CAG)
```
5' - GGT GGT CAG GAC TTC GAC AGC CGT CTG ATC - 3'
5' - --G GCT GTC GAA GTC CTG ACC ACC CAG GTG G-- - 3'
```

Asp231 to Asn (GAC to AAC)
```
5' - GGT GAA AAC TTC GAC AGC CGT CTG ATC AAC TAT CTG - 3'
5' - --G ACG GCT GTC GAA GTT TTC ACC ACC CAG G-- - 3'
```

Ser234 to Ala (AGC to GCC)
```
5' - --C TTC GAC GCC CGT CTG ATC AAC TAT CTG GTT GAA G-- - 3'
5' - GTT GAT CAG ACG GGC GTC GAA GTC TTC ACC ACC - 3'
```

5.9 References

1. Bukau, B.; Horwich, A. L. The Hsp70 and Hsp60 Chaperone Machines. *Cell* **1998**, *92*, 351-366.

2. McClellan, A. J.; Tam, S.; Kaganovich, D.; Frydman, J. Protein Quality Control: Chaperones Culling Corrupt Conformations. *Nat. Cell Biol.* **2005**, *7*, 736-741.

3. Albanèse, V.; Yam, A. Y.-W.; Baughman, J.; Parnot, C.; Frydman, J. Systems Analyses Reveal Two Chaperone Networks with Distinct Functions in Eukaryotic Cells. *Cell* **2006**, *124*, 75-88.

4. Kampinga, H. H.; Craig, E. A. The HSP70 Chaperone Machinery: J Proteins as Drivers of Functional Specificity. *Mol. Cell Biol.* **2010**, *11*, 579-592.

5. Ciocca, D. R.; Calderwood, S. K. Heat Shock Proteins in Cancer: Diagnosis, Prognostic, Predictive, and Treatment Implications. *Cell Stress Chaperones* **2005**, *10*, 86-103.

6. Powers, M. V.; Clarke, P. A.; Workman, P. Death by Chaperone: HSP90, HSP70, or Both? *Cell Cycle* **2009**, *8*, 518-526.

7. Evans, C. G.; Chang, L.; Gestwicki, J. E. Heat Shock Protein 70 (Hsp70) as an Emerging Drug Target. *J. Med. Chem.* **2010**, *53*, 4585-4602.

8. Patury, S.; Miyata, Y.; Gestwicki, J. E. Pharmacological Targeting of the Hsp70 Chaperone. *Curr. Top. Med. Chem.* **2009**, *9*, 1337-1351.

9. Bork, P.; Sander, C.; Valencia, A. An ATPase Domain Common to Prokaryotic Cell Cycle Proteins, Sugar Kinases, Actin and Hsp70 Heat Shock Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 7290-7294.

10. Schüler, H. ATPase Activity and Conformational Changes in the Regulation of Actin. *Biochim. Biophys. Acta.* **2001**, *1549*, 137-147.

11. Boorstein, W. R.; Ziegelhoffer, T.; Craig, E. A. Molecular Evolution of the HSP70 Multigene Family. *J. Mol. Evol.* **1994**, *38*, 1-17.

12. Mayer, M. P.; Brehmer, D.; Gassler, C. S.; Bukau, B. Hsp70 Chaperone Machines. *Adv. Protein Chem.* **2001**, *59*, 1-44.

13. Daugaard, M.; Rohde, M.; Jaattela, M. The Heat Shock Protein 70 Family: Highly Homologous Proteins with Overlapping and Distinct Functions. *FEBS Lett.* **2007**, *581*, 3702-3710.

14. Harrison, C. J.; Hayer-Hartl, M.; Di Liberto, M.; Hartl, F.; Kuriyan, J. Crystal Structure of the Nucleotide Exchange Factor GrpE Bound to the ATPase Domain of the Molecular Chaperone DnaK. *Science* **1997**, *276*, 431-435.

15. Sousa, M. C.; McKay, D. B. The Hydroxyl of Threonine 13 of the Bovine 70-kDa Heat Shock Cognate Protein Is Essential for Transducing the ATP-Induced Conformational Change. *Biochemistry* **1998**, *37*, 15392-15399.

16. O'Brien, M. C.; Flaherty, K. M.; McKay, D. B. Lysine 71 of the Chaperone Protein Hsc70 Is Essential for ATP Hydrolysis. *J. Biol. Mol.* **1996**, *271*, 15874-15878.

17. Bertelsen, E. B.; Chang, L.; Gestwicki, J. E.; Zuiderwig, E. R. P. Solution Conformation of wild-Type E. coli Hsp70 (DnaK) Chaperone Complexed with ADP and Substrate. *Proc. Nat. Am. Sci. U.S.A.* **2009**, *106*, 8471-8476.

18. Bhattacharya, A.; Kurochkin, A. V.; Yip, G. N. B.; Zhang, Y.; Bertelsen, E. B.; Zuiderweg, E. R. P. Allostery in Hsp70 Chaperones Is Transduced by Subdomain Rotations. *J. Mol. Biol.* **2009**, *388*, 475-490.

19. Zhuravleva, A.; Clerico, E. M.; Gierasch, L. M. An Interdomain Energetic Tug-of-War Creates the Allosterically Active State in Hsp70 Molecular Chaperones. *Cell* **2011**, *151*, 1296-1307.

20. Kityk, R.; Kopp, J.; Sinning, I.; Mayer, M. P. Structure and Dynamics of the ATP-bound Open Conformation of Hsp70 Chaperones. *Mol. Cell* **2012**, *48*, 863-874.

21. Chang, L.; Thompson, A.; Ung, P.; Carlson, H. A.; Gestwicki, J. E. Mutagenesis Reveals the Complex Relationships between ATPase Rate and the Chaperone Activities of *Escherichia coli* Heat Shock Protein 70 (Hsp70/DnaK). *J. Biol. Chem.* **2010**, *285*, 21282-21291.

22. Nicolaï, A.; Senet, P.; Delarue, P.; Ripoll, D. R. Human Inducible Hsp70: Structures, Dynamics, and Interdomain Communication from All-Atom Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2010**, *6*, 2501-2519.

23. Gołaś, E.; Maisuradze, G. G.; Senet, P.; Ołdziej, S.; Czaplewski, C.; Scheraga, H. A.; Liwo, A. Simulation of the Opening and Closing of Hsp70 Chaperones by Coarse-Grained Molecular Dynamics. *J. Chem. Theory Comput.* **2012**, *8*, 1750-1764.

24. Liu, Y.; Gierasch, L. M.; Bahar, I. Role of Hsc70 ATPase Domain Intrinsic Dynamics and Sequence Evolution in Enabling its Functional Interactions with NEFs. *PLoS Comput. Biol.* **2010**, *6*, e1000931.

25. Molecular Operating Environment (MOE), 2005.06; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, **2005**.

26. Case, D. A.; Darden, T. A.; Cheatham, III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvary, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER 10*, University of California, San Francisco.: **2008**.

27. Palleros, D. R.; Reid, K. L.; Shi, L.; Welch, W. J.; Fink, A. L. ATP-Induced Protein Hsp70 Complex Dissociation Requires K+ But Not ATP Hydrolysis. *Nature* **1993**, *365*, 664-666.

28. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins* **2006**, *65*, 712-725.

29. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157-1174.

30. Wang, J.; Wang, W.; Kollman P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247-260.

31. Jakalian, A; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comp. Chem.* **2000**, *21*, 132-146.

32. Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comp. Chem.* **2002**, *23*, 1623-1641.

33. Meagher, K. L.; Redman, L. T.; Carlson, H. A. Development of Polyphosphate Parameters for Use with the AMBER Force Field. *J. Comput. Chem.* **2003**, *24*, 1016-1025.

34. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of *n*-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327-341.

35. Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins* **2004**, *55*, 383-394.

36. Buczynski, G.; Slepenkov, S. V.; Sehorn, M. G.; Witt, S. N. Characterization of a Lidless Form of the Molecular Chaperone DnaK. *J. Biol. Chem.* **2001**, *276*, 27231-27236.

37. Stols, L.; Gu, M.; Dieckman, L.; Raffen, R.; Collart, F. R.; Donnelly, M. I. A New Vector for High-Throughput, Ligation-Independent Cloning Encoding a Tabacco Etch Virus Protease Cleavage Site. *Prot. Expres. Purif.* **2002**, *25*, 8-15.

38. Zheng, L.; Baumann, U.; Reymond, J.-L. An Efficient One-Step Site-Directed and Site-Saturation Mutagenesis Protocol. *Nucleic Acids Res.* **2004**, *32*, e115.

39. Chang, L.; Bertelsen, E. B.; Wisén, S.; Larsen, E. M.; Zuiderweg, E. R. P.; Gestwicki, J. E. High-Throughput Screen for Small Molecules That Modulate the

ATPase Activity of the Molecular Chaperone DnaK. *Anal. Biochem.* **2008**, *372*, 167-176.

40.   Pantoliano, M. W.; Petrella, E. C.; Kwasnoski, J. D.; Lobanov, V. S.; Myslik, J.; Graf, E.; Carver, T.; Asel, E.; Springer, B. A.; Lane, P.; Salemme, F. R. High-Density Miniaturized Thermal Shift Assays as a General Strategy for Drug Discovery. *J. Biomol. Screen.* **2001**, *6*, 429-440.

41.   Cummings, M. D.; Famum, M. A.; Nelen, M. I. Universal Screening Methods and Applications of ThermoFluor. *J. Biomol. Screen.* **2006**, *11*, 854-863.

42.   Kranz, J. K.; Schalk-Hihi, C. Protein Thermal Shifts to Identify Low Molecular Weight Fragments. *Methods Enzymol.* **2011**, *493*, 277-298.

43.   Van Veldhoven, P. P.; Mannaerts, G. P. Inorganic and Organic Phosphate Measurements in the Nanomolar Range. *Anal. Biochem.* **1987**, *161*, 45-48.

44.   Ekman, P.; Jager, O. Quantification of Subnanomolar Amount of Phosphate Bound to Seryl and Threonyl Residues in Phosphoproteins Using Alkaline Hydrolysis and Malachite Green. *Anal. Biochem.* **1993**, *214*, 138-141.

45.   Rowlands, M. G.; Newbatt, Y. M.; Prodromou, C.; Pearl, L. H.; Workman, P.; Aherne, W. High-Throughput Screening Assay for Inhibitors of Heat-Shock Protein 90 ATPase Activity. *Anal. Biochem.* **2004**, *327*, 176-183.

46.   Wisén, S.; Gestwicki, J. E. Identification of Small Molecules that Modify the Protein Folding Activity of Heat Shock Protein 70. *Anal. Biochem.* **2008**, *374*, 371-377.

47.   Sherman, W.; Day, T.; Jacobson, M. P. Friesner, R. A.; Farid, R. Novel Procedure for Modeling Ligand/Receptor Induced Fit Effects. *J. Med. Chem.* **2006**, *49*, 534-553.

48.   Schrödinger Suite (**2009**) Induced Fit Docking Protocol; Glide version 5.5, Schrödinger, LLC, New York, NY, 2009; Prime version 2.1, Schrödinger, LLC, New York, NY, 2009.

49.   Chiappori, F.; Merelli, I.; Colombo, G.; Milanesi, L.; Morra, G. Molecular Mechanism of Allosteric Communication in Hsp70 Revealed by Molecular Dynamics Simulations. *PLOS Comput. Biol.* **2012**, *8*, e1002844.

50.   Woo, H.-J.; Jiang, J.; Lafer, E. M.; Sousa, R. ATP-Induced Conformational Changes in Hsp70: Molecular Dynamics and Experimental Validation of an in Silico Predicted Conformation. *Biochemistry* **2009**, *48*, 11470-11477.

51.   Russell, R.; Jordan, R.; McMacken, R. Kinetic Characterization of the ATPase Cycle of the DnaK Molecular Chaperone. *Biochemistry* **1998**, *37*, 596-607.

52.   Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential Dynamics of Proteins. *Proteins* **1993**, *17*, 412-425.

53. Abseher, R.; Nilges, M. Are There Non-Trivial Dynamic Cross-Correlations in Proteins. *J. Mol. Biol.*, **1998**, *279*, 911-920.

54. Skjaerven, L.; Martinez, A.; Reuter, N. Principle Component and Normal Mode Analysis of Proteins: A Quantitative Comparison Using the GroEL Subunit. *Proteins* **2011**, *79*, 232-243.

55. Gerstein, M.; Lesk, A. M.; Chothia, C. Structural Mechanisms for Domain Movements in Proteins. *Biochemistry* **1994**, *33*, 6739-6749.

56. Page, R.; Lindberg, U.; Schutt, C. E. Domain Motions in Actin. *J. Mol. Biol.* **1998**, *280*, 463-474.

57. Janin, J.; Wodak, S. J. Structural Domains in Proteins and Their Roles in the Dynamics of Protein Function. *Prog. Biophys. Mol. Biol.* **1983**, *42*, 21-78.

58. Bhaskara, R. M.; Srinivasan, N. Stability of Domain Structures in Multi-Domain Proteins. *Sci. Rep.* **2011**, *1*, 40.

59. Hovmöller, S.; Zhou, T.; Ohlson, T. Conformations of Amino Acids in Proteins. *Acta Cryst.* **2002**, *D58*, 768-776.

60. Chang, Y.-W.; Sun ,Y.-J.; Wang, C.; Hsiao, C.-D. Crystal Structures of the 70-kDa Heat Shock Proteins in Domain Disjoining Conformation. *J. Biol. Chem.* **2008**, *283*, 15502-15511.

61. Vogel, M.; Bukau, B.; Mayer, M. P. Allosteric Regulation of Hsp70 Chaperones by a Proline Switch. *Mol. Cell* **2006**, *21*, 359-367.

62. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25*, 3389-3402.

63. Chafekar, S. M.; Wisén, S.; Thompson, A. D.; Echeverria, A.; Walter, G. M.; Evans, C. G.; Makley, L. N.; Gestwicki, J. E.; Duennwald, M. L. Pharmacological Tuning of Heat Shock Protein 70 Modulates Polyglutamine Toxicity and Aggregation. *ACS Chem. Biol.* **2012**, *7*, 1556-1564.

64. Swain, J. F.; Dinler, G.; Sivendran, R.; Montgomery, D. L.; Stotz, M.; Gierasch, L. M. Hsp70 Chaperone Ligands Control Domain Association via an Allosteric Mechanism Mediated by the Interdomain Linker. *Mol. Cell* **2007**, *26*, 27-39.

65. Wisén, S.; Bertelsen, E. B; Thompson, A. D.; Patury, S.; Ung, P.; Chang, L.; Evans, C. G.; Walter, G. M.; Wipf, P.; Carlson, H. A.; Brodsky, J. L.; Zuiderweg, E. R. P.; Gestwicki, J. E. Binding of a Small Molecule at a Protein-Protein Interface Regulates the Chaperone Activity of Hsp70-Hsp40. *Chem. Biol.* **2010**, *5*, 611-622.

66. Chang, L.; Miyata, Y.; Ung, P. M.-U.; Bertelsen, E.B.; McQuade, T. J.; Carlson, H. A.; Zuiderweg, E. R. P.; Gestwicki, J. E. Chemical Screens Against a Reconstituted

Multiprotein complex: Myricetin Blocks DnaJ Regulation of DnaK through an Allosteric Mechanism. *Chem. Biol.* **2011**, *18*, 210-221.

67. Gässler, C. S.; Buchberger, A.; Laufen, T.; Mayer, M. P.; Schröder, H.; Valencia, A.; Bukau, B. Mutations in the DnaK Chaperone Affecting Interaction with the DnaJ Cochaperone. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 15229-15234.

68. Suh, W. C.; Burkholder, W. F.; Lu, C. Z.; Zhao, X.; Gottesman, M. E.; Gross, C. A. Interaction of the Hsp70 Molecular Chaperone, DnaK, with Its Co-chaperone DnaJ. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 15223-15228.

69. Ahmad, A.; Bhattacharya, A.; McDonald, R. A.; Cordes, M.; Ellington, B.; Bertelsen, E. B.; Zuiderweg, E. R. P. Heat Shock Protein 70 kDa Chaperone/DnaJ Cochaperone Complex Employs an Unusual Dynamic Interface. *Proc. Nat. Acad. Sci. U.S.A.* **2011**, *108*, 18966-18971.

CHAPTER 6

Community Structure-Activity Resource

6.1 Abstract

The Community Structure Activity Resource (CSAR) is an effort by the drug design community to improve the methods for docking and scoring. It involves the collection of high-quality data of protein-ligand complexes from industry and academia. This collected data would be used in a community-wide exercise to assess the current state of docking and scoring. The protein-ligand complexes used in the first CSAR data set are PDB entries that have binding data (K_d or K_i) in Binding MOAD and PDBind. The finalized data set consists of 343 diverse protein-ligand complexes and a wide range of protein-ligand binding affinity that spans $\sim$ 14 pK_d. It is recognized that different experimental conditions and inherent experimental errors may limit the possible correlation between scores and measured affinities. Many scoring approaches were assessed in the exercise and their correlation to experimental binding affinity ranged $R^2 =$ 0.14-0.51, yet no particular scoring approach had overwhelming advantage over others.

This suggests a need for CSAR to develop data sets of congeneric series with a range of different physical characteristics for improved training of computational methods.

6.2 Introduction

In the quest of finding drug-like molecules from a vast chemical space, academics and the pharmaceutical industry have been investing in time and resources in developing applications for mining the available data and speeding up the research process. Experimentally, combinatorial chemistry and high-throughput screening were developed in the past decade,[1] while structure-based drug discovery (SBDD) methods have been hotly pursued in the past two decades and are instrumental in the research and development of many pharmaceutically active compounds.[2,3]

Of the many techniques utilized in SBDD, docking and scoring are among the most widely applied. In particular, scoring algorithms play a pivotal role in the success of SBDD because it evaluates the ligand-receptor interactions, a key term that guides the docking of a ligand and the subsequent ranking of the docked ligand poses. However, many of these scoring algorithms were developed independently, and unbiased evaluations and comparisons of their accuracy are difficult to achieve. In order to improve the strengths and reduce the limitations of scoring algorithms, data sets that are used in the development and training of these methods should be as diverse and accurate as possible. In light of these requirements, the Community Structure-Activity Resource (CSAR) was developed.[4,5] One of the goals of the CSAR project was to provide a

diverse, high-quality protein-ligand complex data sets for the drug design community as a standardized platform for testing and development of docking and scoring. The CSAR Center was charged with the organization of a community-wide scoring exercise so that developers of the scoring algorithms can contribute their results to this community effort. The CSAR project analyzed the results from this benchmark exercise and documented the problematic complexes that require new approaches.[4] The study also led to the identification of the need of custom data sets of specific ligands and targets aspects for future CSAR development.[5]

6.3 Computational Methods

Scoring Ligand-Protein Complexes

After filtering PDB entries with properties that fit the CSAR data set requirements (resolution < 2.5 Å, diffraction-data precision indicator < 0.5, binary protein-ligand complex, ligand without internal strain, etc), entries with ligand binding data (K_d, K_i or K_a) were selected (Figure 6.1) for further assessment. Proteins and ligands were inspected with MOE[6] and PyMOL[7] for clashes and missing side chains in the vicinity of the ligand-binding site. AM1-BCC partial charges[8,9] were assigned to ligands through QuacPac version 1.3.1,[10] part of the OpenEye suite of software. All proteins and ligands, including metal ions, were parameterized with the default settings according to the docking/rescoring software.

AutoDock 4.2.3.[11] A force field-based function, default Gasteiger partial charges were used in the rescoring. Parameterization was performed with AutoDockTools 1.5.4.

AutoDock Vina 1.1.1.[12] It uses a knowledge-based scoring function. Default parameters from AutoDockTools 1.5.4 were used.

DrugScore 1.1.[13] Default setting was used for this knowledge-based function.

FRED 2.2.5.[14] Grid box was set to extend 4 Å beyond the ligand on each dimension. The ligand-receptor complexes were evaluated with these scoring functions: ShapeGauss,[15] ChemGauss3, ChemScore,[16] OEChemScore 1.4.2,[14] PLP,[17] and ScreenScore.[18]

Glide 5.0.[19] Proteins and ligands were parameterized with Maestro.[20] As GlideScore[21] has two related scoring settings, the Standard-Precision (SP) and Extreme-Precision (XP), scoring was performed in both settings.

M-Score.[22] Default setting was used for this knowledge-based function.

X-Score.[23] Default setting was used. The reported score was an average of three scores (HPScore, HMScore, and HSScore) calculated by X-Score.

6.4 Results and Discussion

Development of CSAR Data Sets

In our initial development of the protein-ligand data sets, we obtained crystal structures from several public domains: Protein Data Bank,[24] Binding MOAD,[25] and PDBind.[26] These protein entries were filtered to select high-quality (resolution better than 2.5 Å), binary protein-ligand complexes with binding data (K_d, K_i, K_a). The

development of CSAR's first data set until September 2010 is summarized in Figure 6.1. This initial selection resulted in a list of 539 PDB entries, which was divided into two sets: 242 entries in SET 1, and 297 entries in SET 2. The binding affinity of the ligand-protein complexes spans a wide range, ~ 14 pK_a, providing diverse samples for the evaluation of scoring algorithms.

Based on the feedback from participants of the CSAR benchmark exercise, the assessments of the protein-ligand complexes were expanded in the Remediated Release. Crystal structures were checked for poor real-space correlation coefficient. Non-standard amino acids were examined. Histidine (HIS) residues in the binding site ($\leq$ 6.0 Å from bound ligand) were examined and renamed to HIP or HIE to designate the adequate tautomer state, while non-binding site HIS were renamed to HID. The ligands were checked for significant contacts with the symmetry partner ($\leq$ 4.0 Å) to avoid crystal-packing effects. These re-assessments led to the release of the remediated CSAR-HiQuality data set in October 2009, which contained 343 protein-ligand complex structures. Later, our colleagues at the National Research Council of Canada (NRC) minimized the protein-ligand complexes and improved the protonation and tautomerization states of certain residues. The CSAR-NRC Release was published in September 2010.

In the July 2009 Initial Release, the accuracy of the reported binding affinities data (K_d or K_i) of the selected entries was confirmed by manually checking with the original articles. If provided, the margin of error of the binding affinity, binding assay and crystallization conditions were also noted. It is found that the binding affinity data was collected through a range of experimental methods: biochemical assay, fluorescence,

isothermal calorimetry, chromatography, stop-flow kinetics, and surface plasmon resonance. The accuracy of each of these methods may differ significantly; methods such as biochemical or fluorescence assay can have up to > 20% experimental error. Furthermore, in many incidents the affinity assay and crystallization conditions differ, most notably the pH. Such alteration in experimental conditions may affect the protein-ligand interactions, leading to conformational changes in the binding pocket and the ligands. These challenges may complicate the CSAR data set as they may limit the possible correlation between scores and measured affinities.

Evaluation of Scoring Functions

With the release of the September 2010 remediated CSAR-NRC protein-ligand complex data set, we evaluated different scoring algorithms, using this standardized platform. Currently, there are over 10 different scoring algorithms widely used and being actively developed in various research groups around the world. They can be divided into three categories: force-field based potentials (AutoDock 4.2.3, DOCK 4.0,[27] MedusaScore,[28] MOE,[29] and SIE[30]) that are derived from molecular mechanics force fields such as AMBER,[31] CHARMM,[32] and OPLS-AA;[33] empirical potentials (eHiTS,[34] FRED-ChemGauss3, GlideScore, Lead Finder,[35] and S2[36]) that use scaled, additive components to approximate binding free energy; and knowledge based potentials (AutoDock Vina, DrugScore, Gold-GhemScore,[37,38] ITScore,[39] M-Score, and X-Score) that statistically fit pairwise interactions to a large data set of crystallographic data.

We generated rescoring results with several algorithms in-house (AutoDock 4.2.3, AutoDock Vina, DrugScore, FRED-ChemGauss3, GlideScore, M-Score, MOE, and X-

227

Score) to understand their performance. We found that Glide required extensive protein preparation before scoring; thus, Glide docking preparation was also used as a tool for detecting protein irregularity, such as the presence of non-standard amino acids, metal ions, clashing atoms, etc.

We received AutoDock 4.2.3 scores from Garrett Morris for comparison; Morris is a major developer of the AutoDock source codes with Art Olson at the Scripps Research Institute. We found that our in-house scoring using AutoDock 4.2.3 and the latest AutoDockTools, version 1.5.4, generated results that are more correlated to the experimental values than Morris'. The discrepancy in the scoring results may be the result of using an older version of AutoDockTools, which could generate inconsistent parameters for the proteins and ligands.

The presence of crystallographic waters in the ligand binding pocket may have significant impact on docking and scoring accuracy. Although many docking and scoring algorithms are designed to perform without the presence of water molecules in the binding pocket, some allow the inclusion of waters. We examined the effect of including crystallographic water molecules in the scoring process with two algorithms, Glide and FRED, which allow the inclusion of crystallographic waters in the rescoring workflow. However, the inclusion of water did not improve the correlation of the scores and the binding affinity data. As shown in Table 6.1, some of the scoring results deteriorated with the inclusion of water. This discrepancy may arise due to two factors: (*1*) uncertainties of the presence and the coordinates of water molecules included in the crystal structures, and (*2*) insufficient parameterization of the scoring algorithm when water molecules are included. Hence, to maintain the consistency of the rescoring protocols and results, no

228
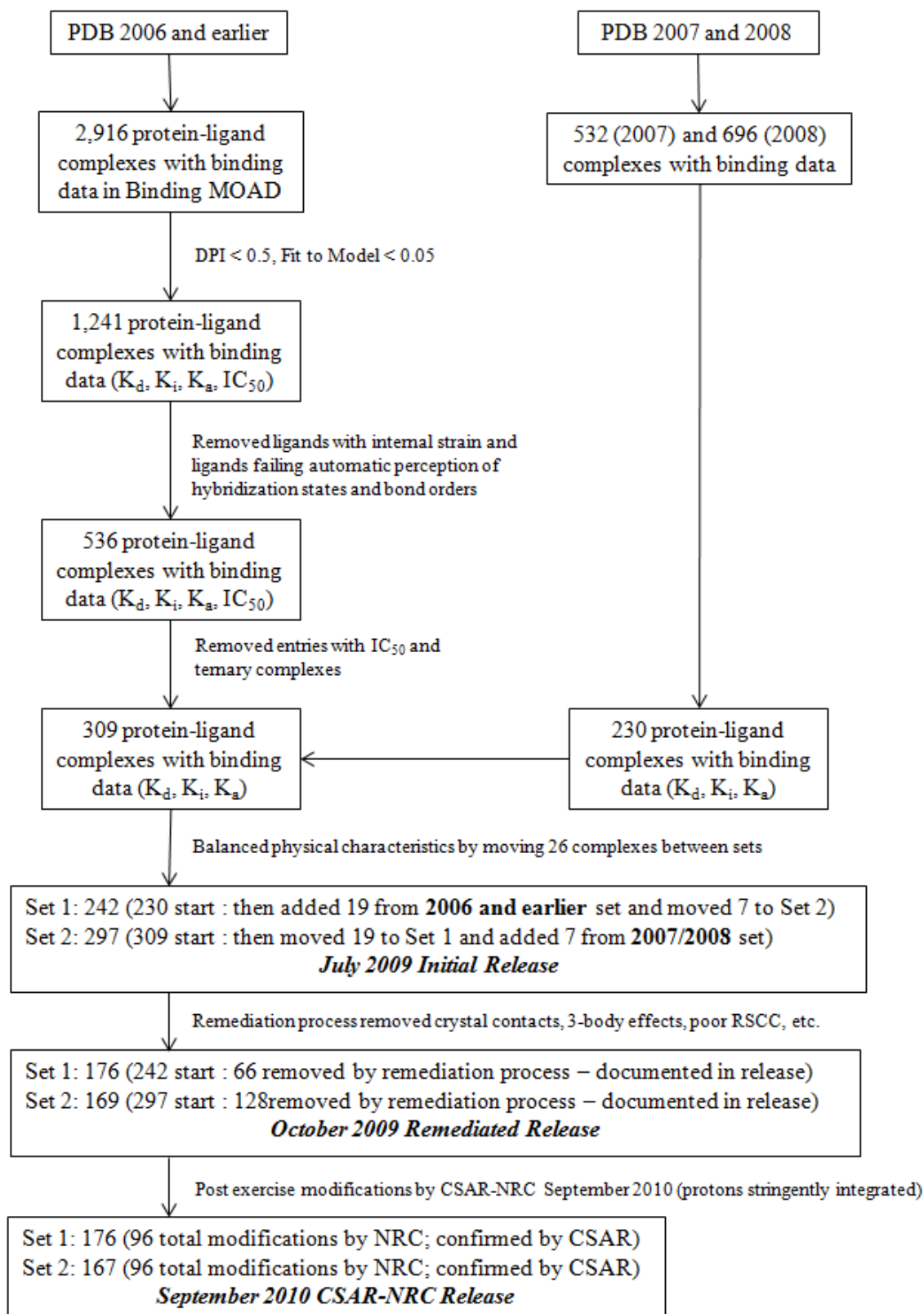
Figure 6.1. Flow chart of the development of the CSAR data sets.

Table 6.1. Comparison of scoring results with and without inclusion of crystallographic waters

	Without Waters					
	Glide-SP	Glide-XP	FRED-CG3	FRED-CS	FRED-PLP	FRED-ScrS
R^2	0.336	0.251	0.418	0.207	0.399	0.401
Pearson r	-0.580	-0.501	-0.646	-0.455	-0.632	-0.633
	With Waters					
	Glide-SP	Glide-XP	FRED-CG3	FRED-CS	FRED-PLP	FRED-ScrS
R^2	0.360	0.150	0.405	0.105	0.182	0.336
Pearson r	-0.600	-0.388	-0.636	-0.324	-0.427	-0.580

crystallographic water was included in the protein setup for scoring in the comparison across all methods.

Scoring results on the CSAR-NRC HiQuality data set are shown in Table 6.2. Glide provides two scoring schemes, GlideScore-XP (extreme precision) and GlideScore-SP (standard precision). GlideScore-SP is less stringent than GideScore-XP as it "softens" the repulsive term in the Lennard-Jones potential calculations. GlideScore-XP (extreme precision) did not perform as well as its less stringent SP version, likely due to the significant penalty of atomic clashing.

We observed that with this data set, knowledge-based scoring algorithms (AutoDock Vina, X-Score, DrugScore) performed marginally better than force-field based and empirical scoring algorithms. It is noted that the scoring algorithms used by X-Score and AutoDock Vina were parameterized with the PDBind data set, one of the input data sets used in the creation of the CAAR-NRC HiQuality data set. Hence, the increase in scoring performance may be the result from the use of overlapping data sets used in the training of the scoring algorithms and this evaluation.

As summarized in Table 6.2, scoring algorithms assessed in this CSAR benchmark exercise have a wide range of correlation to the experimentally determined

ligand binding affinity, $R^2 = 0.14 - 0.51$. The average correlation of SET 1 and SET 2 are $R^2 = 0.28$ and $R^2 = 0.37$, respectively. Each scoring algorithm had examples of high and low correlation to the experimental; however, the variation in the correlation was relatively insignificant as no particular scoring algorithm had overwhelming advantage over others. The outliers of this observation are GlideScore-XP and FRED-ChemScore, both performed significantly worse than other algorithms within the same data set.

For GlideScore, its performance in SET 1 and SET 2 are $R^2 = 0.14$ and $R^2 = 0.25$, respectively, lower than the average of the sets. This may be due to XP's use of "hard" repulsive term in the Lennard-Jones potential calculation, which penalizes any atomic clash between the protein and ligand. This problem is especially significant for GlideScore-XP since the protein-ligand complexes were not specifically optimized for Glide, which requires minimization of residue side chains with OPLS-AA force-field prior to rescoring. The lack of Glide-specific optimization was done out of necessity to maintain the same coordinates used for all methods.

FRED-ChemScore also underperformed in the CSAR data sets; SET 1 and SET 2 correlations are well below the average, at $R^2 = 0.16$ and $R^2 = 0.21$, respectively. ChemScore is an empirical scoring function composed of several additive terms: binding energy, clash penalty, and internal torsion terms. FRED-ChemScore may suffer from the clash penalty as the structure of the complexes was not optimized, similar to the issue faced by GlideScore-XP. It is difficult to claim these scoring functions have in sufficiency when modifications of the setup would lead to better correlations. This underscores a trend identified during this work: when groups did careful setup modifications, specific to their approach, their correlations to experiment significantly

Table 6.2. Statistics of In-house Rescoring on CSAR-NRC-HiQuality Datasets.

SET 1

	-logK	AD42	AD-Vina	Glide-SP	Glide-XP	Drug-Score	M-Score	X-Score	FRED-CG3	FRED-CS	FRED-PLP	FRED-ScrS
mean	6.24	-9.08	-10.9	-9.03	-9.60	-446575	-289	6.38	-84.96	-28.6	-66.8	-127
med	6.22	-8.95	-10.6	-8.98	-9.30	-422041	-300	6.17	-81.77	-29.0	-67.8	-125
min	-0.15	-21.9	-22.8	-16.7	-19.61	-1078200	-558	3.75	-162.7	-61.1	-129.8	-314
max	13.00	-1.86	-2.31	-2.83	-2.93	-118460	-27	9.79	-14.54	-5.00	-12.6	-19.3
sig	2.28	3.53	4.48	2.49	3.13	196475	120	1.47	33.57	10.4	24.7	57.2
R^2		0.25	0.27	0.35	0.14	0.30	0.22	0.41	0.36	0.16	0.30	0.31
Pearson		-0.50	-0.52	-0.59	-0.37	-0.54	-0.47	0.64	-0.60	-0.39	-0.55	-0.56
Y-int		3.31	3.38	1.37	3.73	3.42	3.68	-0.13	2.77	3.78	2.85	3.41
slope		-0.32	-0.26	-0.54	-0.27	0.00	-0.01	1.00	-0.041	-0.09	-0.05	-0.02

SET 2

	-logK	AD42	AD-Vina	Glide-SP	Glide-XP	Drug-Score	M-Score	X-Score	FRED-CG3	FRED-CS	FRED-PLP	FRED-ScrS
mean	6.24	-8.59	-10.6	-8.81	-9.63	-440247	-292	6.21	-82.75	-28.9	-65.0	-127
med	6.22	-8.45	-9.65	-8.69	-9.71	-393191	-286	5.89	-78.89	-29.1	-61.5	-124
min	-0.15	-15.6	-23.7	-15.9	-17.85	-1058600	-610	3.59	-168.8	-56.5	-122.5	-255
max	13.00	-1.97	-0.22	3.07	-2.57	-110360	-22	9.63	6.31	-3.85	7.8	-2.3
sig	2.28	3.02	4.55	2.70	2.94	201552	124	1.47	34.79	10.1	23.3	53.4
R^2		0.41	0.44	0.34	0.25	0.44	0.27	0.51	0.42	0.21	0.40	0.40
Pearson		-0.64	-0.66	-0.58	-0.50	-0.66	-0.52	0.71	-0.65	-0.46	-0.63	-0.63
Y-int		2.12	2.71	1.96	2.53	2.92	3.39	-0.48	2.73	3.23	2.24	2.79
slope		-0.46	-0.32	-0.47	-0.37	0.00	-0.01	1.05	-0.04	-0.10	-0.06	-0.03

AD, AutoDock; SP, standard precision; XP, Extreme precision; CS3, ChemGauss3; CS, ChemScore, ScrS, ScreenScore

improved. This altered coordinates from their custom setup yielded **NO** improvement when used with other methods.

In summary, the scoring data on the CSAR data set suggests that most scoring algorithms have low correlation to the experimental binding affinity data, as evident by the weak R^2 (< 0.5). The complete study, as described by Smith *et al.*,[5] involved a full list of 19 different methods (with the addition of DOCK 4, eHiTS, Gold-ChemScore, ITScore, Lead Finder, MedusaScore, S2, and SIE). The authors found that the correlation to experiment for the various methods ranged $R^2 = 0.12 - 0.58$, similar to the range ($R^2 = 0.14 - 0.51$) seen with 11 scoring algorithms (Table 6.2). No one particular algorithm has decisive advantage over others. Of the 343 complexes in the CSAR data set, 63 complexes were identified to be scored poorly by most scoring algorithms and 123 complexes were scored well by most. Comparison of these two sets of complexes found that the major differences came from the hydrogen bonding between proteins and ligands and torsional strain of ligands. These results suggest that future development of the scoring algorithms would need to improve the accuracy of the hydrogen-bonding and torisonal-strain terms. Furthermore, future development of CSAR should include data sets of congeneric series with a range of hydrogen-bonding and hydrophobic characteristics and a range of rotatable bonds.

6.5 Conclusion

The development of CSAR's data sets receives contributions from many different scientists from the computational modeling and crystallographic communities. Feedback, input, and remediation from users have helped shape the development and goals of CSAR. Participants of this CSAR exercise performed various scoring approaches on this comprehensive data set and the scoring results were collected and analyzed to determine the current state of the scoring algorithms. The knowledge gained from this exercise has helped shape the future development and approach to computational drug design.

6.6 Acknowledgements

6.7 References

1. Macarron, R. Critical Review of the Role of HTS in Drug Discovery. *Drug Dis. Today* **2006**, *11*, 277-279.

2. Congreve, M.; Murray, C. W.; Blundell, T. L. Keynote Review: Structural Biology and Drug Discovery. *Drug Dis. Today* **2005**, *10*, 895-907.

3. Scapin, G. Structural Biology and Drug Discovery. *Curr. Pharm. Des.* **2006**, *12*, 2087-2097.

4. Dunbar, J. B.; Smith, R. D.; Yang, C.-Y.; Ung, P. M.-U.; Lexa, K. L.; Khazanov, N. A.; Stuckey, J. A.; Wang, S.; Carlson, H. A. CSAR Benchmark Exercise of 2010: Selection of the Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2011**, *51*, 2036-2046.

5. Smith, R. D.; Dunbar, J. B.; Ung, P. M.-U.; Esposito, E. X.; Yang, C.-Y.; Wang, S.; Carlson, H. A. CSAR Benchmark Exercise of 2010: Combined Evaluation Across All Submitted Scoring Functions. *J. Chem. Inf. Model.* **2011**, *51*, 2115-2131.

6. Molecular Operating Environment (MOE), version 2010.10; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, **2010**.

7. PyMOL Molecular Graphics System, Version 1.5.0.1 Schrödinger, LLC.

8. Jakalian, A; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comp. Chem.* **2000**, *21*, 132-146.

9. Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comp. Chem.* **2002**, *23*, 1623-1641.

10. QuacPac version 1.3.1, OpenEye Scientific Software, Inc., Santa Fe, NM, **2008**.

11. Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olsen, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30*, 2785-2791.

12. Trott, O.; Olson, A. J. Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461.

13. Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions. *J. Mol. Biol.* **2000**, *295*, 337–356.

14. FRED; version 2.2.5, OpenEye Scientific Software, Inc., Santa Fe, NM, **2009**.

15. McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian Docking Functions. *Biopolymers* **2003**, *68*, 76-90.
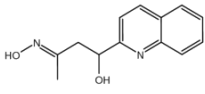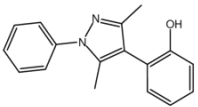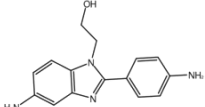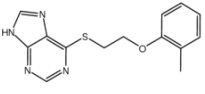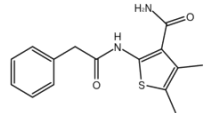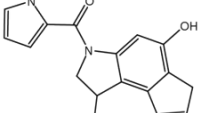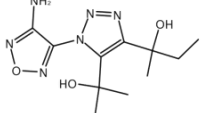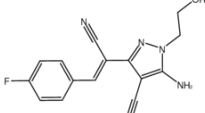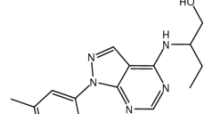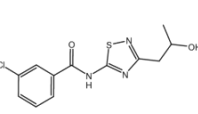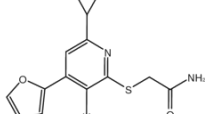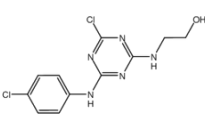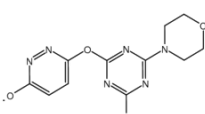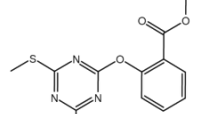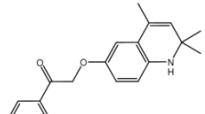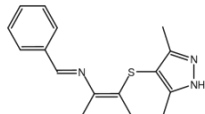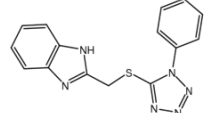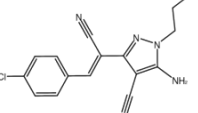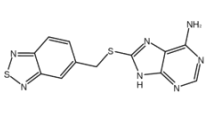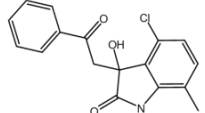
16. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical Scoring Functions: I. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425–445.

17. Verkivker, G. M.; Bouzida, D.; Gehlaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering Common Failures in Molecular Docking of Ligand-Protein Complexes. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 731-751.

18. Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44*, 1035-1042.

19. Glide, version 5.5, Schrödinger, LLC, New York, **2009**.

20. Mastro, version 9.0, Schrödinger, LLC, New York, **2009**.

21. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.

22. Yang, C.-Y.; Wang, R.; Wang, S. M-Score: a Knowledge-Based Potential Scoring Function Accounting for Protein Atom Mobility. *J. Med. Chem.* **2006**, *49*, 5903–5911.

23. Wang, R.; Lai, L.; Wang, S. Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. *J. Comput.-Aid. Mol. Des.* **2002**, *16*, 11–26.

24. Berman, H. M.; Westbrook, J; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

25. Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother of All Databases). *Proteins* **2005**, *60*, 333–340.

26. Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47*, 2977–2980.

27. Shoichet, B. K.; Bodian, D. L.; Kuntz, I. D. Molecular Docking Using Shape Descriptors. *J. Comput. Chem.* **1992**, *13*, 380–397.

28. Yin, S.; Biedermannova, L.; Vondrasek, J.; Dokholyan, N. V. MedusaScore: An Accurate Force Field-Based Scoring Function for Virtual Drug Screening. *J. Chem. Inf. Model.* **2008**, *48*, 1656–1662.

29. Goto, J.; Kataoka, R.; Muta, H.; Hirayama, N. ASEDock Docking Based on Alpha Spheres and Excluded Volumes. *J. Chem. Inf. Model.* **2008**, *48*, 583–590.

30. Naim, M.; Bhat, S.; Rankin, K. N.; Dennis, S.; Chowdhury, S. F.; Siddiqi, I.; Drabik, P.; Sulea, T.; Bayly, C. I.; Jakalian, A.; Purisima, E. O. Solvated Interaction
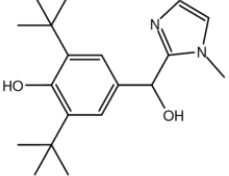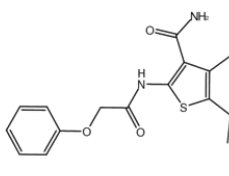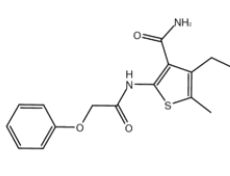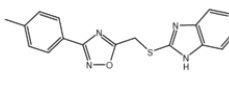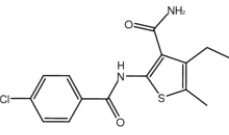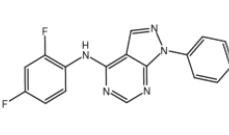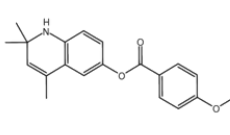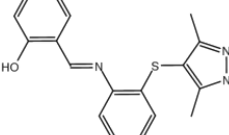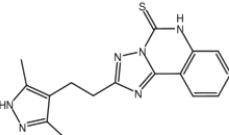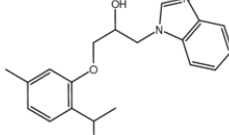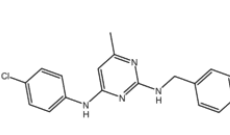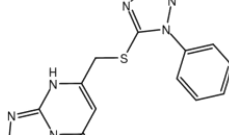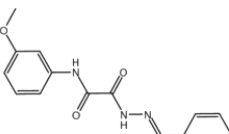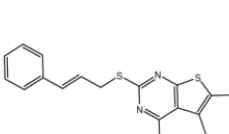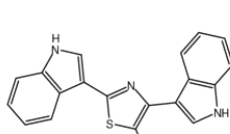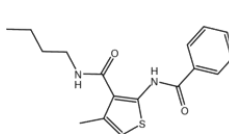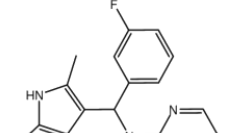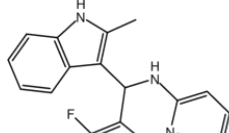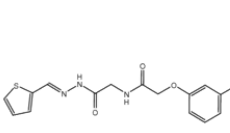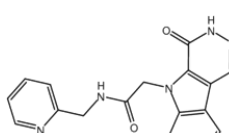
Energy (SIE) for Scoring Protein-Ligand Binding Affinities. 1. Exploring the Parameter Space. *J. Chem. Inf. Model.* **2007**, *47*, 122–133.

31. Cornell, W. D.; Cieplak, P; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc*. **1995**, *117*, 5179-5197.

32. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586-3616.

33. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225-11236.

34. Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S. B.; Johnson, A. P. eHiTS: A New Fast, Exhaustive Flexible Ligand Docking System. *J. Mol. Graph. Model.* **2007**, *26*, 198–212.

35. Stroganov, O. V.; Novikov, F. N.; Stroylov, V. S.; Kulkov, V.; Chilov, G. G. Lead Finder: An Approach to Improve Accuracy of Protein-Ligand Docking, Binding Energy Estimation, and Virtual Screening. *J. Chem. Inf. Model.* **2008**, *48*, 2371–2385.

36. Rahaman, O.; Estrada, T.; Doran, D.; Taufer, M.; Brooks, C.; Armen, R. Evaluation of Several Two-Step Scoring Functions Based on Linear Interaction Energy, Effective Ligand Size, and Empirical Pair Potentials for Prediction of Protein-Ligand Binding Geometry and Free Energy. *J. Chem. Inf. Model.* **2011**, *51*, 2047-2065.

37. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein-Ligand Docking Using GOLD. *Proteins* **2003**, *52*, 609–623.

38. Kramer, C.; Gedeck, P. Global Free Energy Scoring Functions Based on Distance-Dependent Atom-Type Pair Descriptors. *J. Chem. Inf. Model.* **2011**, *51*, 707–720.

39. Huang, S.-Y.; Zou, X. An Iterative Knowledge-Based Scoring Function to Predict Protein-Ligand interactions: I. Derivation of Interaction Potentials. *J. Comput. Chem.* **2006**, *27*, 1866–1875.
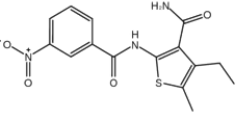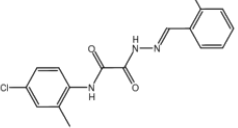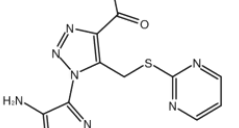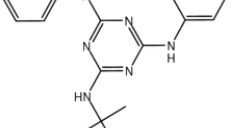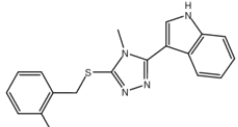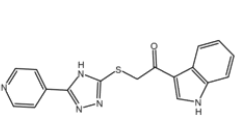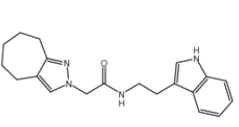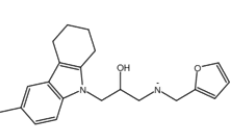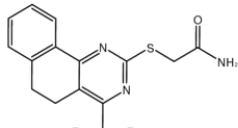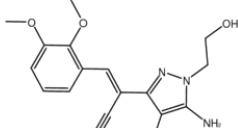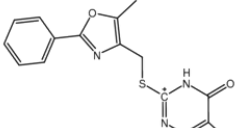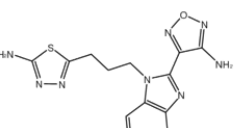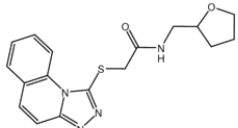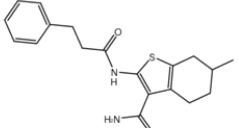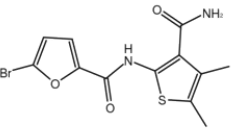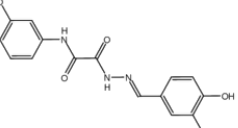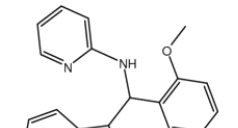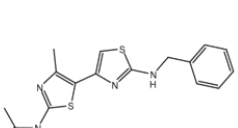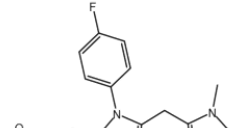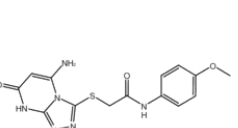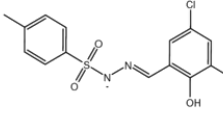
APPENDICES
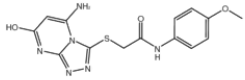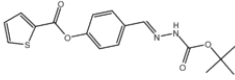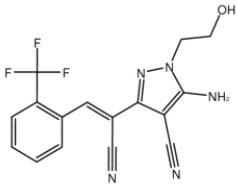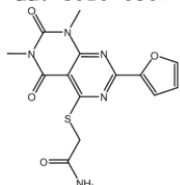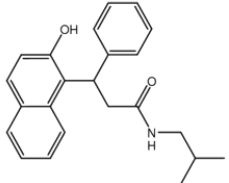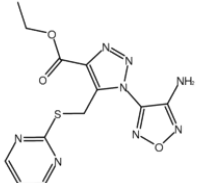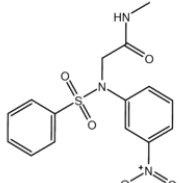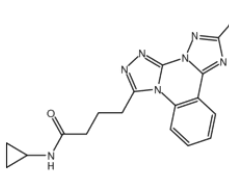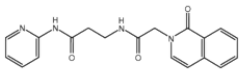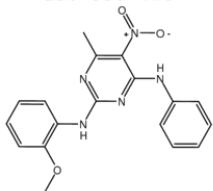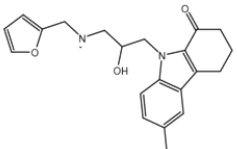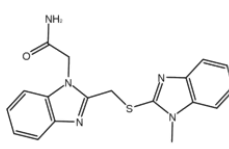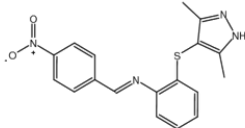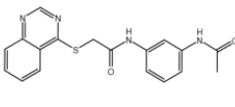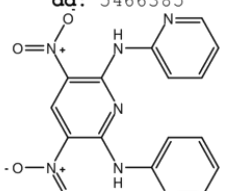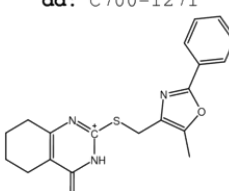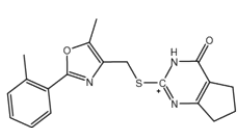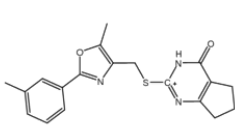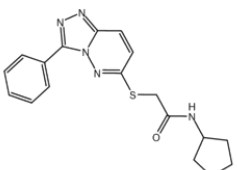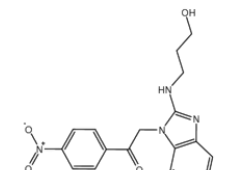
APPENDIX 1
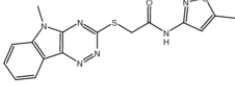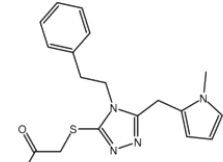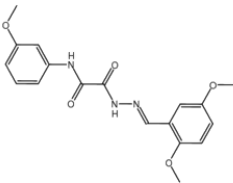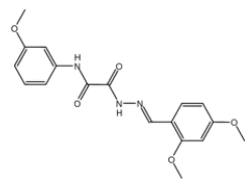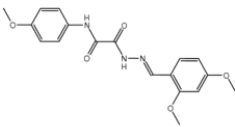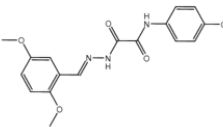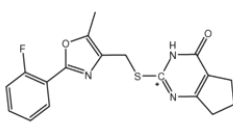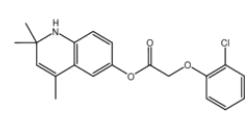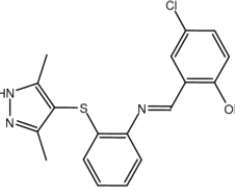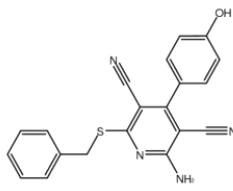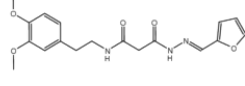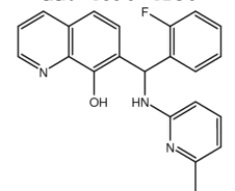
Appendix 1.1. List of the original 93 Eye pharmacophore compounds from Damm *et al.*

1 **dd:** 5562086	**2** **dd:** 1499-0598	**3** **dd:** 3769-2047	**4** **dd:** 1465-0022
Weight: 230.267	**Weight:** 264.328	**Weight:** 268.32	**Weight:** 286.359
5 **dd:** 5524366	**6** **dd:** 5493403	**7** **dd:** 1307-0007	**8** **dd:** 5561288
Weight: 288.371	**Weight:** 293.366	**Weight:** 296.331	**Weight:** 297.293
9 **dd:** K402-0900	**10** **dd:** C200-1280	**11** **dd:** 3343-2780	**12** **dd:** K890-0002
Weight: 297.362	**Weight:** 297.766	**Weight:** 299.354	**Weight:** 300.149
13 **dd:** 6989-0080	**14** **dd:** 3379-0757	**15** **dd:** 5483428	**16** **dd:** 5483692
Weight: 304.29	**Weight:** 306.346	**Weight:** 307.393	**Weight:** 307.421
17 **dd:** 5491484	**18** **dd:** 5524374	**19** **dd:** 1323-0109	**20** **dd:** 5056-0111
Weight: 308.369	**Weight:** 313.748	**Weight:** 315.385	**Weight:** 315.756

239

21 dd: 5529870 Weight: 316.445	**22** dd: 5539991 Weight: 318.397	**23** dd: 5555993 Weight: 318.397	**24** dd: 5371-4626 Weight: 322.392
25 dd: 5546084 Weight: 322.816	**26** dd: K402-0245 Weight: 323.306	**27** dd: 5303843 Weight: 323.392	**28** dd: 5493267 Weight: 323.42
29 dd: K280-0545 Weight: 324.412	**30** dd: 3910-0351 Weight: 324.424	**31** dd: 4466-2067 Weight: 324.815	**32** dd: 6456-0643 Weight: 326.344
33 dd: 5522450 Weight: 327.34	**34** dd: 4236-0436 Weight: 327.476	**35** dd: 6597798 Weight: 329.427	**36** dd: 5373736 Weight: 330.452
37 dd: 5228-4175 Weight: 331.394	**38** dd: 5228-4184 Weight: 331.394	**39** dd: 5468265 Weight: 331.396	**40** dd: C884-0070 Weight: 333.351

240
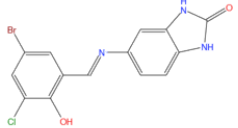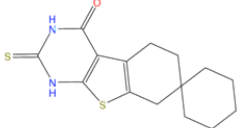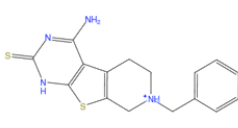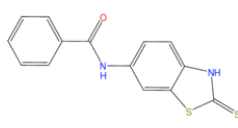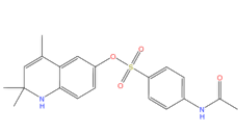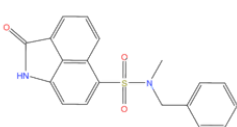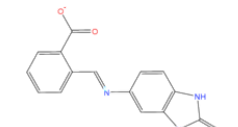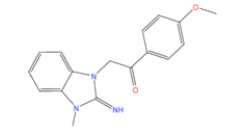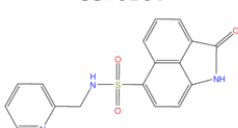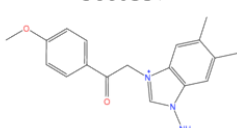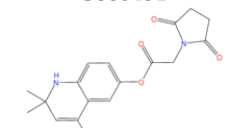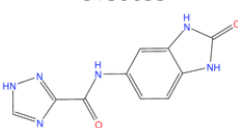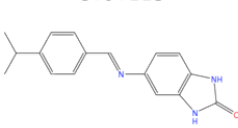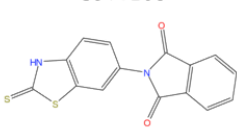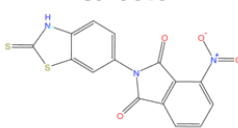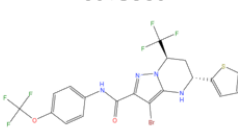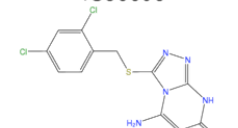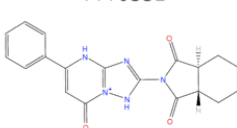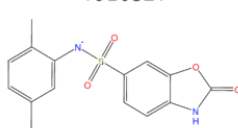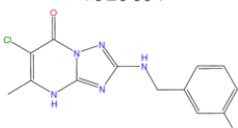
41 dd: 5552988 Weight: 333.368	**42** dd: 5526815 Weight: 333.75	**43** dd: 6703-1500 Weight: 334.32	**44** dd: 5469915 Weight: 334.427
45 dd: C592-0086 Weight: 334.447	**46** dd: 6585837 Weight: 335.391	**47** dd: C881-1035 Weight: 336.439	**48** dd: 6658-0012 Weight: 337.443
49 dd: 6466-0177 Weight: 339.341	**50** dd: 5626286 Weight: 339.355	**51** dd: C700-1363 Weight: 339.419	**52** dd: 3448-7570 Weight: 342.387
53 dd: 5867-3931 Weight: 342.423	**54** dd: 5574999 Weight: 342.463	**55** dd: 5542560 Weight: 343.201	**56** dd: 5545228 Weight: 343.339
57 dd: 5228-4182 Weight: 343.43	**58** dd: 6504928 Weight: 344.463	**59** dd: C677-0232 Weight: 345.402	**60** dd: 6737-0622 Weight: 346.371
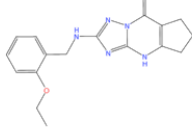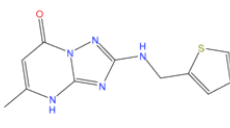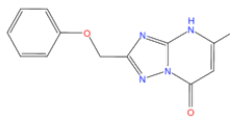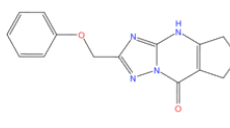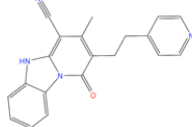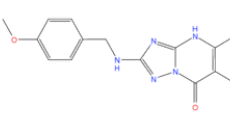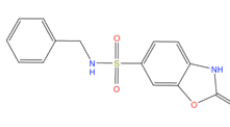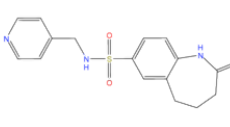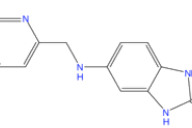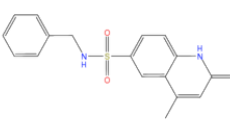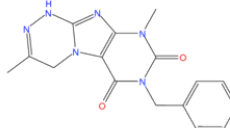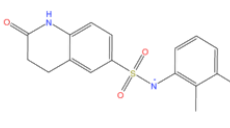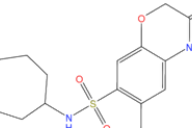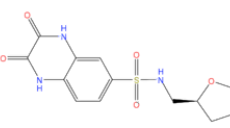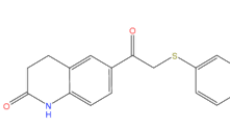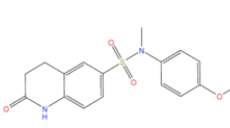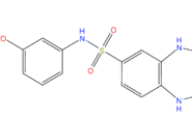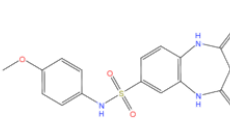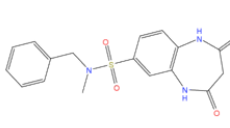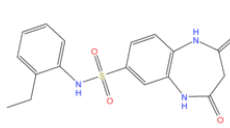
61 **dd:** 6143-0142 **Weight:** 346.371	**62** **dd:** 6261334 **Weight:** 346.407	**63** **dd:** 5522294 **Weight:** 347.3	**64** **dd:** C610-0367 **Weight:** 347.355
65 **dd:** 5489937 **Weight:** 347.458	**66** **dd:** 6703-1782 **Weight:** 348.347	**67** **dd:** 6197605 **Weight:** 349.367	**68** **dd:** C684-0061 **Weight:** 349.398
69 **dd:** C066-1486 **Weight:** 350.378	**70** **dd:** 5307415 **Weight:** 351.366	**71** **dd:** 5926-0037 **Weight:** 351.426	**72** **dd:** 3448-9010 **Weight:** 351.434
73 **dd:** 5486930 **Weight:** 352.418	**74** **dd:** C725-0034 **Weight:** 352.418	**75** **dd:** 5466385 **Weight:** 353.298	**76** **dd:** C700-1271 **Weight:** 353.446
77 **dd:** C700-1355 **Weight:** 353.446	**78** **dd:** C700-1362 **Weight:** 353.446	**79** **dd:** C742-0059 **Weight:** 353.45	**80** **dd:** 5557939 **Weight:** 354.366
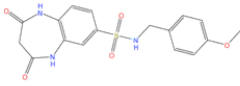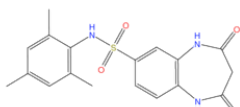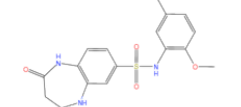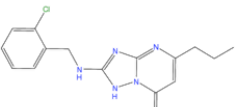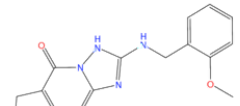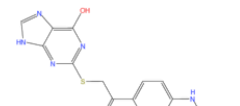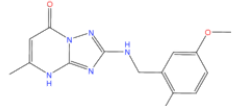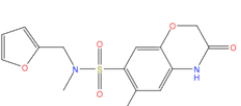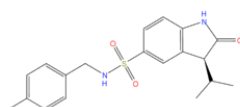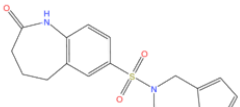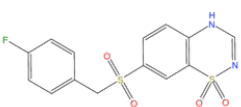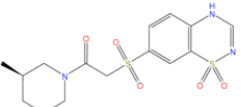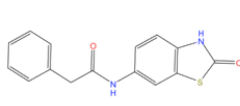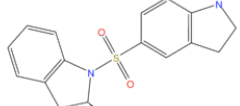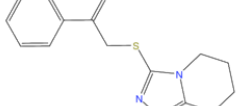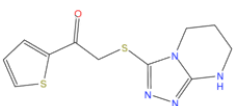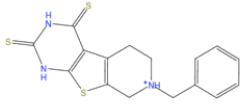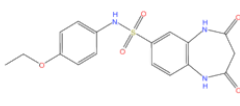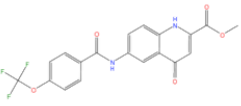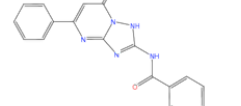
242

81	82	83	84
dd: 6218637	**dd:** C677-0010	**dd:** 5523166	**dd:** 5539286
Weight: 354.394	**Weight:** 355.466	**Weight:** 357.366	**Weight:** 357.366
85	86	87	88
dd: 5556943	**dd:** 5559290	**dd:** C700-1364	**dd:** 5469524
Weight: 357.366	**Weight:** 357.366	**Weight:** 357.409	**Weight:** 357.837
89	90	91	92
dd: 5485189	**dd:** 5600269	**dd:** 3448-5797	**dd:** 5655180
Weight: 357.865	**Weight:** 358.225	**Weight:** 358.425	**Weight:** 359.382
93			
dd: 4896-4230			
Weight: 359.404			

243
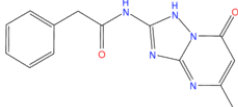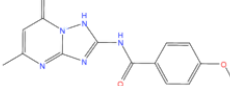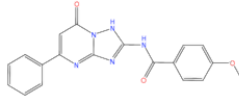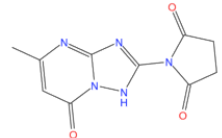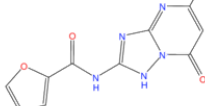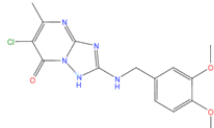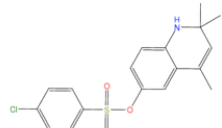
Appendix 1.2. 67 compounds found in compound **1** Markush search of CCG, ChemBridge, ChemDiv, and MayBridge.

1 5202063 **Acti (%):** 80 +/- 5 Supplier: ChemBridge Weight: 367	**2** 5228968 **Acti (%):** -- Supplier: ChemBridge Weight: 306	**3** 5343019 **Acti (%):** -- Supplier: ChemBridge Weight: 329	**4** 5373492 **Acti (%):** 65 +/- 2 Supplier: ChemBridge Weight: 286
5 5470681 **Acti (%):** 80 +/- 6 Supplier: ChemBridge Weight: 386	**6** 5482954 **Acti (%):** -- Supplier: ChemBridge Weight: 352	**7** 5522020 **Acti (%):** 88 +/- 8 Supplier: ChemBridge Weight: 280	**8** 5566364 **Acti (%):** 76 +/- 5 Supplier: ChemBridge Weight: 296
9 5570164 **Acti (%):** -- Supplier: ChemBridge Weight: 339	**10** 5660337 **Acti (%):** 79 +/- 2 Supplier: ChemBridge Weight: 310	**11** 5660491 **Acti (%):** -- Supplier: ChemBridge Weight: 328	**12** 5738653 **Acti (%):** 84 +/- 5 Supplier: ChemBridge Weight: 244
13 5797113 **Acti (%):** 77 +/- 7 Supplier: ChemBridge Weight: 279	**14** 5977108 **Acti (%):** 80 +/- 4 Supplier: ChemBridge Weight: 312	**15** 5979646 **Acti (%):** 37 +/- 6 Supplier: ChemBridge Weight: 357	**16** 6075860 **Acti (%):** -- Supplier: ChemBridge Weight: 555
17 7386606 **Acti (%):** 79 +/- 9 Supplier: ChemBridge Weight: 342	**18** 7770331 **Acti (%):** 82 +/- 6 Supplier: ChemBridge Weight: 364	**19** 7910527 **Acti (%):** 88 +/- 4 Supplier: ChemBridge Weight: 317	**20** 7929697 **Acti (%):** 71 +/- 7 Supplier: ChemBridge Weight: 369

21 7932774	22 7935081	23 7936269	24 7936917
Acti (%): 75 +/- 4	Acti (%): 64 +/- 4	Acti (%): 79 +/- 10	Acti (%): 58 +/- 5
Supplier: ChemBridge	Supplier: ChemBridge	Supplier: ChemBridge	Supplier: ChemBridge
Weight: 325	Weight: 261	Weight: 256	Weight: 282
25 7937150	26 7940247	27 7989211	28 9022153
Acti (%): 82 +/- 7	Acti (%): 89 +/- 5	Acti (%): 74 +/- 6	Acti (%): 78 +/- 10
Supplier: ChemBridge	Supplier: ChemBridge	Supplier: ChemBridge	Supplier: ChemBridge
Weight: 328	Weight: 299	Weight: 304	Weight: 331
29 9058373	30 9058855	31 9112729	32 9118639
Acti (%): --	Acti (%): 68 +/- 6	Acti (%): 65 +/- 7	Acti (%): 72 +/- 9
Supplier: ChemBridge	Supplier: ChemBridge	Supplier: ChemBridge	Supplier: ChemBridge
Weight: 240	Weight: 328	Weight: 324	Weight: 333
33 9123846	34 9154080	35 BTB 09244	36 C224-0681
Acti (%): 79 +/- 12	Acti (%): --	Acti (%): 95 +/- 8	Acti (%): 122 +/- 7
Supplier: ChemBridge	Supplier: ChemBridge	Supplier: MayBridge	Supplier: ChemDiv
Weight: 338	Weight: 325	Weight: 297	Weight: 346
37 C698-0346	38 C698-0347	39 C698-0348	40 C698-0378
Acti (%): 98 +/- 7	Acti (%): 107 +/- 13	Acti (%): 100 +/- 7	Acti (%): 96 +/- 5
Supplier: ChemDiv	Supplier: ChemDiv	Supplier: ChemDiv	Supplier: ChemDiv
Weight: 361	Weight: 361	Weight: 359	Weight: 359

41	42	43	44
C698-0391	**C698-0447**	**C698-0540**	**D043-0170**
Acti (%): 97 +/- 24	**Acti (%):** 79 +/- 6	**Acti (%):** 93 +/- 5	**Acti (%):** 87 +/- 6
Supplier: ChemDiv	Supplier: ChemDiv	Supplier: ChemDiv	Supplier: ChemDiv
Weight: 375	Weight: 373	Weight: 375	Weight: 318

45	46	47	48
D043-0196	**D701-0025**	**D298-0425**	**E245-0514**
Acti (%): 92 +/- 7	**Acti (%):** 80 +/- 7	**Acti (%):** 82 +/- 5	**Acti (%):** 92 +/- 5
Supplier: ChemDiv	Supplier: ChemDiv	Supplier: ChemDiv	Supplier: ChemDiv
Weight: 311	Weight: 343	Weight: 315	Weight: 336

49	50	51	52
E966-0066	**E977-0196**	**G428-0017**	**G428-0524**
Acti (%): --	**Acti (%):** 88 +/- 5	**Acti (%):** 81 +/- 11	**Acti (%):** 95 +/- 9
Supplier: ChemDiv	Supplier: ChemDiv	Supplier: ChemDiv	Supplier: ChemDiv
Weight: 358	Weight: 334	Weight: 354	Weight: 385

53	54	55	56
G645-0045	**G851-1179**	**HTS 12769**	**HTS 12771**
Acti (%): 91 +/- 14	**Acti (%):** 93 +/- 6	**Acti (%):** 92 +/- 8	**Acti (%):** 84 +/- 7
Supplier: ChemDiv	Supplier: ChemDiv	Supplier: MayBridge	Supplier: MayBridge
Weight: 284	Weight: 314	Weight: 274	Weight: 280

57	58	59	60
HTS 12793	**K906-2240**	**SEW 04939**	**STOCK3S-76897**
Acti (%): 89 +/- 12	**Acti (%):** 82 +/- 4	**Acti (%):** --	**Acti (%):** --
Supplier: MayBridge	Supplier: ChemDiv	Supplier: MayBridge	Supplier: IBScreen
Weight: 347	Weight: 375	Weight: 406	Weight: 345

246

61 STOCK3S-77617	**62** STOCK3S-80036

61 STOCK3S-77617

Acti (%): 104 +/- 7
Supplier: IBScreen
Weight: 283

62 STOCK3S-80036

Acti (%): 107 +/- 6
Supplier: IBScreen
Weight: 299

63 STOCK3S-80366

Acti (%): --
Supplier: IBScreen
Weight: 361

64 STOCK3S-84057

Acti (%): 95 +/- 4
Supplier: IBScreen
Weight: 247

65 STOCK3S-90451

Acti (%): 98 +/- 5
Supplier: IBScreen
Weight: 259

66 STOCK5S-32389

Acti (%): --
Supplier: IBScreen
Weight: 350

67 T0502-1854

Acti (%): 81 +/- 9
Supplier: Enamine
Weight: 364

247

APPENDIX 2

Appendix 2.1. HIV-1 protease activity analysis code.

HIVp_percentage-activity.pl

```perl
#!/usr/bin/perl -w
use strict;
use Statistics::Descriptive;

#########################################################################
##
## Peter Ung @ UMich COPharm / 15 September 2007
## v1.1 30Oct07
## v1.2 23Feb08  -- change time-point used from 0-5min to 2-5min
##               -- revised the way ARGV is read
##     24Feb08  -- minor change in print-out format
## v1.3  2Aug08  -- added special case: #Sat
##
## Calculate HIV %activity and %st devation from FRET probing
##
## Process file exported from SepectraMax M5.
## Positive control repeats must be placed in the first couple columns.
## All repeats must be placed next to each other.
##
## Initial time-point starts at 2min.
## Internal maximum for data period analyzed is 300s.
## Time format must be in XX:XX or XX:XX:XX
## Column name will be in number if the positive control is not name (+).
##
#########################################################################
## Constants

my $R_2   = 0.90;        # R^2
my $RFU   = 60;          # RFU limit
my $upACT = 110;         # upper activity limit
my $loACT = 0;           # lower activity limit
my $DEV   = 15.0;        # deviation limit
my $init_time = 120; # initial time-point (sec) used for slope calculation
my $max_time  = 420; # ending time-point (sec) used for slope calculation

#########################################################################

die "\n    Usage: x.pl <File> <no. (+) repeat> <no. test repeat> [option]
              -Controls must be the in first columns.
              -Repeats must be placed next to each other.
              -An Asterisk (*) is placed if: initial RFU >= $RFU
                                             Data R^2     <= $R_2
                                             Activity    >  $upACT% or < $loACT%
                                             % Deviat    >= $DEV%
              -Optional: -np -- Inhibit printing of the re-formated data
                         -i=<xx> -- initial time-point other than $init_time s
```

```perl
                                  -e=<xx> -- endpoint of analysis other than $max_time
s\n\n"
  unless @ARGV >= 3;

open FILE, "< $ARGV[0]" or die "\n    Error: File not found.\n\n";
print "## $ARGV[0]\n";

my ($posNum, $repeat, $prtOpt) = (0, 0, 1);
# Number of Positive-Control Run
if ($ARGV[1] =~ /\d+/) { $posNum = $ARGV[1]; }
else { die "\n    Error: <no. (+) repeat> must be integer.\n\n"; }

# Number of Repeat for Each Compound
if ($ARGV[2] =~ /\d+/) { $repeat = $ARGV[2]; }
else { die "\n    Error: <no. test repeat> must be integer.\n\n"; }

# Options
if (@ARGV > 3) {
  my @newARGV = splice(@ARGV, 3, $#ARGV);
  for (@newARGV) {
    $prtOpt = 0 if $_ =~ /-np/i;
    if (/-i=/) {            # user-defined initial time-point
      @_ = split /=/, $_;
      if ($_[1] =~ /\d+/) { $init_time = $_[1]; }
      else { die "\n    Error: <-i> option must be a number.\n\n"; }
    }
    if (/-e=/) {            # user-defined ending time-point
      @_ = split /=/, $_;
      if ($_[1] =~ /\d+/) { $max_time = $_[1]; }
      else { die "\n    Error: <-e> option must be a number.\n\n"; }
    }
  }
}

############################################################################
## Convert File Format
my (@name, @col, @rowCol, @x) = ((), (), (), ());
my ($name, $rec, $colNum, $rowNum, $temper) = (0, 0, 0, 0, 0);
while (<FILE>)
{
  # Record the title line preceeding data
  if (/^Time/)
  {
    $rec = 1;
    my @temp = split;
    shift @temp;
    $temper = 1 if $temp[0] =~ /temp/i; # exclude Temperature if occurs
    shift @temp if $temper == 1;

    if ($temp[0] =~ /\+/)      # check if (+) present
    {
      $name = 1;
      my @tempName = ();
      # Save the name of the positive control
      for (my $i = 0; $i < $posNum; $i++)
      { push @tempName, shift @temp; }
      my $z = $tempName[$posNum-1];
      if ($tempName[0] =~ /\+/i or /$z/i)
      { push @name, $tempName[0]; }
      else
      { die "\n    Error: Number of compound column differs from input
                  or error in compound's name.
                  Error in \"Check\" if (+) present\n\n"; }
```

```perl
        # Save the name of each set of compound tested
        while (@temp)
        {
          @tempName = ();
          for (my $i = 0; $i < $repeat; $i++)
          { push @tempName, shift @temp; }
          my $z = $tempName[$repeat-1];
          if ($tempName[0] =~ /$z/i)
          { push @name, $tempName[0]; }
          else
          { die "\n    Error: Number of compound column differs from input
                     or error in compound's name: $z.\n\n"; }
        }
      }
      $_ = <FILE>;
    }
last if /^~End/;
    if ($rec)      # Start to collect data after /^Time/
    {
      @col = split;
      next if @col == 0;
    last if @col < 3;                 # stop reading when readings are terminated
      push @x, shift @col;            # remove and save the x-value: Time
      shift @col if $temper == 1; # remove Temperature value
      $colNum = $#col + 1;           # number of column input -- number of well
      push @rowCol, [@col];
    }
}
close FILE;
$rowNum = $#rowCol + 1; # number of row input -- number of time point

##########################################################################
## Re-format the List for Calculation; Transpose the Data "Matrix"
my (@list, @initVal) = ((), ());
# transpose the "Matrix"
for (my $j = 0; $j < $colNum; $j++)      # orig data column number
{
  my @row = ();
  for (my $i = 0; $i < $rowNum; $i++)    # orig data row number
  {
    my $rowPt = $rowCol[$i];
    my @tempRow = @$rowPt;
    push @row, $tempRow[$j];
  }
  push @initVal, $row[0];
  push @list, [@row];
}

# reformat time step in @x from 00:00(:00) into actual minute
my ($data_init_bloc, $time_total) = (0, 0);
my @time = ();
for (my $i = 0; $i < $rowNum; $i++)
{
  my $sec = 0;
  my @tmp = split /:/, $x[$i];
  # time in XX : XX : XX
  if (@tmp == 3)
  {
    if ($tmp[2] =~ /0\d/i)     # check if the second of XX is > 9
    {                          # if < 9, cut out the leading 0
      @_ = split //, $tmp[2];
      $tmp[2] = $_[2];
    }
```

250

```perl
    if ($tmp[1] =~ /0\d/i)
    {
      @_ = split //, $tmp[1];
      $tmp[1] = $_[1];
    }
    $sec = $tmp[0]*3600 + $tmp[1]*60 + $tmp[2];
  }

  # time in XX : XX
  elsif (@tmp == 2)
  {
    if ($tmp[1] =~ /0\d/i)        # same as above
    {
      @_ = split //, $tmp[1];
      $tmp[1] = $_[1];
    }
    $sec = $tmp[0]*60 + $tmp[1];
  }
  else
  { die "\n    Error: Time format must be in 00:00 or 00:00:00.\n\n"; }

last if $sec > $max_time;        # stop storing Time larger than $max_time
  $data_init_bloc = $i if $sec >= $init_time && $data_init_bloc == 0;
  push(@time, $sec/60)           # convert time in sec to min and save
    if $data_init_bloc > 0;
   $time_total++;
}
#########################################################################

# Calculate the Slope for each well
my (@rowOut, @slope, @R2) = ((), (), ());
for (@list)
{
  my $saturate = 0;                      # if RFU is saturated, slope = 0
  my @well = @$_;
  splice(@well, $time_total, $#well);   # take off data beyond $max_time
  splice(@well, 0, $data_init_bloc);    # take off data before $init_time
  for (@well) {
    $saturate = 1 if /Sat/;
  }
  if ($saturate == 0) {
    my $stat = Statistics::Descriptive::Full->new();
    $stat->add_data(@well);    # $y
    my ($q, $m, $r, $rms) = $stat->least_squares_fit(@time);  # $y = $m*$x + $q
    my $R2 = $r**2;          # squared correlation coefficient; equiv. Excel's R^2
    push @slope, $m;
    push @R2, $R2;
  } else {
    push @slope, 0;
    push @R2, 0;
  }
  push @rowOut, [@well];
}

#########################################################################

## Optional: Print reformated Data
if ($prtOpt == 1)
{
  # Title
  printf "%6s | ", "No.";
  printf "  %7s"x($#time+1), @time;
  printf " | %6s  %6s\n", "Slope", "R^2";
```

```perl
  print  "---------"x($#time+4)."\n";

  # Print positive control
  for (my $j = 0; $j < $posNum; $j++)
  {
    my $saturate = 0;     #for special case #Sat
    my $rowPt = $rowOut[$j];
    my @row = @$rowPt;
    printf "%6s | ", $name[0] if $name == 1;
    printf "%6s | ", "(+)"    if $name == 0;
    for (@row) {
      $saturate = 1 if /Sat/;
    }
    printf "  %7.2f"x($#row+1), @row if $saturate == 0;
    printf "  %7s"x($#row+1), @row if $saturate == 1;
    printf " | %6.3f  %6.4f\n", $slope[$j], $R2[$j];
  }
  my ($k, $l, $line) = (0, 0, 0);
  for (my $i = $posNum; $i <= $#list; $i++)
  {
    my $saturate = 0;    # for special case #Sat
    my $rowPt = $rowOut[$i];
    my @row = @$rowPt;
    printf "%6s | ", $name[1+$l] if $name == 1;
    printf "%6s | ", 1+$l         if $name == 0;
    for (@row) {
      $saturate = 1 if /Sat/;
    }
    printf "  %7.2f"x($#row+1), @row if $saturate == 0;
    printf "  %7s"x($#row+1), @row if $saturate == 1;
    printf " | %6.3f  %6.4f\n", $slope[$i], $R2[$i];
    $k++;
    if ($k == $repeat) { $k = 0; $l++; $line++; }

    print  "---------"x($#time+4)."\n" if $line == 2;   # separate line
    $line = 0 if $line == 2;
  }
  print "---------"x($#time+4)."\n";
  print "\n\n";
}

#########################################################################
## Calculate Positive-Control Activity and Standard Deviation
my (@out, @print) = ((), ());

  my (@pos, @init, @R2_val) = ((), (), ());
  my ($posMean, $posStdv, $init, $R2) = (0, 0, 0, 0);
  # Take out positive controls in the first couple columns
  for (my $i = 0; $i < $posNum; $i++)
  {
    push @pos,     shift @slope;     # slope of line
    push @init,    shift @initVal;   # initial RFU of $y
    push @R2_val,  shift @R2;        # rms of slope
  }

  # Statistics of Postive Controls
  my $stat = Statistics::Descriptive::Full->new();
  $stat->add_data(@pos);
  $posMean = $stat->mean();
  $posStdv = $stat->standard_deviation();
#  my $dev = sqrt( 2*(($posStdv/$posMean)**2) )*100;
  my $dev = ($posStdv/$posMean)*100;
```

```perl
  # Calculate the Mean of the Initial Positive Controls
  $stat = Statistics::Descriptive::Full->new();
  $stat->add_data(@init);
  $init = $stat->mean();

  # Calculate the Mean of the Slope RMS Positive Controls
  $stat = Statistics::Descriptive::Full->new();
  $stat->add_data(@R2_val);
  $R2 = $stat->mean();
  # (+) control values
  @print = ($name[0], $posMean, $posStdv, $R2, $init, 100, $dev)
    if $name == 1;
  shift @name if $name == 1;
  @print = ("(+)", $posMean, $posStdv, $R2, $init, 100, $dev)
    if $name == 0;
  push @out, [@print];

########################################################################

## Calculate % Activity and % Standard Deviation
my $num = 0;
while (@slope)
{
  my (@data, @init, @R2_val) = ((), (), ());
  my ($mean, $stdv, $init, $R2, $activity, $actiStdv) = (0, 0, 0, 0, 0);

  for (my $j = 0; $j < $repeat; $j++)
  {
    push @data,    shift @slope;        # slope of each well
    push @init,    shift @initVal;      # initial data points
    push @R2_val,  shift @R2;           # rms of each slope
  }

  # Statistics for each data point with its repeat(s)
  my $stat = Statistics::Descriptive::Full->new();
  $stat->add_data(@data);
  $mean = $stat->mean();
  $stdv = $stat->standard_deviation();

  # Mean of the initial data point
  $stat = Statistics::Descriptive::Full->new();
  $stat->add_data(@init);
  $init = $stat->mean();
  # Mean of RMS of data point
  $stat = Statistics::Descriptive::Full->new();
  $stat->add_data(@R2_val);
  $R2 = $stat->mean();

  # Standard deviation for multiply/division: dz/z = sqrt((dx/x)2 + (dy/y)2)
  $mean = 1 if $mean == 0;
  $activity = ($mean/$posMean)*100;
  $actiStdv = sqrt( ($stdv/$mean)**2 + ($posStdv/$posMean)**2 )*$activity;

  @print = ($name[$num], $mean, $stdv, $R2, $init, $activity, $actiStdv)
    if $name == 1;
  @print = ($num+1, $mean, $stdv, $R2, $init, $activity, $actiStdv)
    if $name == 0;
  push @out, [@print];
  $num++;
}

########################################################################
## Print Result
```

253

```perl
print "initial DFU >= $RFU Data R^2 <= $R_2 Activity > $upACT% || < $loACT%".
      " Deviat >= $DEV%\n";
print "Time Analysed: ".$init_time." -- ".$max_time." s\n\n";
printf "%6s  %6s  %7s  %6s  %4s  %8s  %6s\n",
  "No.", "slope", "StDev", "R^2", "Init", 'Activity', "StDev";
print "-"x62, "\n";

foreach (@out)
{
  my @prt = @$_;
  printf "%6s  %6.2f  %7.4f  %6.4f  %4d  %8.2f  %6.2f  ",
    $prt[0], $prt[1], $prt[2], $prt[3], $prt[4], $prt[5], $prt[6];

  print "R" if $prt[3] <= $R_2;          # slope rms not exceeds $R_2
  print "I" if $prt[4] >= $RFU;          # initial RFU not exceeds xx
  print "A"                              # activity never exceeds xx%
    if $prt[5] > $upACT || $prt[5] < $loACT;
  print "D"                              # activity deviation exceeds xx%
    if $prt[6] > $DEV || $prt[6] < 0;
  print "\n";
}
print "\n";
```

APPENDIX 3

Appendix 3.1. Essential dynamics extraction codes.

ED_eigenVector_porcupine_split.pl

> ➢ Extract the individual eigenvector information from the *ptraj*-generated eigenvector output `<evec_all.pev>`.

```perl
#!/usr/bin/perl -w
use strict;
use Math::Vec;


##############################################################################
##
##      Peter M.U. Ung          @ UMich COPharm
##      v1.0    11.03.01
##      v2.0    11.03.02 -- output eigenvalue
##      v3.0    11.03.09 -- bugfix: the last eigenvector not read in: no ****
##      v4.0    12.02.29 -- bugfix: add ./ to PDB output for VMD to locate
##
##      Purpose: read in PTRAJ calculated matrix PCA eigenvectors and output
##               the average C-alpha PDB structure. For the vectors on the
##               backbone atoms of each residue (@N,CA,C,O), vector addition of
##               all four is done and is then scaled (default = 75). The vector
##               is then projected onto the @CA of the residue.
##
##      Need:
##      1) aerage_structure.pdb -- pdb structure file with only the C-alpha of
##                      each residue. The coordinates will be replaced
##      2) porcupine_template.vmd
##      3) evec.pev -- correlation PCA of trajectory
##      4) essential.ptraj -- see Interactive Essential Dynamics for reference
##
##############################################################################

die "\n   Usage: x.pl <ptraj eigenVector> <output prefix> <PDB template> <VMD
template>\n\n"
  unless @ARGV == 4;

my ($vectIn, $prefix, $pdbTmpl, $vmdTmpl) = @ARGV;

## Constants
my $scaling = 75;       ## Scaling factor to elongate Vector for visualization
in VMD
```

```perl
## Read in the PTRAJ eigenvector file, usually .pev file
open IN, "< $vectIn"
  or die "\n     Error: Cannot read in <ptraj eigenVector> file: $vectIn\n\n";

my $vector = 0;
my (@line, @coord, @vector, @eigenVec) = ((), (), (), ());
my $c = 0;
## Separate the Average Structure and the EigenVectors by **** in the file
while (<IN>) {
  ## Store the information before/between each ****
  if (/\*\*\*\*/) {
    $vector = 1;
    push @eigenVec, [@vector] unless @vector == 0;
    undef @vector;
    next;
  }

  ## Average structure of the trajectories: at the beginning of the file
   # backbone atom order: @N(x,y,z),@CA(x,y,z),@C(x,y,z),@O(x,y,z)
  if (!$vector) {
    next if /Eigenvector file|^ \d/;
    @line = split;
    foreach (@line) { push @coord, $_; }
    undef @line;
    next;
  }

  ## Split the eigenvectors and store each of them separately
   # ED mode number and eigenvalue
  if ($vector == 1) {
    $vector = 2;
    @line = split;
    push @vector, [(@line)];
    undef @line;
    next;
  }
   # ED mode vectors of @N(x,y,z),@CA(x,y,z),@C(x,y,z),@O(x,y,z)
  if ($vector == 2) {
    @line = split;
    foreach (@line) { push @vector, $_; }
    undef @line;
  }

   # the last line of the last eigenVector, not followed by ****
  if (eof) {
    push @eigenVec, [@vector] unless @vector == 0;
    undef @vector;
  }
}
close IN;

################################################################################
use Math::Vec qw(:terse);
use Math::Trig;
```

256

```perl
## Check number of atom in Coordinates and pick out the atomtype for PDB
generation
my (@mol, @atomList);
my $pickAtom = 1;          # @N,CA,C,O = (0,1,2,3) = superimpose unified atom to
@CA=1

 # Check number of atom in the coordinates: has to be multiple of 3 (x,y,z)
times
## 4 (@N,CA,C,O)
die "\n    Error: Number of atom in Coordinates is not multiple of 3*4 --
".($#coord+1)."\n\n"
  if ($#coord+1)%12 != 0;

 # Pick out the (x,y,z) of each atom and put them in the same set
for (my $x = 0; $x <= $#coord; $x += 3) {
  my $y = $x;
  push @mol, V($coord[$y], $coord[$y+1], $coord[$y+2]);
}
 # pick out the coordinates of specific atomtype from @N,CA,C,O --> (0,1,2,3)
 # here using @CA as vector template, so chose $pickAtom = 1
for (my $x = $pickAtom; $x <= $#mol; $x += 4) { push @atomList, $mol[$x]; }


open PDB, "< $pdbTmpl"
  or die "\n    Error: Cannot open PDB C-alpha template file: $pdbTmpl\n\n";
open OUT, "> $prefix.pdb"
  or die "\n    Error: Cannot write to PDB C-alpha file: $prefix.pdb\n\n";

## Print out the Average Structure of the Trajectories by using a template PDB
file
## with all @CA and residue info ready
my $count = 0;
while (<PDB>) {
  if (/^ATOM/) {
    my $newCoord = sprintf "%8.3f%8.3f%8.3f",
      $atomList[$count]->[0], $atomList[$count]->[1], $atomList[$count]->[2];
    substr($_, 30, 24, $newCoord);
    $count++;
  }
  print OUT;
}
close OUT;

#for (my $x = 0; $x <= $#atomList; $x++) {
#print "$atomList[$x]   $x\n";#}
#print "atomlist: $#atomList\n";

###############################################################################

## Combine Eigenvectors and project them to the atom coordinates
my @eigenVal;
foreach (@eigenVec) {
  my ($eigenVal, @atomVec, @vecSum, @vectList);
  @vector = @$_;
```

257

```perl
  # Check the number of vector in each Eigenvector: has to be 3 (x,y,z) times
  # 4 (@N,CA,C,O)
  die "\n    Error: Number of value is not multiple of 3*4 -- ".$#vector."\n\n"
    if ($#vector)%12 != 0;

  # Put out the eigenvector number and the assoicated eigenvalue
  $eigenVal = shift @vector;

  # Separate the numbers into (x, y, z) format
  for (my $x = 0; $x <= $#vector; $x += 3) {
    my $y = $x;
    push @atomVec, [($vector[$y], $vector[$y+1], $vector[$y+2])];
  }
  print "\n  Atom number = ".(($#atomVec+1)/4)."\n\n";

  # Combine the vectors on @N,CA,C,O by VectorSum and then scale the unified
vector
  for (my $x = 0; $x <= $#atomVec; $x += 4) {
    my $y = $x;
    my $cv = V($atomVec[$y][0],   $atomVec[$y][1],   $atomVec[$y][2]  ) +
             V($atomVec[$y+1][0], $atomVec[$y+1][1], $atomVec[$y+1][2]) +
             V($atomVec[$y+2][0], $atomVec[$y+2][1], $atomVec[$y+2][2]) +
             V($atomVec[$y+3][0], $atomVec[$y+3][1], $atomVec[$y+3][2]);
    push @vecSum, V($cv->ScalarMult($scaling));
  }

  # Add the unified vector to the @CA coordinates to create an endpoint of the
vector
  for (my $x = 0; $x <= $#vecSum; $x++) {
#     my $coord = V($atomList[$x]->[0], $atomList[$x]->[1], $atomList[$x]->[2]);
    my $vecED = $vecSum[$x] + $atomList[$x];
    push @vectList, $vecED;
  }

  ## *** checking the vector sum
#   my $x = 0; my $y = 0;
#   foreach (@vecSum) {
#     $x++;
#     print "  $x ".(abs($_)*37.5)."\n";
#     $y += abs($_);
#   }

  ## Write out coordinates of vector for VMD Porcupine Plot display
  open VMD, "< $vmdTmpl"
    or die "\n    Error: Cannot read the template VMD script for Porcupine
Plot: $vmdTmpl\n\n";
  open OUT, "> $prefix.$eigenVal->[0]"
    or die "\n    Error: Cannot write out VMD script for Porcupine Plot:
$prefix.$eigenVal->[0]\n\n";

   # Write lines to create PCA mode vectors
  while (<VMD>) {
    if (/^CALPHAPDB/) {
#       print OUT "mol new {$prefix.pdb} type pdb\n";
#       print OUT "mol modstyle 0 top tube\nmol modcolor 0 top beta\n";
```

258

```perl
        print OUT "mol new {./$prefix.pdb} type pdb\n";
        print OUT "mol modstyle 0 top lines\nmol modcolor 0 top name\n";
        next;
      }
      # Write out combined vector of mode1 and mode2 with dot-product (scale 10)
    if (/^GRAPHICS/) {
      print OUT "# dotVec\ngraphics top color 2\n";        # "red" color for
dotVec
      for (my $x = 0; $x <= $#vectList; $x++) {
#print "$atomList[$x]->[0], $vectList[$x]->[0], $x\n";
        printf OUT "graphics top cone {%8.3f %8.3f %8.3f} {%8.3f %8.3f %8.3f}
radius 0.30000 resolution 10\n", $atomList[$x]->[0], $atomList[$x]->[1],
$atomList[$x]->[2], $vectList[$x]->[0], $vectList[$x]->[1], $vectList[$x]->[2];
      }
      next;
    }
    print OUT;
  }
  close VMD;
  close OUT;


  push @eigenVal, $eigenVal;
}


############################
## Write out the normalized eigenvalues of the used eigenvectors

open VAL, "> $prefix.eval"
  or die "\n    Error: Cannot open EigenValue file: $prefix.eval\n\n";
my $eigenValSum = 0;
my $per = 0;
foreach (@eigenVal) { $eigenValSum += $_->[1]; }
foreach (@eigenVal) {
  my $val =  ($_->[1]/$eigenValSum);
    $per += ($val*100);
  printf VAL "%5d\t%7.5f\t%5.1f%%\n", $_->[0], $val, $per;
}
close VAL;
```

Appendix 3.2. Essential dynamics generation code.

ED_vector_gen.pl
> ➢ Extract the eigenvector information from the *.vmd* file generated by the *perl* script
> `<ED_eigenVector_procupine_split.pl>` to generate an output file that contains
> the starting and ending points of the vector on each of the C-alpha atom.

```perl
#!/usr/bin/perl -w
use strict;


#############################################################################
##
##      Peter M.U. Ung  @ UMich @ COPharm
##      v1.0    10.08.10
##      v2.0    10.09.01 -- change the output vector to end-point coordinates,
##                          vector will be calculated in the ED_vector_dot.pl
##      v3.0    10.09.17 -- use an aligned C-alpha PDB to modified the
coordinates
##                          and vector end-point
##
##      Purpose: read in the C-alpha only PDB of a protein structure and
##               extract ED eignvector of each C-alpha and generate an output
file
##               that contains the two points of the vector on the C-alpha.
This
##               should correspond to the C-alpha coordinates from the
##               Porcupine_$i.vmd file generated by DynaTraj:
##      http://s12-
ap550.bioch.ox.ac.uk:8078/dynamite_html/dynatraj_tutorial.html
##
##      Related Scripts:
##      1) ED_vector_dot.pl -- calculate dot products of 2 different sets of
##                             eignvectors to determine degree of overlap
##      2) porcupine_$i.vmd -- VMD session file that contains the eignvectors
##                             and initial coordinates of the ED
##
#############################################################################

die "\n    Usage: x.pl <C-alpha PDB> <porcupine vmd> <output>\n\n"
  unless @ARGV == 3;

open PDB, "< $ARGV[0]"
  or die "\n    Error: Cannot open <C-alpha PDB>: $ARGV[0]\n\n";

my @pdb;
my ($resname, $resid);

## Read in the PDB and only consider C-alpha coordinates
while (<PDB>) {
  next unless /^ATOM/ && substr($_, 13, 2) eq "CA";

  $resname = substr($_, 17, 3);
  $resid   = substr($_, 22, 4);
  my $pdbX = substr($_, 30, 8); my $pdbY = substr($_, 38, 9);
```

```perl
  my $pdbZ = substr($_, 46, 8);

  push @pdb, [($resid, $resname, $pdbX, $pdbY, $pdbZ)];
}
close PDB;


## Read in VMD with vector information
open VMD, "< $ARGV[1]"
  or die "\n    Error: Cannot open <porcupine vmd>: $ARGV[1]\n\n";

my @vector;
my $count = 0;  # counting of C-alpha in PDB

while (<VMD>) {
  next unless /^graphics top cone/;
  s/{//g;        s/}//g;
  @_ = split;
  ## pick out the end-point (coordinates) of the vector on each C-alpha
  my $porX = $_[6];
  my $porY = $_[7];
  my $porZ = $_[8];

  ## add vector data to the PDB data
  push @vector, [($pdb[$count]->[0], $pdb[$count]->[1],
                  $pdb[$count]->[2], $pdb[$count]->[3], $pdb[$count]->[4],
                  $porX, $porY, $porZ)];

  $count++;
}
undef @pdb;
close VMD;

open OUT, "> $ARGV[2]"
  or die "\n    Error: Cannot write to <output>: $ARGV[2]\n\n";

foreach my $vec (@vector) {
  my @vec = @$vec;

#($resID, $resName, $coord[X], $coord[Y], $coord[Z], $vec[X], $vec[Y], $vec[Z])
  printf OUT "%5d %3s %8.3f %8.3f %8.3f %10.6f %10.6f %10.6f\n",
    $vec[0],$vec[1],$vec[2],$vec[3],$vec[4],$vec[5],$vec[6],$vec[7];
}
close OUT;
```

261

Appendix 3.3. Essential dynamics eigenvector dot-product code.

ED_vector_dot.pl
> ➢ Perform dot-product calculation between two eigenvectors extracted from the
> eigenvector file generated by PTRAJ <`evec_all.pev`> and processed by the *perl*
> script <`ED_vector_gen.pl`>.

```perl
#!/usr/bin/perl -w
use strict;
use Math::Vec;
use Statistics::Basic qw(:all);


#############################################################################
##
##      Peter M.U. Ung  @ UMich @ COPharm
##      v1.0    10.08.11
##      v2.0    10.09.01 -- input porcupine data are 2 sets of coordinates
##                          instead of 1 anchor coordinates and 1 vector
##      v3.0    10.09.03 -- create a C-alpha PDB with dot-product as B-factor
##      v4.0    10.09.06 -- rescale PCA vector magnitude to 0-10 and
incorporate
##                          vector magnitude into final representation
##      v5.0    10.09.07 -- remove PCA vector scaling dependency on dot product
##                          -- use B-factor column as dot product scale -1 - 0 - 1
##      v6.0    10.09.08 -- use the C-alpha coordinate of PCA_2
##      v6.1    10.09.09 -- add Histogram function for dot product distribution
##      v6.2    10.09.10 -- add directory of C-alpha PDB file
##      v7.0    10.09.13 -- add median calculation of overall dot product
##      v7.1    12.02.24 -- REMARK and # recognition at readin
##      v8.0    12.02.29 -- change the PDB write-out method to substr
##      v9.0    12.03.01 -- add calculation of cumulative overlap
##      v10.0   12.05.15 -- add calculation of Pearson skewness coefficient


##      Purpose: Using the coordinates and eignvectors extracted from 2
##               different sets of Porcupine plots (porcupine_$i.vmd) to
##               perform dot product to deduce the level of vector overlapping
##               of 2 eignvector sets.
##               !!!! all frames should have the same reference frame/prmcrd!!!
##      Related Scripts:
##      1) ED_vector_gen.pl -- generate the output list needed by this script
##      2) porcupine_template.vmd -- template file for dot product porcupine
##                                      plot
##      3) average_structure.pdb -- C-alpha PDB that procupine.vmd will read in
##
#############################################################################

die "\n   Usage: x.pl <PCA 1> <PCA 2> <template vmd> <template PDB> <output
prefix>\n\n"
  unless @ARGV == 5;

# Read in PCA_1 data from processed output from ED_vector_gen.pl
open V_1, "< $ARGV[0]"
  or die "\n    Error: Cannot open <PCA 1>: $ARGV[0]\n\n";
```

```perl
my @PCA_1;        # all PCA_1 vector start/end-points of each avail. C-alpha
my @list;         # temp. array to hold split data
## Save the coordinates of <PCA 1>
while (<V_1>) {
  next if /^#|^REMARK/;
  @list = split;
  ## @PCA_1 = ( $resID, $resName, ($coord 1-1), ($coord 1-2) )
  push @PCA_1, [( $list[0], $list[1],
                 [($list[2],$list[3],$list[4])],
                 [($list[5],$list[6],$list[7])]  )];
}
close V_1;


# Read in PCA_2 data from processed output from ED_vector_gen.pl
open V_2, "< $ARGV[1]"
  or die "\n   Error: Cannot open <PCA 2>: $ARGV[1]\n\n";

my @PCA_2;        # all PCA_2 vector start/end-points of each avail. C-alpha
## Save the coordinates of <PCA 2>
while (<V_2>) {
  next if /^#|^REMARK/;
  @list = split;
  ## @PCA_2 = ( $resID, $resName, ($coord 2-1), ($coord 2-2) )
  push @PCA_2, [( $list[0], $list[1],
                 [($list[2],$list[3],$list[4])],
                 [($list[5],$list[6],$list[7])]  )];
}
close V_2;


################################################################################
my $countx = 0;
my @resid_avail;
## Put residue with available vector together (some resisdues are not
calculated
## and missing in the porcupine_$i.vmd, although they are there in the initial
input
foreach my $pca1 (@PCA_1) {
## $resid   = $pca1->[0];      ##  $resname = $pca1->[1];

  foreach my $pca2 (@PCA_2) {
    if ($pca2->[0] == $pca1->[0] && $pca2->[1] eq $pca1->[1]) { # $resID,
$resname
      $countx++;
  ## @available_coord = ( $resID,$resName,($coord 1-1),($coord 1-2),($coord 2-
1),
  ##                         ($coord 2-2) )
      push @resid_avail, [($pca1->[0],$pca1->[1],$pca1->[2],$pca1->[3], $pca2-
>[2], $pca2->[3])];
      next;
    }
  }
}
print "countx = $countx\n";

my (@Vec, @dotVec, @pt_1_2, @vec_1_2);
```

```perl
my ($dot_1_2, $vec_1_2, $dot_ang);
my ($cross_1_2, $cross_mag);
my ($maxAmp_1, $maxAmp_2, $ampSum_1, $ampSum_2) = (0, 0, 0, 0); # vector
amptitude
my ($dot_pos, $dot_neg) = (0, 0);
use Math::Vec qw(:terse);
use Math::Trig;

foreach my $pt (@resid_avail) {
  my @pt = @$pt;

  # start/end-points of PCA_1 and PCA_2 vector
  my $a = V($pt[2]->[0], $pt[2]->[1], $pt[2]->[2]);      # (coord 1-1)
  my $b = V($pt[3]->[0], $pt[3]->[1], $pt[3]->[2]);      # (coord 1-2)
  my $c = V($pt[4]->[0], $pt[4]->[1], $pt[4]->[2]);      # (coord 2-1)
  my $d = V($pt[5]->[0], $pt[5]->[1], $pt[5]->[2]);      # (coord 2-2)

  ## vector 1 ($b-$a) and vector 2 ($c-$a) -- both have the same origin
  my $vec_1 = $b - $a;
  my $vec_2 = $d - $c;
  my $Amp_1 = abs($vec_1);       # vector length
  my $Amp_2 = abs($vec_2);       # vector length

  ## collect max. magnitude
  $maxAmp_1 = $Amp_1 if $Amp_1 > $maxAmp_1;      # find the max. vector length
  $maxAmp_2 = $Amp_2 if $Amp_2 > $maxAmp_2;      # find the min. vector length
  $ampSum_1 += $Amp_1;  $ampSum_2 += $Amp_2;     # sum of all vector lengths

  ## @Vec = ($resID, $resName ($coord 1-1), $vec_1, $vec_2, $Amp_1, $Amp_2)
  push @Vec, [($pt[0], $pt[1], $pt[2], $vec_1, $vec_2, $Amp_1, $Amp_2)];
}
# print "xxx $pt[0] $pt[1]\n";
# print "xxx ".abs($b-$a)." xxx ".abs($c-$a)."\n\n";

my ($dot_sum, $dot_num, $CO2_1_2, $div) = (0, 0, 0);
my $avg_1_2;
my (@div, @dot_histo) = ((), (0,0,0,0,0,0,0,0,0,0,0,0));
my @dot_library;        # store all dot-products
## Vector calculations
foreach my $vec (@Vec) {
  my ($pt_0, $pt_1, $pt_2, $vec_1, $vec_2, $Amp_1, $Amp_2) = @$vec;
  my $a = V($pt_2->[0], $pt_2->[1], $pt_2->[2]);

  ## Scaling factor of $vec_1 + $vec_2, incorporating amptitude of both vectors
  my $scale = (($Amp_1/$maxAmp_1) * 10) * (($Amp_2/$maxAmp_2) * 10);

  # Dot product (a numeric value) of $vec_1 and $vec_2 to see their degree of
overlap
  $dot_1_2 = U($vec_1)->Dot(U($vec_2));          # unit vector dot product
  $avg_1_2 = V($vec_1->Plus($vec_2));            # $vec_1 + $vec_2
  $vec_1_2 = V($avg_1_2->ScalarMult(1/2));       # scaling of ($vec_1 + $vec_2)
  @pt_1_2  = $vec_1_2->Plus($a);                 # ($vec_1 + $vec_2) vector end-
point

  ## Build Histogram of dot-product angle distribution
```

264

```perl
  $dot_ang = acos($dot_1_2)*(180/pi);   # convert normalized dot-product to
angle
  $div = sprintf "%4.1f", ($dot_ang/10);# construct histogram
  @div = split /\./, $div;
  $dot_histo[$div[0]]++;                       # construct histogram -- take only the
integer

  ## Counting for global overlap calculation
  $dot_sum += $dot_1_2;                        # add up all dot-product
  $CO2_1_2 += ($dot_1_2)**2;                   # Cumulative overlap, CO = (Sum -
                                               #
($dot_1_2)^2)^0.5
  $dot_num++;                                  # count number of dot-product added
  $dot_pos++ if $dot_1_2 >= 0;         # number of dot-product >= 0
  $dot_neg++ if $dot_1_2 <  0;         # number of dot-product <  0

  ## Save dot product for median calclulation
  push @dot_library, $dot_1_2;

  ## @dotVec = (<PCA_1 origin pt>, <PCA_1 vec pt>, <PCA_2 vec pt>, <vec>,
  ##            $dot_prod, $dot_angle)
  push @dotVec, [($pt_0, $pt_1, $pt_2, [@pt_1_2], $dot_1_2, $dot_ang)];

  # Cross product (a vector) of $vec_1 and $vec_2
#  $cross_1_2 = U($vec_1->Cross($vec_2));        # Vector of cross product
#  $cross_mag = abs(U($vec_1)->Cross(U($vec_2)));# Magnitude of cross product;
                                                # perpendicularity


}
my $dot_mean   = mean([@dot_library]  );
my $dot_median = median([@dot_library]);
my $dot_stddev = stddev([@dot_library]);

## Write out C-alpha PDB with dot-product as B-factor
## dot-product = 1 (red) when total overlap in same direction, = -1 (blue) when
total
## overlap in opposite direction, = 0 (white) when perpendicular.
open PDB, "< $ARGV[3]"
  or die "\n    Error: Cannot open <template C-alpha PDB>: $ARGV[3]\n\n";
open OUT, "> $ARGV[4].pdb"
  or die "\n    Error: Cannot open <C-alpha PDB>: $ARGV[4].pdb\n\n";

my $g = 0;
while (<PDB>) {
  next unless /^ATOM/;

  @_ = split;
  if ($dotVec[$g]->[0] == $_[4]) {
    my $x = sprintf "%6.2f", (100 - (1+$dotVec[$g]->[4])*50);
    substr($_,60,6,$x);
    print OUT $_;
  }
  $g++;
}
close PDB;
```

```perl
close OUT;


my ($print_PCA_1, $print_PCA_2) = (0, 0);
## Write out new porcupine plot vmd input
open TMPL, "< $ARGV[2]"
  or die "\n    Error: Cannot open <template vmd>: $ARGV[2]\n\n";
open OUT , "> $ARGV[4].vmd"
  or die "\n    Error: Cannot open <output vmd>: $ARGV[4]\n\n";

my $atomPt;    my @atom;
while (<TMPL>) {
  ## Write line to read in modified C-alpha PDB
  if (/^CALPHAPDB/) {
    print OUT "mol new {./$ARGV[4].pdb} type pdb\n";
    print OUT "mol modstyle 0 top tube\nmol modcolor 0 top beta\n";
    print OUT "mol new {./$ARGV[4].pdb} type pdb\n";
    print OUT "mol modstyle 0 top lines\nmol modcolor 0 top name\n";
    next;
  }

  ## Write lines to create PCA mode vectors
  if (/^GRAPHICS/) {

    ## Write out combined vector of mode1 and mode2 with dot-product (scale 10)
    print OUT "# dotVec\ngraphics top color 2\n";        # "red" color for
dotVec
    for (my $i = 0; $i <= $#dotVec; $i+= 2) {
      $atomPt = $dotVec[$i];    @atom = @$atomPt;
      printf OUT "# %8.3f\t%8.2f degree\t$atom[1] $atom[0]\n", $atom[4],
$atom[5];
      printf OUT "graphics top cone {%8.3f %8.3f %8.3f} {%8.3f %8.3f %8.3f}
radius 0.30000 resolution 10\n",
        $atom[2]->[0], $atom[2]->[1], $atom[2]->[2], $atom[3]->[0], $atom[3]-
>[1], $atom[3]->[2];
      undef @atom;

      ## Print statistics
      if ($i == $#dotVec-1) {
print "xxx\n";
        printf OUT "## Average Ampitude for PCA_1: %6.2f\tfor PCA_2: %6.2f\n",
          ($ampSum_1/$dot_num), ($ampSum_2/$dot_num);
        printf OUT "## Overall (average) dot product: %6.3f (%6.1f deg) out of
%4s counts\n",
          abs($dot_mean),   (acos(abs($dot_mean))*(180/pi)),   $dot_num;
        printf OUT "## Overall (median) dot product:  %6.3f (%6.1f deg) out of
%4s counts\n",
          abs($dot_median), (acos(abs($dot_median))*(180/pi)), $dot_num;
        printf OUT "## Overall (mode) dot product:    %6.3f (%6.1f deg) out of
%4s counts\n",
          (&histo_mode([@dot_histo])*10), 1.1,   $dot_num;
        printf OUT "## Standard deviation dot product: %6.3f\n",
          abs($dot_stddev);
        printf OUT "## Normalize Cumulative Overlap: %6.3f out of %4s
counts\n",
```

266

```perl
            (($CO2_1_2/$dot_num)**0.5), $dot_num;
        printf OUT "## Positive dot product: %5.1f %%\n",
(($dot_pos/$dot_num)*100);
        printf OUT "## Negative dot product: %5.1f %%\n",
(($dot_neg/$dot_num)*100);

        print OUT "## Histogram of dot product: 10-degree increment\n";
        for (my $x = 0; $x <= $#dot_histo; $x++) {
          print OUT "##\n" if $x == 5 or $x == 13;
          printf OUT "## %3s - %3s %4.1f %% | ",
             ($x*10), (($x+1)*10), (100*$dot_histo[$x]/$dot_num);
          $dot_histo[$x] = sprintf "#" x $dot_histo[$x];
          print OUT "$dot_histo[$x]\n";
        }
        print OUT "##\n";
    }
  }
```

Appendix 4.1. Residue backbone dihedral angle analysis codes

Note: Three sets of codes, one perl script, one python script, and one GNUPlot input file, are used concurrently to produce the result.

rama_histo_gen.pl
 ➢ Generate heat-map

```perl
#!/usr/bin/perl -w
use strict;


#########################################################################
##
##   Peter MU Ung @ Umich COPharm
##
##  v1.0 -- 22 Feb 2009 / original script
##
##
##   Use the pre-generated torsion angle HISTOGRAM to construct
##   heat-map Ramachandran plot.
##   This is mainly for the analysis of torsion angle cluster of 1 particular
##   residue over a time period.
##
##   Require Histogram file with dihedral angle and the relative population.
##
##   Use with:
##     -) combine_column.pl
##     -) ecoDnaK.fasta
##     1) rama_dih_gen_csh
##     2) phi_psi_dih_combine.csh
##     3) rama_histo_gen.py
##     4) rama_histo_sample.gnu
##     5) rama_histo_gen.pl
##
##   Output format will be PNG; JPEG is larger w/ lower resolution.
##
##   Note: Require a sequence file (fasta) of the protein
##          Input filename must be: $prefix-$run.dih.$resid.out
##
#########################################################################

#########################################################################
```

```perl
## Setup Parameters

my $folder = 'adph';                    ## Folder name with data files
my ($runStart, $runEnd)     = (1, 5);   ## Run of each simulation type
my ($residStart, $residEnd) = (5, 378); ## Residues used for analysis

my $denMax    = "1.5";                  ## Highest density (red) displayed
my $fasta     = "ecoDnaK.fasta";        ## Protein seq in FASTA

my $outPrefix = "ADPH";                 ## Prefix: $outPrefix.dih.$resid.png
my $outFolder = "histo";                ## Output folder

############################################################################
## Constants

# Temp. files
my $gnuTemplt = "rama_histo_sample.gnu";## Template GNUPlot input file
my $gnuFinal  = "tmp4.gnu";             ## Temp. GNUPlot input file
my $dataFile  = "tmp4.data";            ## Temp. data of all runs
my $histoFile = "tmp4.histo";           ## Temp. Histogram file

## Python calling
my $python = "/users/mlerner/software/bin/python rama_histo_gen.py ";
#my $python = "/usr/bin/python rama_histo_gen.py ";

# hash of 1-letter/3-letter AA code
my %code = (
'A' => 'Ala', 'C' => 'Cys', 'D' => 'Asp', 'E' => 'Glu',
'F' => 'Phe', 'G' => 'Gly', 'H' => 'His', 'I' => 'Ile',
'K' => 'Lys', 'L' => 'Leu', 'M' => 'Met', 'N' => 'Asn',
'P' => 'Pro', 'Q' => 'Gln', 'R' => 'Arg', 'S' => 'Ser',
'T' => 'Thr', 'V' => 'Val', 'W' => 'Trp', 'Y' => 'Tyr'
);

############################################################################
## Read in FASTA sequence

open FASTA, "< $fasta"
  or die "\n    Error: Cannot open input FASTA file: $fasta.\n\n";

my @SEQ = ();          # FASTA Sequence array
while (<FASTA>) {
  next if /#/;
  @SEQ = split //, $_;
}

############################################################################
##

# Iterate through each residue
for (my $resid = $residStart; $resid <= $residEnd; $resid++) {

  my $cat = "";
  # Iterate through all runs
  for (my $run = $runStart; $run <= $runEnd; $run++) {
```

```perl
      $cat .= "$folder/$folder-$run.dih.$resid.out ";
  }

  system "cat $cat > $dataFile";
  my $py = $python." $dataFile $histoFile";
  system($py);


  ## output figure filename
  my $output = "$outFolder/$outPrefix.histo.$resid.png";


  ## Title name:
  my $title = $outPrefix.": ".$code{$SEQ[$resid-1]}." ".$resid;

  open TEMPLATE, "< $gnuTemplt"
    or die "\n    Error: Cannot open file: $gnuTemplt.\n\n";

  open OUTPUT, "> $gnuFinal.$resid"
    or die "\n    Error: Cannot open file: $gnuFinal.$resid.\n\n";

  while (<TEMPLATE>) {
    s/DENMAX/$denMax/;
    s/HISTOPLOT/$histoFile/;
    s/OUTPUTPNG/$output/;
    s/TITLE/$title/;
    print OUTPUT;
  }
  close TEMPLATE;
  close OUTPUT;

  ## System call -- GNUPlot
  my $sys = "gnuplot "."$gnuFinal.$resid > x.x";
  system $sys;
  print "  Complete $resid\n";


}
```

Rama_histo_gen.py

> ➤ Supporting python script to calculate the frequency of occurance at the grid point for heat-map generation. Use with the perl script <rama_histo_gen.pl>.

```python
#!/users/mlerner/software/bin/python

#############################################################################
## 2 Feb 2009 Peter MU Ung @ UMich @ COPharm
##
## from
http://maple.rsvs.ulaval.ca/mediawiki/index.php/Making_density_maps_using_Gnuplot
## use to convert dihedral angle of a residue over time (snapshots) to
## HISTOGRAM format for GNUPlot
##
## Required input file format (for each frame): <time> <phi> <psi>
##
```

```
##   Use with:
##     -) combine_column.pl
##     -) ecoDnaK.fasta
##     1) rama_dih_gen_csh
##     2) phi_psi_dih_combine.csh
##     3) rama_histo_gen.py
##     4) rama_histo_sample.gnu
##     5) rama_histo_gen.pl
##
#############################################################################

from numpy import *
import sys, os

data_file = sys.argv[1]
histogram_file = sys.argv[2]

# Here, angles are defined from [-180 to 180[ degrees.
# You could choose another domain. Points will be automatically
# wrapped inside that domain. This can be useful for symetrical
# side chains, for instance. Be careful to use an appropriate domain,
# otherwise the wrapping will produce meaningless data.
x_min, x_max, y_min, y_max = -180.0, 180.0, -180.0, 180.0

# Number of 2D regions in which the plot is divided.
x_resolution, y_resolution = 180, 180

def read_angles(line):
    tokens = line.split()
    x = float(tokens[1])
    y = float(tokens[2])
    while x < x_min:
        x = x_max - (x_min - x)
    while x >= x_max:
        x = x_min + (x - x_max)
    while y < y_min:
        y = y_max - (y_min - y)
    while y >= y_max:
        y = y_min + (y - y_max)
    return [x, y]

points = [read_angles(line) for line in open(data_file)]
count = len(points)
histogram = zeros([x_resolution, y_resolution])
x_interval_length = (x_max - x_min) / x_resolution
y_interval_length = (y_max - y_min) / y_resolution
interval_surface = x_interval_length * y_interval_length
increment = 1000.0 / count / interval_surface

for i in points:
    x = int((i[0] - x_min) / x_interval_length)
    y = int((i[1] - y_min) / y_interval_length)
    histogram[x,y] += increment

x_intervals = arange(x_min, x_max, (x_max - x_min) / x_resolution)
```

271

```
y_intervals = arange(y_min, y_max, (y_max - y_min) / y_resolution)

o = open(histogram_file, 'w')
for i, x in enumerate(x_intervals):
    for j, y in enumerate(y_intervals):
        o.write('%f %f %f \n' % (x, y, histogram[i,j]))
    o.write('\n')
print histogram.max()
```

rama_histo_sample.gnu
> Gnu input script

```
#############################################################
##
## 2 Feb 2009 Peter MU Ung @ UMich @ COPharm
##
## Template GNUPlot input file for generating a series of Ramachandran Plot
## with heat-map.
##
## Hightest density (DENMAX) is set to 1.5 for all Ramachandran Plot generated.
##
##  Use with:
##    -) combine_column.pl
##    -) ecoDnaK.fasta
##    1) rama_dih_gen_csh
##    2) phi_psi_dih_combine.csh
##    3) rama_histo_gen.py
##    4) rama_histo_sample.gnu
##    5) rama_histo_gen.pl
##
#############################################################

set term png
set size square
set xlabel 'Phi'
set xtic -180., 60., 180.
set xrange [-180:180]
set ylabel 'Psi'
set ytic -180., 60., 180.
set yrange[-180:180]
set format xy "%4.0f"
set border
set size 0.7, 1.0

#########################################

set xzeroaxis lt -1
set yzeroaxis lt -1

#########################################

set pm3d map interpolate 8,8
set cbrange [0:DENMAX]
```

```
set palette defined (0. "white", 0.01 "blue", 0.02 "green", 0.03 "yellow", 0.04
"red")
splot "HISTOPLOT" notitle

set output "OUTPUTPNG"
set title "TITLE"

replot
```

APPENDIX 5

Appendix 5.1. Code for analyzing conserved residues in a protein sequence.

Seq_conserv.pl

```perl
#!/usr/bin/perl -w
use strict;

die "\n    Usage: x.pl <sequence file>\n\n"
  if @ARGV == 0;

my $input = $ARGV[0];

open IN, "< $input"
  or die "\n    Error: Cannot open input file $input.\n\n";

##################################
#  Originally for use to calculate residue conservation and % mutation
#  of each input residue on the genome aligned
#
#  Original source of protein sequence: Biology Workbench - BlastP
#
#  Required format
#
#  ACDEGHK
#  .......
#  ......A
#  ...D...
#  ...D...
#  .......
#  -..-...
#
# where (.) is conserved, (A) is mutant and (-) is absent
#
##################################

my (@seq, @seqRst, @tmpSeq) = ((),());
my ($line, $noM, $consvM, $noncvM, $absent) = (-1,0,0,0);

while (<IN>) {
  chomp;
  if ($line == -1) {
    my @tmp;
    @_ = split //, $_;
    for (@_) {
      @tmp = ($_, 0, 0, 0);
      push @seqRst, [@tmp];
```

```perl
        undef @tmp;
     }
     $line++;
   }
else {
  $line++;
  $consvM = $noncvM = $absent = 0;        # reset counter
  @tmpSeq = split //, $_;
  for (my $i = 0; $i <= $#tmpSeq; $i++) {
     $seqRst[$i][1]++ if $tmpSeq[$i] =~
/A|C|D|E|F|G|H|I|K|L|M|N|P|Q|R|S|T|V|W|Y/;
     $seqRst[$i][2]++ if $tmpSeq[$i] =~
/a|c|d|e|f|g|h|i|k|l|m|n|p|q|r|s|t|v|w|y/;
     $seqRst[$i][3]++ if $tmpSeq[$i] =~ /-/;     # absent
  }
}  ## else
last if eof;
}

#############################
foreach my $residPt (@seqRst) {
  my @resid = @$residPt;
  ## number of conserved
  my $conserv = ($line - $resid[1] - $resid[2] - $resid[3])*100/$line;
  $consvM  = ($resid[1])*100/$line;
  $noncvM  = ($resid[2])*100/$line;
  $absent  = ($resid[3])*100/$line;

  printf "--%1s:\n", $resid[0];
  printf "  Conserved Residue      = %6.2f%%\n  Conserved Mutation     =
%6.2f%%\n  Non-conserved Mutation = %6.2f%%\n  Absent                 =
%6.2f%%\n\n", $conserv, $consvM, $noncvM, $absent;

}
```