ABSTRACT

Title of Dissertation: COMBINATORIAL LIBRARY DESIGN OF MUTATION-RESISTANT HIV PROTEASE INHIBITORS.

Sripriya Chellappan, Doctor of Philosophy, 2006

Dissertation directed by: Professor Michael K. Gilson
Department of Molecular and Cellular Biology

The emergence of HIV strains that are resistant to current HIV protease inhibitors in the past few years has become a major concern in AIDS treatment. The goal of this project is to design a combinatorial library of potential lead compounds that can bind to both the wild-type and mutant proteases and that can resist further mutations. A recent crystallographic study of complexes of HIV protease with its substrates has provided structural insights into the differential recognition of the substrates and inhibitors. It has been proposed that clinical resistance is a consequence of inhibitors failure to stay within the consensus substrate volume. In this work, we devised a quantitative indicator of the degree to which a candidate ligand falls outside the consensus substrate volume, and determined its correlation with the inhibitor's sensitivity to clinically relevant resistant mutations. The validation of this hypothesis has encouraged us to use this strategy in our design of a combinatorial library of inhibitors.

The compounds in a typical combinatorial library are built around a common structural scaffold possessing multiple connection points where substituents can be added by reliable synthetic steps. As the number of compounds encompassed by such a combinatorial scheme frequently exceeds what can actually be synthesized and tested, virtual screening methods are sought to shortlist the compounds. Even though these methods require only seconds to minutes of CPU time per compound, exhaustive screening of an entire virtual combinatorial library is computationally demanding. We therefore implemented a simple algorithm of combining substituents that have been optimized independently for the substituent sites. This method was compared with Genetic Algorithm, a global optimization method and was found equally efficient. This simple method was hence chosen for the design process.

A combinatorial library based on these ideas and methods has been synthesized and tested. It includes four compounds with nanomolar inhibition constants. Two of them were shown to have retained affinity against a panel of treatment-resistant mutations.

COMBINATORIAL LIBRARY DESIGN OF MUTATION-RESISTANT HIV PROTEASE INHIBITORS

by

Sripriya Chellappan

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2006

Advisory Committee:

Dr. Michael K. Gilson, Chairman/Advisor
Dr. Amitabh Varshney, Dean’s Representative
Dr. John Moult
Dr. Stephen M. Mount
Dr. Sergei Sukharev

DEDICATION

This thesis is dedicated to my husband, my parents and my brother
for all their love and support.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude towards my advisor Prof. Michael K. Gilson, for his patience, encouragement and guidance throughout my PhD. I find no words that are good enough to express my thanks for his advice and kind suggestions which have significantly improved my communication skills.

I would like to thank my committee members for their critical comments and suggestions, which continuously improved the quality of this work.

I thank our collaborators for their cooperation in this project to design new HIV protease inhibitors. This has helped me a lot in learning the language of medicinal chemistry. My heartfelt thanks to Mike for providing me this opportunity. This experience is definitely an asset for my research career.

I would like to thank Visvaldas Kairys, Chia-En Chang, and Miguel Fernandes, for helping me to get acquainted with the lab during my initial years as a graduate student. I also thank Visvaldas and Miguel for proof reading my papers.

I thank Robert Jorrisen for his professional and personal help. Without his support the time I spent in this lab would have been less enjoyable. I especially thank him for being my chauffeur at various times.

I am also thankful to Wei Chen, Eugene Melamud and Mike Potter, for their valuable suggestions in setting up the calculations for my project. I sincerely thank my other lab members for their suggestions and help throughout my research project.

I owe a special thanks to Vivek Gopalan, who worked with me for several hours, helping me format my thesis. Without his help, I would have never been able to complete my thesis on time.

My sincere thanks to my parents for giving me this wonderful life and helping in pursuing my dreams. I also thank my brother for his constant love and encouragement.

Last but not least, I am greatly indebted to my loving husband, who stood by me in the times of thick and thin. Without his unrelenting emotional support this body of work would never have been completed.

List of Figures

List of Tables

List of Abbreviations

AIDS	-	Acquired Immunodeficiency Syndrome
CPU	-	Central Processing Unit
DNA	-	Deoxy Ribonucleic Acid
ELISA	-	Enzyme Linked Immunosorbent Assay
FDA	-	Food and Drug Administration
GA	-	Genetic Algorithm
GP	-	Glycoprotein
HIV	-	Human Immunodeficiency Virus
HIVP	-	Human Immunodeficiency Virus Protease
ITC	-	Isothermal Calorimetry
LAM	-	Local Area Multi Computer
MC	-	Monte Carlo
MPI	-	Message Passing Interface
mRNA	-	Messenger Ribonucleic Acid
NC	-	Nucleocapsid
NNRTI	-	Non Nucleoside Reverse Transcriptase Inhibitors
NRTI	-	Nucleoside analogs of Reverse Transcriptase Inhibitors
OI	-	Opportunistic Infections
PDB	-	Protein Data Bank
PI	-	Protease Inhibitor
RMSD	-	Root Mean Square Distance
RNA	-	Ribonucleic Acid
SDF		Structure Definition File

Chapter 1. Introduction

1.1 AIDS and HIV

1.1.1 AIDS epidemiology

Acquired Immunodeficiency Syndrome (AIDS) is a severe, immune deficiency syndrome due to impaired T-cell function, which results in serious opportunistic infections such as tuberculosis, kaposi sarcoma and herpes (1). The UNAIDS reported in 2005 that there are about 38 million people worldwide affected by AIDS. Just last year, about 3 million people lost their lives to this disease (2). In 2005, at the G8 nations meeting and the United Nations World Summit, world leaders recognized AIDS as a major epidemic and committed to provide universal access to AIDS treatment for all who need it (2).

1.1.2 AIDS infection

In 1984, a cytocidal retrovirus was identified as an infectious etiologic agent for AIDS (3). This virus was named the Human Immunodeficiency Virus (HIV) by the International Committee on Taxonomy of Viruses in 1986. AIDS is a communicable disease which is transmitted through direct contact of a mucous membrane with a bodily fluid of an HIV infected person, such as blood, semen, or vaginal fluid. The main routes of HIV transmission are transfusion of blood and blood products, sexual contact, transmission from mother to child, and the use of contaminated hypodermic needles. Upon infection, HIV attacks host helper T-cells and macrophages, which are major components of the host immune system. The virus gains entry into the host cell through the cell surface receptors CD4 (cluster of differentiation) (4) and co-receptors such as the cytokine receptor CCR5 (5). Once inside the cell, HIV disrupts the normal cellular function by hijacking the cellular machinery to produce viral proteins. It also affects the integrity of the host cell by copious budding during its replication. This eventually leads to the destruction of helper T-cells and macrophages. There is an initial decline in the count of helper T-cells, which is followed by a short term recovery to a nearly normal level. After this recovery, there is an average annual loss of about 60 T-cells/µl. When the count falls below 200/µl (the normal count is between 800-1200/µl), the patient is diagnosed with AIDS (6). The time period between the recovery and this state corresponds to clinical latency.

The symptoms of acute HIV infection include headache, sore throat, muscle pain and other virus-like symptoms. Non-pruritic macular erythematous rashes of the trunk and extremities can distinguish HIV from other infectious diseases. During this acute infection phase, there is a high level of infectious viruses with heterogeneity in strains in the blood, which can be detected by Enzyme Linked Immunosorbent Assay (ELISA) detection kits. After the acute phase, patients can be asymptomatic for years. Because of the latency involved in the clinical manifestation of this syndrome, the virus was classified as lentivirus, a sub-family of retrovirus (7).

When the helper T-cell count drops below 200/µl, the patient becomes highly susceptible to opportunistic infections (OIs). Because of the highly compromised immune system, the individual becomes susceptible to a wide range of opportunistic pathogens, which are harmless to the normal population. Candidiasis, Kaposi sarcoma, herpes and cytomegalovirus are relatively common OIs among AIDS patients. Neoplasms and neurological symptoms such as aseptic meningitis, myelopathies are also not rare in AIDS patients (7). The OIs are chiefly responsible for the morbidity and mortality in AIDS.

1.1.3 HIV structure

The HIV virion is roughly spherical, with a diameter of approximately 100nm. The surface is made up of a lipid bilayer and envelope glycoproteins, which occur as trimers or tetramers. These glycoproteins include the external surface envelope protein gp120, and a trans-membrane protein, gp41, which interact covalently with each other. Gp120 has binding sites for the host cellular receptors. Beneath the lipid bilayer, there is a viral membrane, whose inner structure is supported by a myristoylated matrix protein. This protein is important for the viral structure and hence for the integrity of the virion. Apart from structural stability, the matrix protein has also been shown to play an important role in the incorporation of envelope proteins gp120 and gp41 into the mature virion. Inside the virus, there is a cone shaped core structure termed the capsid or nucleoid, which is composed of a capsid protein. The nucleoid contains two identical RNA strands and the associated proteins viral RNA dependent DNA polymerase and nucleocapsid protein (7) . Figure 1.1 shows a simple cartoon diagram of an HIV virion.

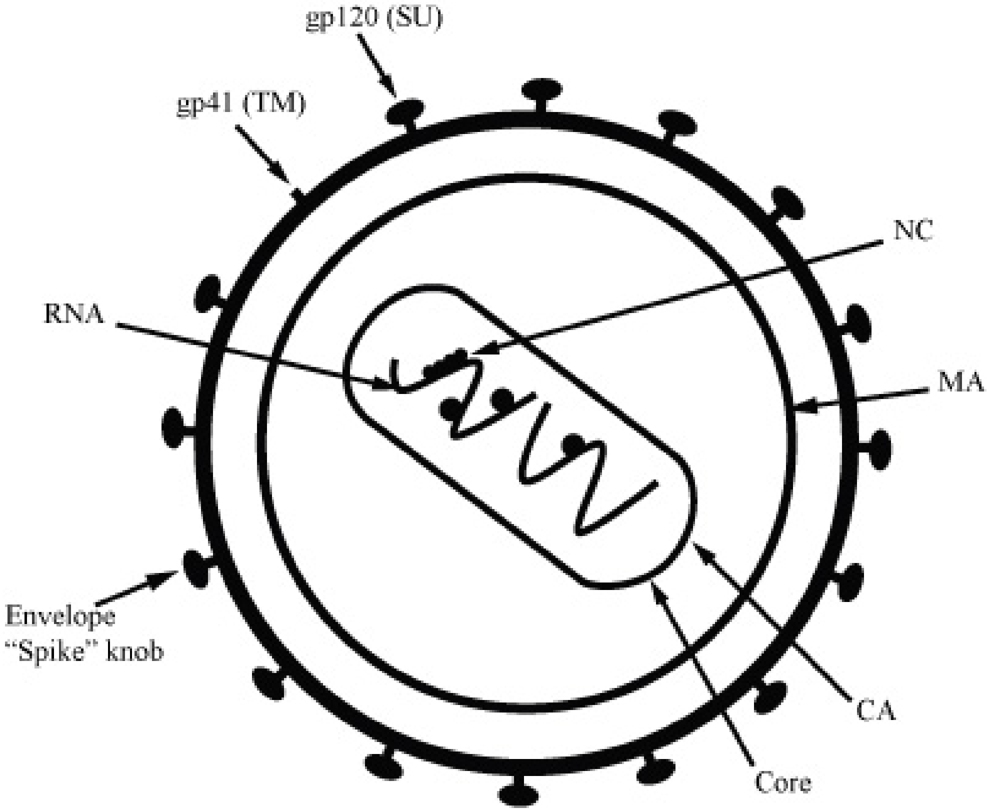


Figure 1.1: Structure of HIV.
Important structural features are highlighted. Reprinted from (7).

1.1.4 The life cycle of HIV

The life cycle of HIV starts with the attachment of viral particles to the target host cells, which are mainly helper T-cells and macrophages. Gp120 aids in the attachment of HIV to the host cell by interacting with its cell surface protein CD4. This interaction leads to a conformational change in the envelope protein, by which the virus gains entry into the host cell. During the process of entry, gp120 is displaced and cleaved by cellular proteases, exposing gp41 fusion domain, which has been suggested to be important in viral envelope shedding and cleavage. Alternative sites for viral entry through complement and F_c receptors have also been proposed (7). After the shedding and cleavage of the viral envelope, ribonucleocapsid is released into the host cell. The viral RNA undergoes reverse transcription, using its associated RNA-dependent DNA polymerase and RNase H proteins, forming double stranded DNA. This viral DNA then migrates to the nucleus and integrates with the host chromosomes. Some of the earliest mRNA species have regulatory genes, particularly tat, rev and nef. These genes determine the state of HIV virus as dormant or active. The primary full length viral mRNA is translated into Gag, Pol and Env polyproteins. This early transcription relies primarily on the cellular transcription factors. Hence the state of the host cell (differentiation or quiescent phase) affects the viral replication cycle. The Gag precursor protein on cleavage yields smaller structural proteins p25, p17, p9 and p6, whereas the Pol yields functional proteins reverse transcriptase, protease and integrase and Env yields two envelope proteins gp120 and gp41. Other viral proteins are produced by alternate splicing events. Viral genomic RNA is then incorporated into a capsid forming at the host cell surface. The processing of Gag, Pol and Env polyproteins by HIV protease occurs at the cell surface or in budding virion. Viral envelope proteins get inserted on the viral cell surface. HIV then buds off the host cell, completing the HIV replication cycle. A cartoon representation of HIV life cycle is shown in Figure 1.2. There are several life stages in HIV replication cycle that can be considered as targets for anti-AIDS therapy. They are viral entry, reverse transcription of viral RNA, integration of viral DNA into host DNA, processing of Gag, Pol, Env polyproteins, viral assembly and budding. Gp41 (needed for viral fusion), reverse transcriptase and HIV protease are the targets for the existing FDA approved drugs (7).

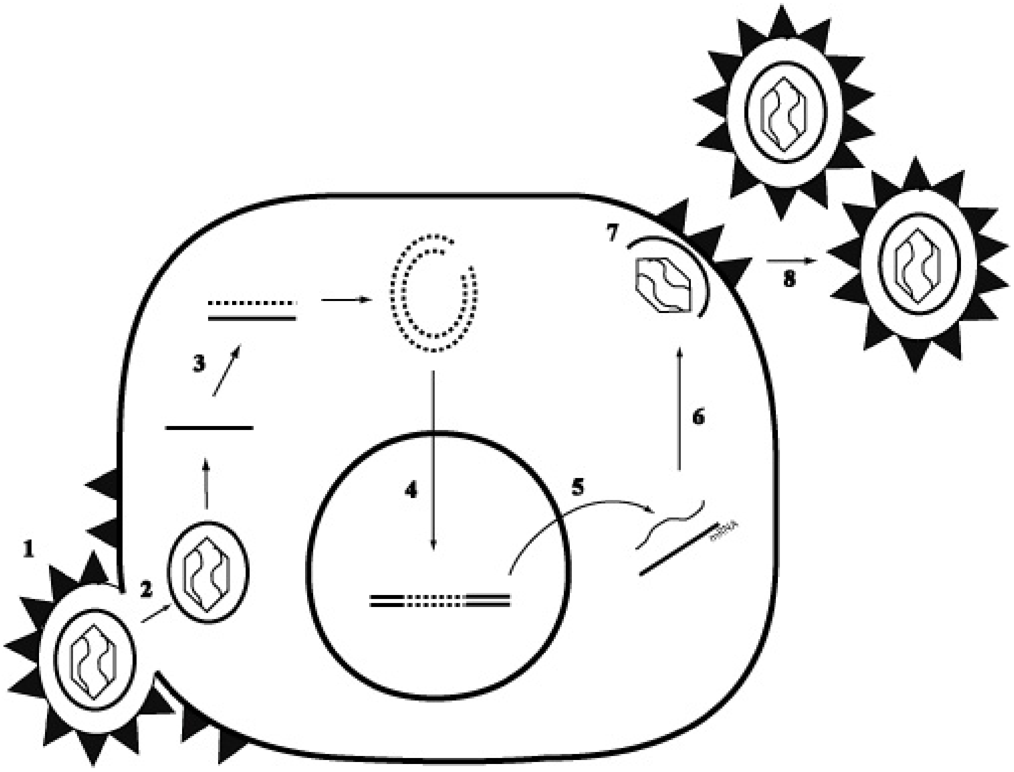


Figure 1.2: Diagram of the life cycle of HIV.

(1) Viral entry and fusion. (2) Envelope shedding and cleavage. (3) Reverse transcription. (4) Integration. (5) Cellular transcription and translation. (6) Post processing of precursor polyproteins. (7) Viral assembly. (8) Viral budding. Reprinted from (7)

1.1.5 Anti-HIV therapy

The four major therapeutic classes of anti-HIV drugs include are nucleoside analogues of reverse transcriptase inhibitors (NRTI), non-nucleoside reverse transcriptase inhibitors (NNRTI), protease inhibitors (PI) and fusion inhibitors. The first two classes act by inhibiting the RNA-dependent DNA polymerase (reverse transcriptase): the NRTIs competitively inhibits the binding of nucleosides to reverse transcriptase and thereby prevents elongation of viral DNA strand, while the NNRTIs non-competitively inhibit the viral reverse transcriptase. FDA approved NRTIs include Zidovudine (8), Didanosine (9) and Lamivudine (10), Zalcitabine (11), Stavudine (12) and Abacavir (13). Combinations of two or more NRTIs such as Combivir (Lamivudine & Zidovudine) and Trizivir (Abacavir & Lamivudine & Zidovudine) are also therapeutically used. Nevirapine (14), Delavirdine (15) and Efavirenz (16) are the NNRTIs that are FDA approved. Protease inhibitors act by competitively inhibiting HIV protease, an enzyme that is essential for virulence. Amprenavir (17), Indinavir (18), Saquinavir (19), Nelfinavir (20), Ritonavir (21), Lopinavir (22) and Atazanavir (23) are FDA approved PIs. A combination of Lopinavir and Ritonavir is also in clinical use. Fusion inhibitors act by binding to viral protein gp41, which is important for the viral entry and fusion. Enfuvirtide (24) is the only FDA approved drug that belongs to this class. It is advocated only as an add-on drug.

A combination of one PI or one NNRTI and two NRTIsis routinely used in AIDS therapy to overcome the emergence of drug resistant mutations. Such combination therapy, termed Highly Active Anti Retroviral Therapy (HAART), has been reported to be responsible for a rapid decline in the morbidity and mortality rate associated with HIV infection (25).

1.1.6 HIV protease

1.1.6.1 Structure

HIV protease is an enzyme that processes the precursor polyproteins into mature proteins. This posttranslational processing is one of the vital processes of the HIV replication cycle. Mutation of catalytic Asp25 in HIV protease to asparagine destroys the enzyme's catalytic activity, supporting its classification as an aspartyl protease, a class that also includes renin and pepsin. The virion produced by such mutants lacks virulence and infectivity (26). A synthetic peptide resembling an HIV protease substrate but with a non-hydrolysable replacement of the scissile amide bond was found to inhibit HIV protease activity. It was also found to inhibit the viral replication in blood lymphocytes. These studies suggested that HIV protease could be a suitable target for therapeutic intervention (27-29).

HIV protease shares sequence homology around the active site with other retroviral proteases. It is a homodimer, with 99 amino acids in each monomer. It exhibits C2 symmetry in the absence of ligands. A cartoon representation of HIV protease is given in Figure 1.3. The amino and carboxyl termini from both monomers form a four-stranded anti-parallel beta sheet at the dimer interface. This structure is stabilized by ionic interactions between the N- termini (residues 1-4 of beta strand *a*) of each monomer and the C-termini (residues 96 –99 of beta strand *q*) of other monomer (30). The beta strand *b* (residues 9-15) continues to another beta strand *c*, through a loop, which terminates in the active site triad. The catalytic triad Asp (25, 25'), Thr (26,26'), Gly (27,27') is located at the bottom of the dimer interface. Each monomer contributes one aspartic acid to the catalytic triad. Coplanar arrangement of catalytic aspartates is due to the network of hydrogen bonds resembling "fireman's grip", between the loops bearing catalytic triad. Thr26 and Thr24 from both the monomers are involved in this network of hydrogen bonds. Beta chain *d* (residues 30-35) which follows the strand *c*, is terminated at a distorted loop (residues 36-42). There is an approximate two fold intra-molecular symmetry in the monomer, making second half of the molecule topologically similar to the first half of the molecule. The *a'* beta strand (resides 43 - 49) and a part (residues 52-58) of longer *b'* beta strand (residues 52-66) form a flap. These glycine rich flap regions, from both the monomers, cover the active site. The beta chain *c'* (residues 69-78) is connected to another beta chain *d'* (residues 83-85), through a loop (residues 79-82). A well-defined helix (residues 86-94) follows the beta strand *d'*, which is in turn followed by the C-terminal beta strand *q*. Aψ-shaped beta sheet in the molecular core, is formed by four of the beta strands (*c*, *d* and *d'* ; *c*',*d* and *d'*). This feature is a characteristic for the family of aspartic proteases(31).

There are at least three distinct subsites on either side of the catalytic triad. The numbering of sub sites begins from the catalytic site and continues on either side (Figure 1.4). The residues that comprise S_1 and S_1' site are mostly hydrophobic. Most of the crystal structures of inhibitors and substrates show hydrophobic moieties interacting with this subsite, though exceptions are not rare (31). Subsites S_2 and S_2' are also hydrophobic but still accommodate both hydrophobic and hydrophilic residues (31). The distal subsites are not as well-defined, and can accommodate a wide variety of substituents.

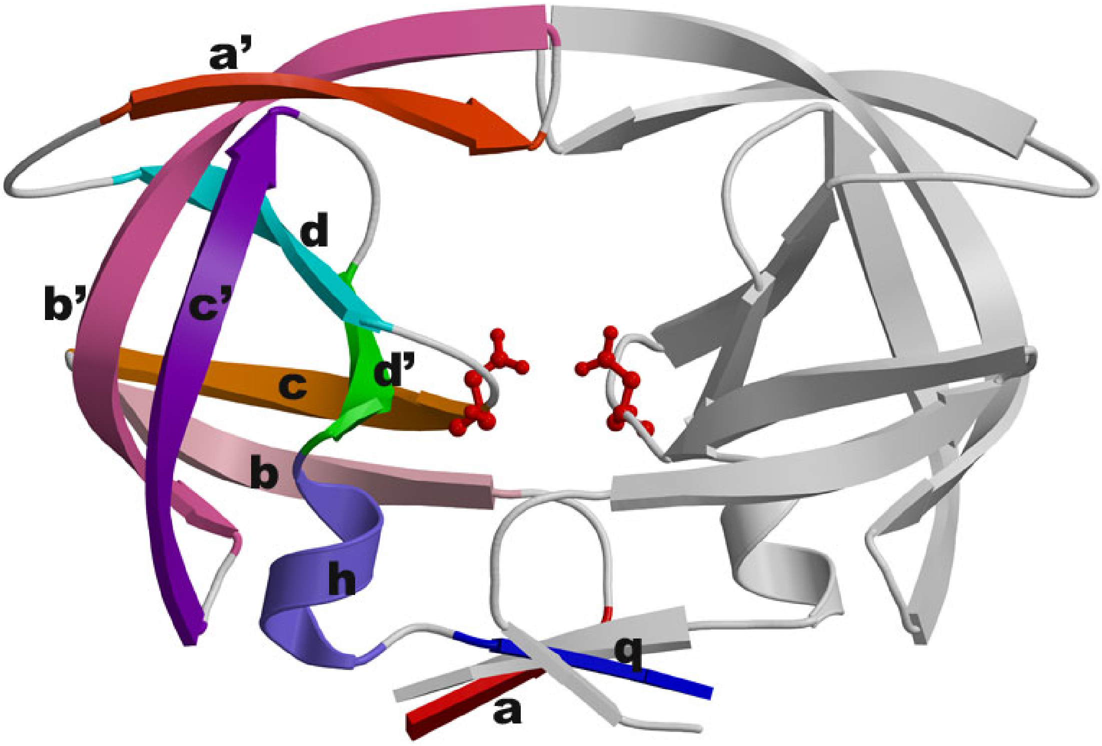


Figure 1.3: A cartoon representation of HIV protease.
Substructures are labeled as described in the text. Catalytic aspartates are represented in ball and stick model. Redrawn from (31).

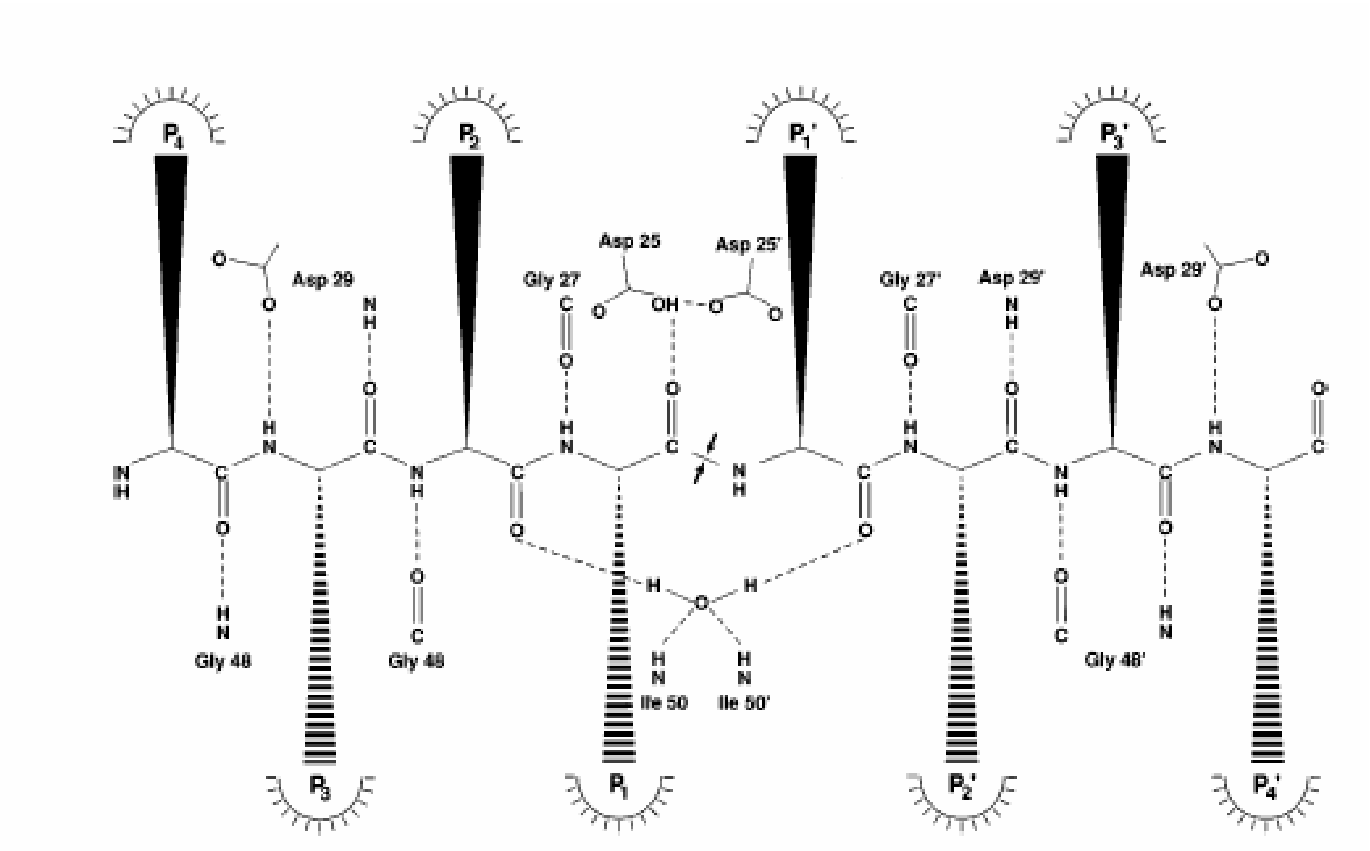


Figure 1.4: Hydrogen bonds between HIV-1 Protease and a modeled substrate. Reprinted from (32).

1.1.6.2 Substrate recognition

HIV protease processes Gag, Pol and Env polyprotein into structural and functional proteins. These substrates require a large-scale flap opening in HIV protease to access the active site. This movement in the flap region can be observed in structures of ligand bound and free HIV proteases with more heterogeneous flap structures in the Protein Data Bank (PDB) (33), varying from closed to open form. The nature of this movement in flap regions is debatable. Some studies have modeled it as rigid lever movement (34), while others have described it as a curling of flap tips (35).

The binding, cleavage, and release of substrates involve large-scale movement of certain regions in HIV protease (36, 37). The protease molecule can be conceptually divided into four regions based on their mobility (Figure 1.5). They are the fulcrum (residues 11-21 and 11’-21’), fireman’s grip (residues 22-28 and 22’-28’), flaps (residues 34-59 and 34’-59’) and cantilever (residues 64-74 and 64’-74’) (Figure 5). The fireman’s grip region is considered rigid because of the extensive hydrogen-bond network in this region. Movements of the flap, fulcrum and cantilever regions was observed to correlate with each other and with the movement of substrates in a recent study using Molecular Dynamics (MD) simulations (36).

Substrates bound to the protease form a parallel β-sheet with one monomer and anti-parallel β-sheet with the other. The transition between these two parts of the substrate generates a kink in the substrate at the center of the active site, directly above the catalytic aspartates, which is thought to facilitate cleavage (38). The hydrogen bonds between the substrate and the backbone amide nitrogen of residues I50 through the flap water are thought to be important for the catalytic action (39). These hydrogen bonds were suggested to exert strain on the scissile amide bond by causing it to rotate out of the plane and lose double-bond character, assisting in catalysis. Most of the hydrogen bonds between the substrates and the enzyme involve the backbone of the substrates.(38) Therefore the substrate specificity arises mainly from nonpolar interactions between the side-chains of the substrate and the residues lining the corresponding sub-sites (37).

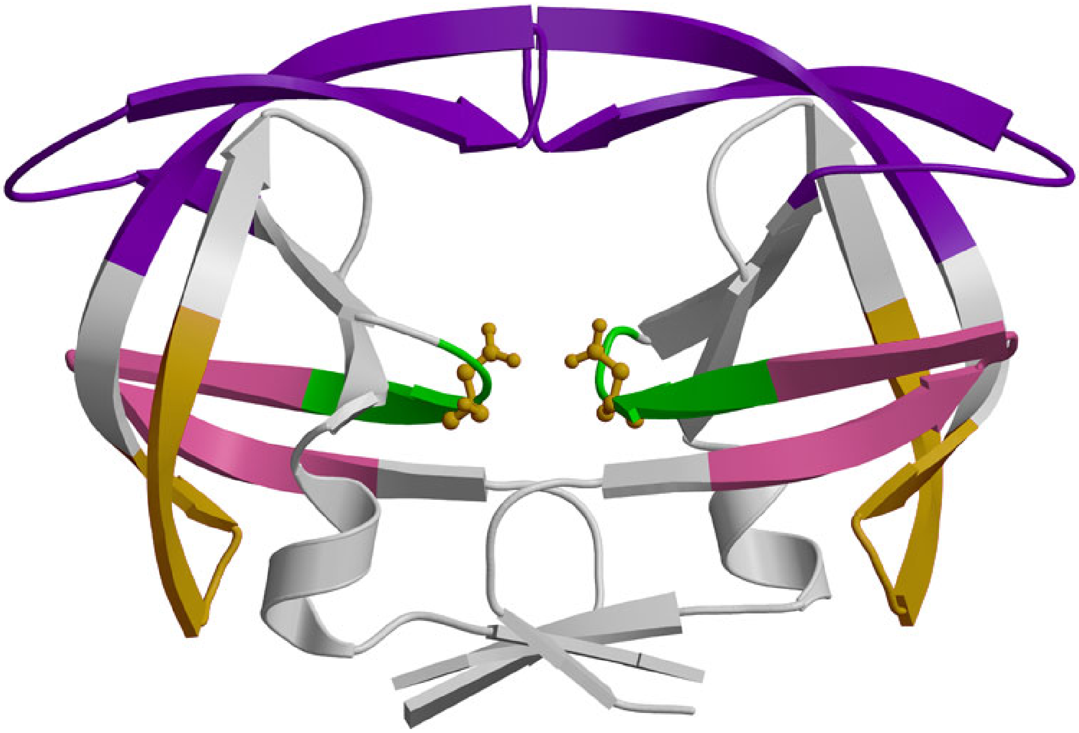


Figure 1.5: Flexible regions of HIV Protease.
Violet: Flap. Gold: Cantilever. Pink: Fulcrum. Green: Fireman’s grip. Catalytic aspartates are shown in ball and stick model. Redrawn from (32, 36).

1.1.6.3 HIVP substrate specificity

HIVP cleaves the Gag-Pol polyprotein at least at eight different locations; Table 1.1 presents the P_4 to P_3' amino acid sequences flanking these cleavage sites. The substrates can be grouped into two types, based on the amino acids immediately flanking the scissile bond: "–aromatic * Pro-" , and "–hydrophobic * hydrophobic-". The interactions of these substrates with HIVP are typically analyzed in terms of seven subsites, S_4-S_3', within the binding site (40). For example, the branched amino acids (Val or Ile) are preferred at the P_2 position in substrates of the –hydrophobic * hydrophobic- type, while Asn is prefered at this position in the –aromatic * Pro- types. Interestingly, although though there is no S_5 subsite, Lys at P_5 position significantly enhances the catalysis of substrates with a -hydrophobic * hydrophobic cleavage site, and also can significantly affect catalysis for the -aromatic*hydrophobic- junction (40). Thus, substrate residues at long distances from the scissile peptide bond contribute to HIVP specificity. However, although it is conceivable that entire flanking domains of the Gag-Pol polyprotine might also be involved in binding to HIVP and to substrate specificity, such interactions to not appear to have been reported in the literature.

JUNCTION	**P_4**	**P_3**	**P_2**	**P_1**	*	**P_1’**	**P_2’**	**P_3’**
Aro*Pro	Ser	Gln	Asn	Tyr	*	Pro	Ile	Val
	Ser	Phe	Asn	Phe	*	Pro	Gln	Ile
	Thr	Leu	Asn	Phe	*	Pro	Ile	Ser
Hydrophobic * Hydrophobic	Ala	Arg	Val	Leu	*	Ala	Glu	Ala
	Ala	Thr	Ile	Met	*	Met	Gln	Arg
	Pro	Gly	Asn	Phe	*	Leu	Gln	Ser
	Ala	Glu	Thr	Phe	*	Tyr	Val	Asp
	Arg	Lys	Val	Leu	*	Phe	Leu	Asp

Table 1.1: Substrate sequences from P4 to P3’.
The cleavage site is denoted by *. Reproduced from Griffiths *etal* (41).

1.1.6.4 Mechanism of catalysis

The aspartic proteases catalyze hydrolysis of their substrates through a general acid-base mechanism. This proteolysis takes place in four steps (Figure 1.6). The peptide carbonyl is hydrated by an active-site water in the first step of catalysis (Figure 1.6a). There is also translocation of a proton between the active site aspartates. In the second step, the scissile peptide bond adapts a *gauche* conformation (Figure 1.6b). The flexibility of the hydrated bond leads to this conformational transition. The *gauche* conformation facilitates proton transfer in the third step (Figure 1.6c). This proton exchange involves simultaneous proton transfers from one hydroxyl group of active site water to the charged aspartate and from the second aspartate to the nitrogen lone pair of the hydrated peptide bond. Rotation of the proton donor aspartate around the Cβ–Cγ bond is required for this proton exchange. The final step involves the breakage of C–N peptide bond and regeneration of the initial protonation state of catalytic aspartates (Figure 1.6d)(42).

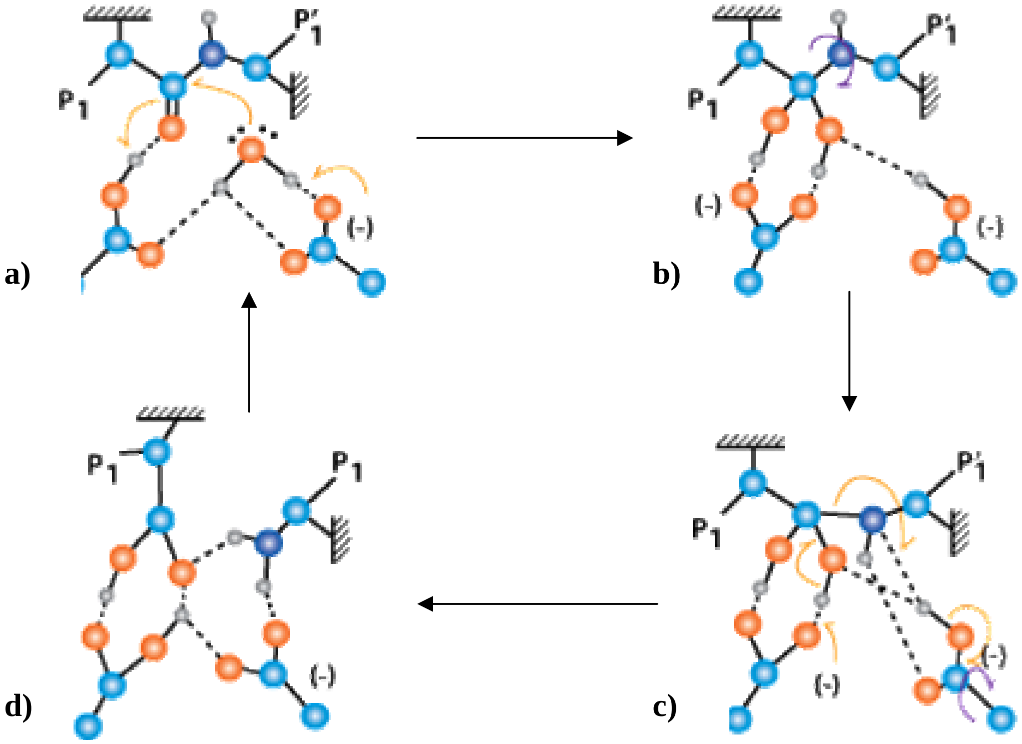


Figure 1.6: A model for HIV protease catalysis.
(a) Hydration of the scissile amide bond. (b) Conformational transition. (c) Simultaneous proton transfer. (d) Cleavage of the scissile bond. Reprinted from **(42)**.

1.1.7 HIV protease inhibitors

HIV protease inhibitors are competitive inhibitors, in which the scissile P1-P1' amide bond is replaced by a non-hydrolysable isostere such as a reduced amide, hydroxyethylene, hydroxyethylamine, azapeptide, etc (Figure 1.7) (43). Crystal structures of HIV protease-inhibitor complexes show that the enzyme structure is well preserved. This implies that the interaction pattern of inhibitors with the main chain of the protein remains the same, in spite of their differences in chemistry and structure. On superimposition, the functional elements were found to have a very good overall alignment. Most of the substrate-based inhibitors bind to the enzyme in an extended conformation(31). Inhibitors with a hydroxyl group at the non scissile junction position it between the catalytic aspartates, within hydrogen-bonding distance to at least one carboxylate oxygen of each aspartate (Figures 1.8a and 1.8c). Interaction of inhibitors with a buried water molecule, that bridges their P2 and P1' carbonyl groups and Ile50 and Ile50' NH groups of the flaps, is one more common feature (Figures 1.8a and 1.8b). This water is completely separated from the bulk solvent. Most inhibitors have hydrophobic moieties occupying P1 and P1' sub sites.

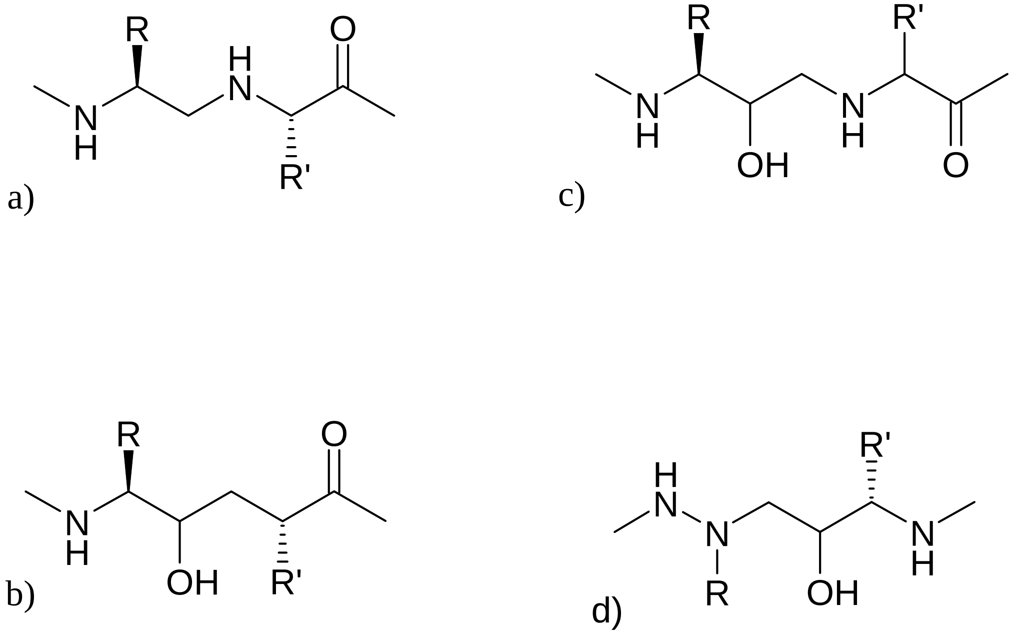


Figure 1.7: Examples of non-hydrolysable isosteres of the peptide bond.
(a) Reduced amide. (b) Hydroxy ethylene. (c) Hydroxy ethylamine. (d) Azapeptide (44).

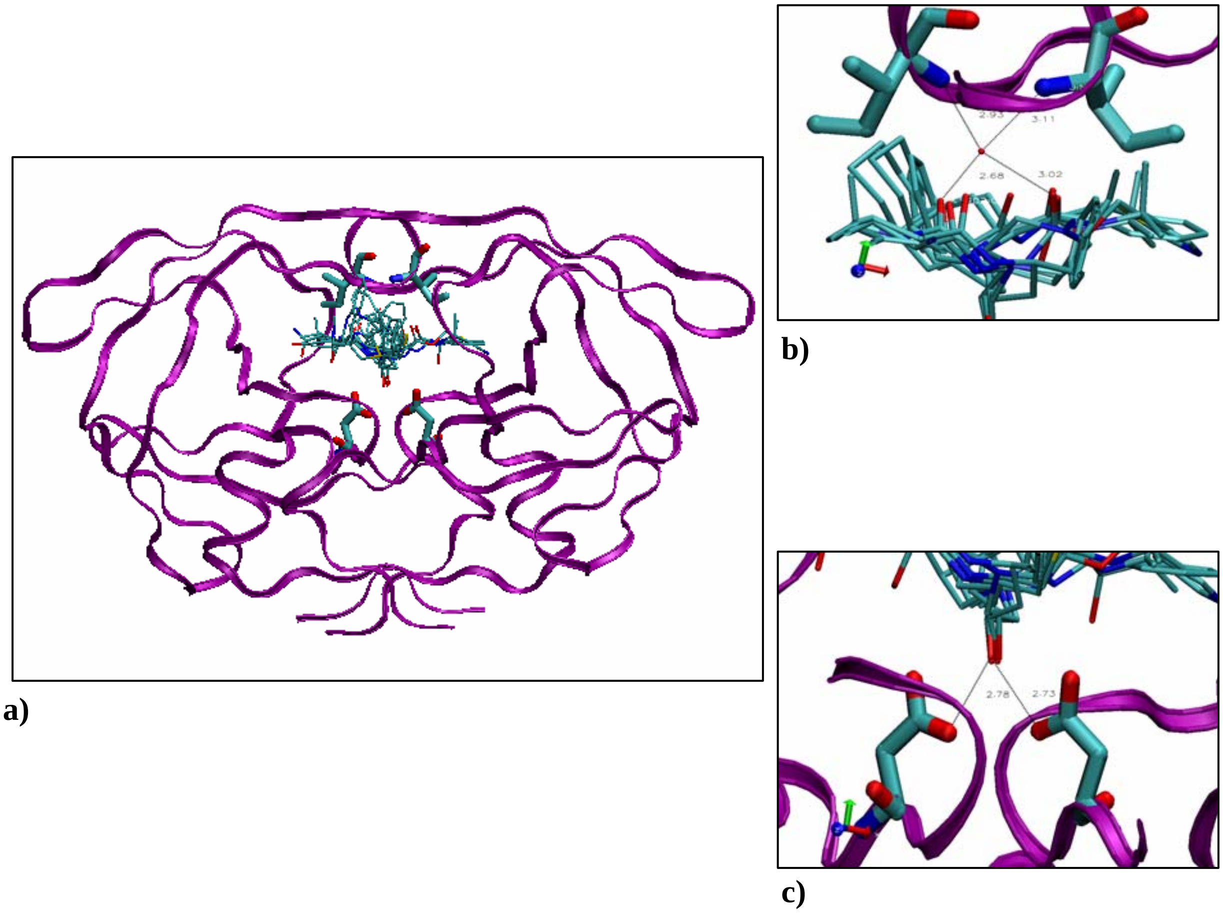


Figure 1.8: Superimposed HIV protease-clinical inhibitor complexes.
Crystal structures of Amprenavir, Indinavir, Saquinavir, Ritonavir and Nelfinavir in complex with HIV protease were superimposed to show the consensus hydrogen bonding interaction with flap water (a) and with catalytic aspartates (b). HIV protease is shown in ribbon model; the inhibitors, flap water and residues Asp25, Asp25’, Ile50 and Ile50’ are shown in licorice model.

Inhibitor-binding induces substantial changes in enzyme conformation. There is an approximate rotation of about 2 Å around the hinge region in the β-sheet interface. This rotation is accompanied by a large ~7 Å motion of the flap regions, which leads to flap closure and tightens the active site (31). This tight conformation of the active site excludes bulk water, providing the dehydrated environment needed for catalysis or for inhibitor binding (45). Upon binding, clinical inhibitors such as Indinavir, nelfinavir, saquinavir and ritonavir bury a large amount of largely hydrophobic surface area. Hence, the main driving force for their binding is, arguably, provided by the hydrophobic effect (45).

1.1.8 Clinical resistance

Despite the initial success of anti-AIDS therapy, nowadays there is a worrisome emergence of viral strains that exhibit resistance (45). It is estimated that almost 14% of new infections in America and 10% in Europe are by treatment-resistant strains. Its high replication rate and error-prone reverse transcriptase makes this virus remarkably prone to mutation. As a consequence, the appearance of HIV variants with decreased susceptibility to clinical inhibitors can be viewed as inevitable (44).

HIV protease, being a small enzyme and a homo-dimer, was initially thought of as an ideal target because of its limited mutational possibilities (46). There was also a speculation that resistance mutations would not readily develop to HIV protease inhibitors, because the protease need to process nine different substrates (47). However, rapid emergence of strains resistant to HIV protease inhibitors completely changed the initial perspective. There are at least 49 residues in HIV protease that undergo mutation, leading to resistance to one or more clinical inhibitors (45, 48). These mutations considerably reduce the affinity to inhibitors while retaining a viable enzymatic profile.

HIV protease mutations can be classified as primary or secondary, based on their order of appearance in patients undergoing treatment. Mutations in the binding site are almost invariably primary mutations. They are generally conservative, preserving the charge and polarity of the active site but not the geometry (49). As few as two such mutations often suffice to reduce the binding affinity of inhibitors several hundred-fold (50).

The effects of active site mutations on binding affinity are often relatively easy to interpret, as they are responsible for the loss of local van der Waals interactions or hydrogen bonds. The active site residues Asp30, Val32, Gly48, Ile50, Val82 and Ile84 (Figure 1.9), are highly prone to mutation. As these sites directly interact with the inhibitors, most of them are signature mutations for specific inhibitors. For example, signature mutations for saquinavir are G48V and I84V; for indinavir and ritonavir are V82A/T/F/S and I84V; for nelfinavir are D30N and I84V; for amprenavir are I50V and I84V; and for lopinavir V82A/T/F/S (45).

These active site mutations also affect the catalytic efficiency and maturation of Gag proteins, leading to impaired infectivity (51). Mutations at sites other than the active site have been found to compensate for this compromised catalytic activity (52). They also affects the binding of clinical inhibitors by causing a distortion in the geometry of binding site (53), and by stabilizing the open form of the protease (50). These effects of non-active site mutations were proposed to have a greater effect on binding of rigid inhibitors rather than the binding of substrates (54). Hence viral strains that show phenotypic and genotypic resistance to multiple drugs have only one or two active site mutations and a constellation of non-active site mutations (53). The diminished catalytic activity was also compensated to a certain degree by the co-evolution of substrates with mutations in Gag precursor p1/p6 and/or NC (p7)/p1cleavage sites (55-57).

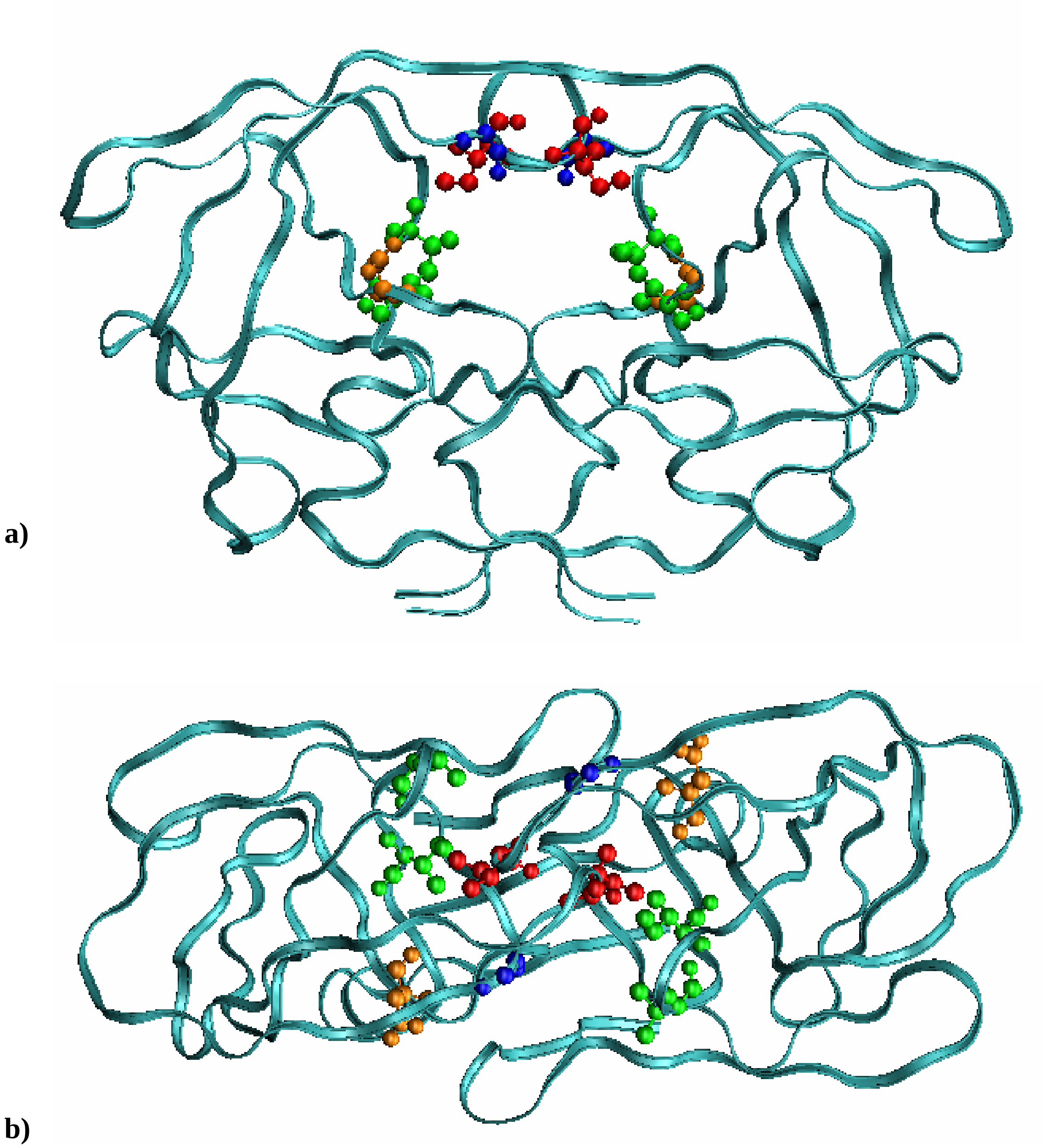


Figure 1.9: Sites of signature mutations.
Blue: Saquinavir. Red: Amprenavir. Orange: Nelfinavir. Green: Residues 82 and 84, that mutate with all clinical inhibitors. (a) Front view (b) Top view

1.1.9 Design strategies for mutation resistant HIV protease inhibitors

The emergence of viral strains that are resistant to existing clinical HIV protease inhibitors has necessitated a search for novel inhibitors with broad specificity against treatment-resistant strains. The goal of our project is to design a combinatorial library of such inhibitors. Several plausible design strategies deserve consideration. This section gives a short list of such strategies and discusses their merits.

The first strategy is to design inhibitors against multi-PI resistant variants. These variants carry various permutations and combinations of mutations in the protease. Hence it is not possible to include all the variants in screening, but it is possible to account for mutations at residues 10, 54, 71, 82 and 84, which occur frequently in multi-PI resistant strains. Therefore the strategy would be to design compounds that might not interact with these residues (58). Compounds having unique contacts with protease, such as TMC126, can be another category of novel mutation resistant inhibitors. The development of resistance to such an inhibitor will follow a different genetic pathway than the one due to the existing clinical protease inhibitors. Hence such inhibitors can be used in cocktail therapy with other protease inhibitors or for salvage therapy (59).

Inhibitors can be designed to have flexibility in the regions that interact with residues that are prone to mutate (60, 61). This idea is based on the observation that highly flexible peptide substrates are more amenable to adapt to backbone rearrangements or subtle conformational changes induced by mutations in the protease. In contrast, more rigid inhibitors lose much of their affinity on minor changes in the geometry of the binding site

(62). This differential susceptibility to mutation leads to clinical resistance. Hence inhibitors that are flexible may evade resistance. It is difficult though not impossible to incorporate this criterion in massive *in silico* screening.

The inhibitors that target open active site can be another set of candidate drugs. Most of the non active site mutations occur at domain interfaces, stabilizing protein in un-liganded (open) state and thus increasing the off rate for existing inhibitors. Therefore drugs targeting open active site can selectively avoid the mutations that confer resistance in this manner (63).

Design of compounds to have a volume consensus with substrates is another design strategy. It has been recently proposed that substrates are recognized by the protease by their volume rather than by their sequence or by their charge distribution. As these compounds would occupy the same volume as that of the substrates, any mutation that affects the binding of inhibitors would also affect the binding of substrates, and thereby affecting the viability of the virus. We selected this strategy for further scrutiny and evaluation in the design of a combinatorial library of mutation resistant HIV protease inhibitors. This strategy was also chosen based on its easy implementation and compatibility with our scoring algorithm (details are given in Chapter 2).

1.2 Combinatorial library design

1.2.1 Combinatorial library and virtual screening

Combinatorial library is a library of compounds that are related to one another, as they are built from same set of building blocks. Each compound in this library is a unique combination of these building blocks. There is a common structural core (combinatorial scaffold) between the members of this library, which has linking points for the building blocks (64). As there are huge number of available building blocks (synthetic reagents), it is not possible to synthesize and screen all possible combinations in the library for bio-activity. Identification of subsets of compounds from this vast combinatorial library that have the best potential for the discovery of new leads is a daunting task. Computational methods (Virtual screening) are sought to pick such promising candidates for synthesis and *in vitro* screening.

Based on the criteria used, virtual screening methods can be broadly classified into three classes. They are as follows: Cheminformatics-based, Ligand-based and Structure-based methods. The first class uses chemical descriptors or other chemical properties to select molecules based on drug-likeness, lead-likeness or diversity. The second class is used only when one or more active compounds were known. This method compares structural features of the screened compounds with the known actives. The last method is used when the structure of biological target and the key features of molecular recognition are known (65). Hence, selection of the virtual screening method depends on the level of available information. The last two methods are more predictable than the first method (65).

Cheminformatics-based methods are used when there is no prior information on the target or the actives for this target. In this situation, it is desirable to screen combinatorial libraries for diverse chemistry with drug and lead-like characteristics and also without undesirable chemical properties such as toxicity(66). This virtual screening method thus involve evaluation of molecular similarity or diversity, which can be assessed by comparing molecular features expressed in binary fingerprints (67). These fingerprints carry information on molecular features such as presence or absence of structural fragments, aromatic character, flexibility, and hydrogen-bonding capacity of molecules. There are a variety of metrics, that are available to compare the fingerprints (68).

Ligand-based methods are based on a pharmacophore model, which correlates molecular architectures with bio-activity. In this method compounds are screened for isofunctional molecular architecture that mimics the pharmacophore pattern in bio-active molecules. This method is especially useful, when there is limited or no structural information on the target. There are several software programs (HARPick (69), MoSELECT (70), TOPAS (71)) that use this method in the design of combinatorial libraries (72).

When the structure of a target protein is known, structure based screening methods are widely chosen because of their potential to discover diverse chemistry. They are also preferred for the structural insights, which can guide the design and optimization of lead compounds. In the structure based methods, suitable combinations of building blocks are suggested based on their predicted binding affinity for a given receptor. These methods can be broadly classified as *de novo* and scaffold based. In the *de novo* based method, optimal building blocks (chemical fragments) that interact favorably with sub sites in the binding cavity of a target protein are linked sequentially to build combinatorial compounds (73). Examples of such *de novo* based design algorithms include CombiSMoG (74),TOPAS (71), MCDNLG (73), LUDI (75). The disadvantage of these methods is that the choice of an optimal fragment for an earlier link in the growth phase may not be optimal further down the growth chain. Certain programs such as CONCERTS (76) circumvent this problem by allowing the links to break and reform during the growth phase.

Scaffold based design method involves linking building blocks (substituents) to an anchoring fragment (combinatorial scaffold) through the linking points (substituent positions). The combinatorial scaffold may be the one that provides the key interactions with target protein or a starting fragment in the synthetic route or the fragment to which other fragments are linked. Examples of this method include CombiDock (77) and PRO_SELECT (78).

Structure-based virtual screening methods predict the binding affinity of a given compound by docking them in the target active site and scoring the docked poses. Examples for the programs that are widely used for this purpose are DOCK(79), GOLD(80), AutoDock(81), FlexX(82), PRO_LEADS(83) and GLIDE(84). The following two sections give a brief introduction on the docking and scoring functions and on VDock(85, 86), one such function that we use in our combinatorial library design.

1.2.2 Docking and scoring

Docking is a computational method that uses the structure of a targeted protein to aid in the discovery of new ligands. It involves generating putative binding poses of candidate ligand in the target's active site and scoring these poses for their predicted binding affinity (65). Kuntz and coworkers pioneered this approach with their geometric method of fitting compounds into the binding site; this method was later named DOCK (87). Since then, numerous other approaches have been developed that vary in their scoring functions, their approaches to treating the flexibility of ligand and protein, and their algorithms for discovering low energy conformations of the ligand- protein complex.

There are three categories of scoring function: force-field based, empirical and knowledge-based. The force-field based methods compute the binding energy in terms of van der Waals and Coulombic potentials from a molecular mechanics force field. Docking programs that use this approach include AutoDock (81), DOCK (79) and VDock (85, 86). An empirical scoring function uses fitted hydrogen bonding, ionic and hydrophobic energy terms, which are calibrated based upon complexes of known affinity. The docking programs FlexX (82), AutoDock (87), ChemScore (88) and LUDI (89) use this approach. Knowledge-based potentials, such as BLEEP (90), PLP (91), PMFScore (92) and DrugScore (93), are based on observed atom-atom distributions among a large set of protein-ligand structures. (65).

Different docking proteins account for the flexibility of ligands and proteins in greater or lesser detail. One approach to handling ligand flexibility is to precompute a set of ligand conformations, and then dock them all as rigid components into the binding site. This approach is employed by a number of methods, including EUDOC (94), FLOG (95) and LigandFit (96). Another approach, used in programs such as FlexX (82) and DOCK (79), considers the ligand as a series of rigid molecular fragments. After positioning a key anchoring fragment in the active site, the other fragments are added to it in a step-wise fashion in such a way as to optimally fit the ligand into the binding site. Finally, a number of other programs use global optimization methods, such as Metropolis Monte Carlo (MC) (97, 98), Tabu search (83) and Genetic Algorithm (GA) (99), to seek the lowest-energy conformation of a flexible ligand in the binding site. Examples of this approach include GOLD (80), AutoDock (81), VDock (85, 86), GLIDE (84) and PRO_LEADS (65).

Docking methods are typically evaluated based upon their ability to reproduce the experimentally observed conformations of bound ligands. In many cases, the predicted conformations lie within a root-mean-square deviation (RMSD) of 2.0 Å On the other hand, the computed structure of lowest RMSD is not always the one with the most favorable docking score, due to errors in the the scoring function (65). Docking programs are also tested for their ability to score the known ligands of a targeted protein higher than a set of background ("decoy") compounds that are not thought to bind the target. When a mixture of known ligands and decoys are scored, most programs provide substantial enrichment of known ligands among the top-ranked compounds (74, 100, 101). However, it is important to recognize that performance on such tests depends upon the details of the comparisons. For example, it is typically easier to assign high scores to ligands that bind the target tightly than to lower-affinity ligands. Also, the known ligands are harder to differentiate from the decoy compounds when the decoys are more drug-like (65).

1.2.3 VDock

VDock is a force-field based docking algorithm developed in our lab. This program uses a rigid protein model. In the calculation of protein-ligand interaction energy, it is computationally expensive to include all non bonded interactions, as they are numerous. The fact that the protein atoms are fixed in the calculations, enables us to pre-compute the potentials (electrostatic potentials, and Lennard-Jones potentials) generated by all protein atoms in advance, and storing these potentials on a grid. Protein-ligand interaction energy can then be computed from the interaction between ligand atoms and grid points. These grids (electrostatic, steric and dispersive grids) with lattice spacings of 0.2 A° are generated using CHARMM 22 force field parameters and an implicit solvent model with a distance dependent dielectric constant, $\varepsilon = 4r$, where r is the distance between the grid point and the receptor atom (88). The grid dimensions are chosen to encompass all regions that mobile ligand atoms might enter.

VDock models the ligand as flexible by explicitly considering all torsional degrees of freedom along with translational and rotational mobility. As stated earlier, it is not possible to thoroughly investigate the huge conformational space that results from the inclusion of all degrees of freedom in the generation of conformers. Hence, VDock uses mining minima optimizer to find low energy conformations from this huge search space.

This mining minima assumes a hierarchical binding energy landscape, where low energy structures are close to each other and clearly separated from high energy structures. This optimizer combines principles from global-under estimator method (89) and GA (85). During the search using a pseudo global-under estimator, a large number of candidate ligand conformations are generated randomly within the sampling range, which is gradually narrowed around the current low-energy conformer. When a new low energy conformation is generated, the center of the sampling range is moved to this new conformer. Once the sampling range is narrowed to zero, a local energy minimum is found. This local minimum is kept in memory for use in later cycles of sampling by GA. A detailed description of GA is given in section 1.2.4.

Initial pseudo global-under estimator search results in multiple local minima, whose number is user defined. In GA, these local minima are recombined with random conformation through cross over. This recombination of partial solutions, resulting from the earlier pseudo global-under estimator search, accelerates the discovery of global optimum (85).

The energy of protein-ligand conformation generated during the search is computed from the sum of vanderWaals and coulombic interactions, using three pre-computed grids of potential fields, as follows.

$$E_{LR} = \sum_{i=1}^{Natoms} q_i \phi_j^{elec} + A_i \phi_j^{steric} + B_i \phi_j^{disp}$$

Equation 1.1

Where E_{LR} is the interaction energy, q_i is charge of atom, ϕ_j^{elec}, ϕ_j^{steric} and ϕ_j^{disp} are electrostatic, steric and dispersion potentials at j^{th} grid point, respectively. $A_i \equiv \sqrt{\varepsilon_i \sigma_i^6} / r_i^{12}$ and $B_i \equiv \sqrt{\varepsilon_i \sigma_i^3} / r_i^6$. ε_i, σ_i are the Lennard-Jones parameters and r is the distance between the atom and the grid point. The interaction energy of a compound is given by the energy of the lowest energy conformation generated in docking runs. These interaction energies serve as fitness for the tested compounds (85).

VDock had been shown to have comparable performance in the reproduction of crystallographic binding pose, with popular docking programs such as PRO_LEADS, AutoDock, FlexX, MCDOCK and GOLD (85). The second criterion for validating a scoring function is the enrichment ability. It is the ability to select known actives against a background of decoy compounds for a target protein. It is tested by docking both active and decoy compounds into the corresponding target protein. The entire set of compounds is then ranked as per their interaction energy. The concentration of known binders in the top of the ranked list determines the enrichment ability of the docking program. If the scoring had been ideal, all the actives will be concentrated in the top of the ranked list, where as if it had been a random selection, the actives will be distributed uniformly through out the ranked list.

Enrichment ability of VDock is shown in Figure 1.10. The reported actives and NCI diversity set were used as known actives and decoys, respectively for this experiment. The random selection of compounds would yield an enrichment plot along the diagonal. i.e, 10% of the actives can be recovered from top 10% of the ranked database. The greater the shift of the line towards left, greater is the enrichment. VDock is shown to have remarkable enrichment ability with our target protein, HIV protease. This justifies the use of VDock in our combinatorial library design.

Even though computational evaluation of a candidate compound by VDock can be fast (seconds to a few minutes of computer time per compound), the computational demands can become problematic for the exhaustive screening of virtual combinatorial libraries. Optimization methods such as GA and simulated annealing methods are therefore, used in selecting an experimentally tractable sub library from among this astronomical number of compounds that could in principle be synthesized.

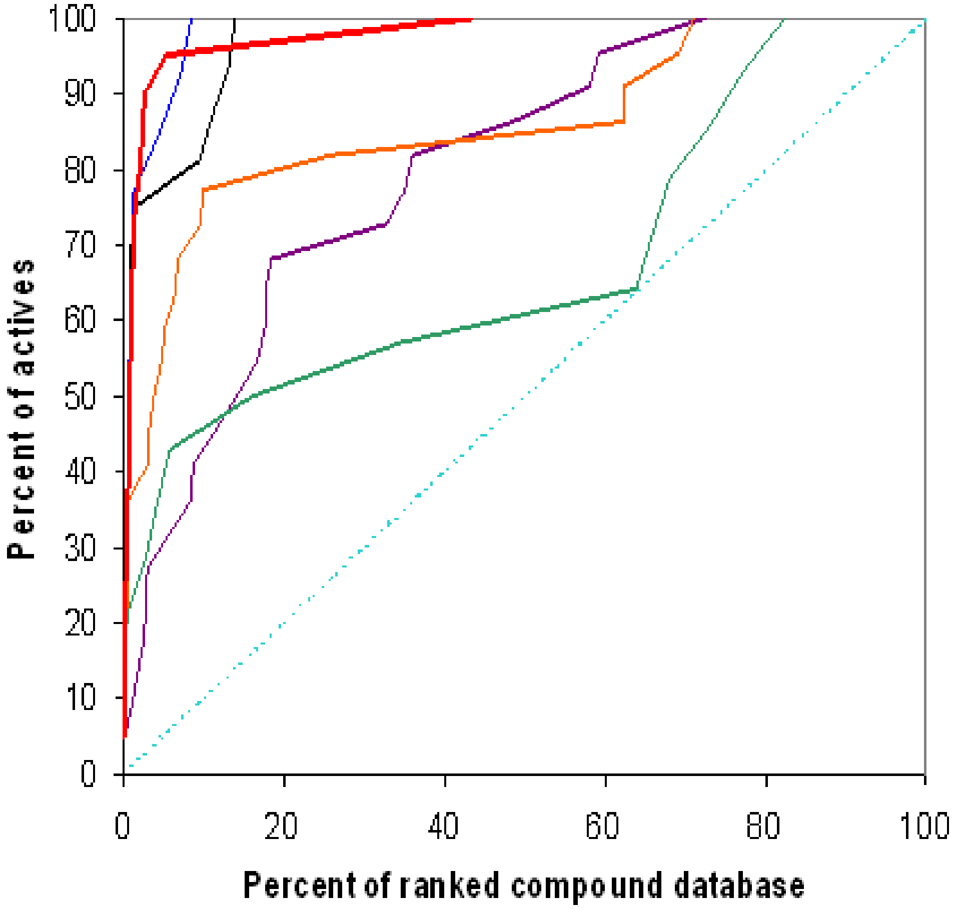


Figure 1.10: Enrichment graph for VDock with several target proteins.
Black: Factor Xa. Purple: Cyclin Dependent Kinase2. Green: Androgen receptor. Blue: Peroxisome Proliferator Activated Receptor. Orange: Neuraminidase. Red: HIV Protease. (Courtesy: Dr Visvaldas Kairys)

1.2.4 Genetic algorithm

Optimization algorithms are widely used in the design of combinatorial libraries that use both ligand-based (90, 91) and structure-based methods for the evaluation of compounds. One such method is genetic algorithm (GA), a stochastic method that mimics Darwinian evolution. In the combinatorial problem, each entity is a unique combination of a set of parameters that defines the position of this entity in the search space. Merit of each entity is a prospective solution to the combinatorial problem. The GA represents each individual entity as a "Chromosome", a linear series of genes. Each parameter is represented by a gene, which has a set of values that are represented by alleles. A random population of chromosomes is created in the first generation of the program, which are then scored for their fitness. A new population of chromosomes is constructed from the few selected, first generation chromosomes through cross over, mutation and reproduction. The chromosome selection for the propagation is biased to favor those with better fitness score. In the cross over operation, selected parent chromosomes swap a part of their chromosome resulting in two new chromosomes with different combination of alleles. Mutation results in change of one or more alleles in the parent chromosome. Reproduction is a mere replication of the parent chromosome. The new population is scored in turn and the iteration continues. Typically the average and maximum fitness score of the population increases with each generation until they converge to some maximum value. GA is especially useful in the search of large combinatorial search space, knowledge of whose terrain is very limited (90).

In the combinatorial library design problem, each chromosome represents a compound in the virtual combinatorial library, each gene corresponding to one substitution site, each allele of a gene corresponding to a candidate substituent (building block) at the site and the fitness of a chromosome corresponding to the interaction energy score predicted by VDock. The implementation of the genetic algorithm in our combinatorial library design is explained in detail under section 3.2.1

1.3 Overview of the thesis

This thesis is organized as follows. Chapter 2 introduces the substrate envelope hypothesis, describes the calculation of the fit of a candidate inhibitor to the substrate envelope with a grid based method, and evaluates the correlation of this score with susceptibility to resistant mutations. Chapter 3 compares a combinatorial library design strategy based on a simple additivity scheme that assumes independence among substitution sites, with a strategy based upon a genetic algorithm. Chapter 4 describes the application of the Additivity design strategy in the discovery of mutation-resistant HIV protease inhibitors. This process is found to yield two tightly binding compounds with desirable resistance profiles. Chapter 5 provides a general discussion and a prospectus for future research work. The major conclusions of the full project are summarized in chapter 6.

Chapter 2. The Substrate Envelope Hypothesis

2.1 Introduction

Inhibitors of Human Immunodeficiency Virus (HIV) protease revolutionized the treatment of patients infected with HIV in the mid-1990s, and remain a mainstay of therapy today (25, 92, 93). However, recent years have seen the emergence of HIV strains that are resistant to protease inhibitors (48, 94). The appearance of resistance is traceable to the selective pressure of therapy, combined with the high replication rate of HIV and the low fidelity with which HIV replicates its genetic information. Overcoming the evolutionary power of this system and maintaining an active armamentarium against HIV may prove to be a substantial challenge. On the other hand, this challenge is circumscribed by the fact that a viable resistance mutant of HIV protease must still bind and hydrolyze the various cleavage sites of the virus's Gag-Pol gene product at an adequate rate to allow viral replication. Accordingly, it has been argued that an inhibitor which forms "substrate-like" interactions with the protease should tend to evade viral resistance, because a mutation that weakens inhibitor-binding should simultaneously weaken substrate-binding, and hence damage the activity of the enzyme.

Recent crystallographic studies of complexes of HIV protease with its substrates provide a basis for pursuing this design concept. The substrates adopt a rather uniform shape when bound, despite the differences among their amino acid sequences, and the border of the consensus volume they occupy has been termed the "substrate envelope" (95). Intriguingly, the consensus volume occupied by a number of bound inhibitors differs significantly from the consensus substrate volume, and key resistance mutations appear to cluster near locations where inhibitors protrude outside the substrate envelope (96). These observations have led to the hypotheses that the protease recognizes its various substrates largely on the basis of their shape, and that inhibitors that fit within the substrate envelope may be less susceptible to mutational resistance (96). Indeed, it has been argued that the fit of inhibitor TMC-114 to the substrate envelope helps explain its ability to retain affinity for clinically relevant protease mutants (97, 98).

These considerations suggest that the substrate envelope hypothesis may be useful as a basis for the design of new inhibitors that will tend to counteract the emergence of resistance mutants. The present study addresses this issue by devising a quantitative indicator of the degree to which a candidate ligand falls outside the substrate envelope, and then determining whether this indicator correlates with the inhibitor's sensitivity to clinically relevant resistant mutations. The resistance analysis is based upon new calorimetric data for the association of various inhibitors with wild-type and mutant proteases, supplemented by additional calorimetric data from the literature.

2.2 Materials and methods

This section describes a method of quantifying the volume of a bound inhibitor falling outside the envelope, then details the measurement of affinities by isothermal titration calorimetry (ITC) for a group of inhibitors and proteases, and summarizes additional binding data drawn from prior publications. Finally, a novel measure of the "clinical relevance" of the mutant proteases is described.

2.2.1 Computational methods

2.2.1.1 Evaluation of the fit of an inhibitor to the substrate envelope

A 3D grid of substrate density in the binding site was generated as follows. The Superimpose module of QUANTA(99) was used to superimpose six crystal structures of HIV protease having bound substrate peptides (1F7A, 1KJ4, 1KJ7, 1KJF, 1KJG, 1KJH (37)) on a crystal structure of HIV protease with indinavir (1HSG(100)), based upon the coordinates of backbone atoms (Figure 2.1). The chemical C2 rotational symmetry of the receptor structure was accounted for by carrying out the symmetry operation and superimposing the six resulting structures on the original six by the same method, for a total of 12 overlaid substrates. Next, a cubic three-dimensional grid with side-length 10 Å and grid spacing 0.2 Å was centered on the active site, and an initial value of 0 was assigned to each grid point. Then a value *g(i,j,k)* was incremented by 1 for every substrate structure that contains the grid point *(i,j,k)*, where a grid point was considered to be contained by a substrate if it lies within the CHARMm (101) van der Waals radius of any non-hydrogen atom of the substrate. Because there are 12 overlaid substrates, the resulting grid values vary between 0 (outside all substrates) to 12 (inside all substrates).

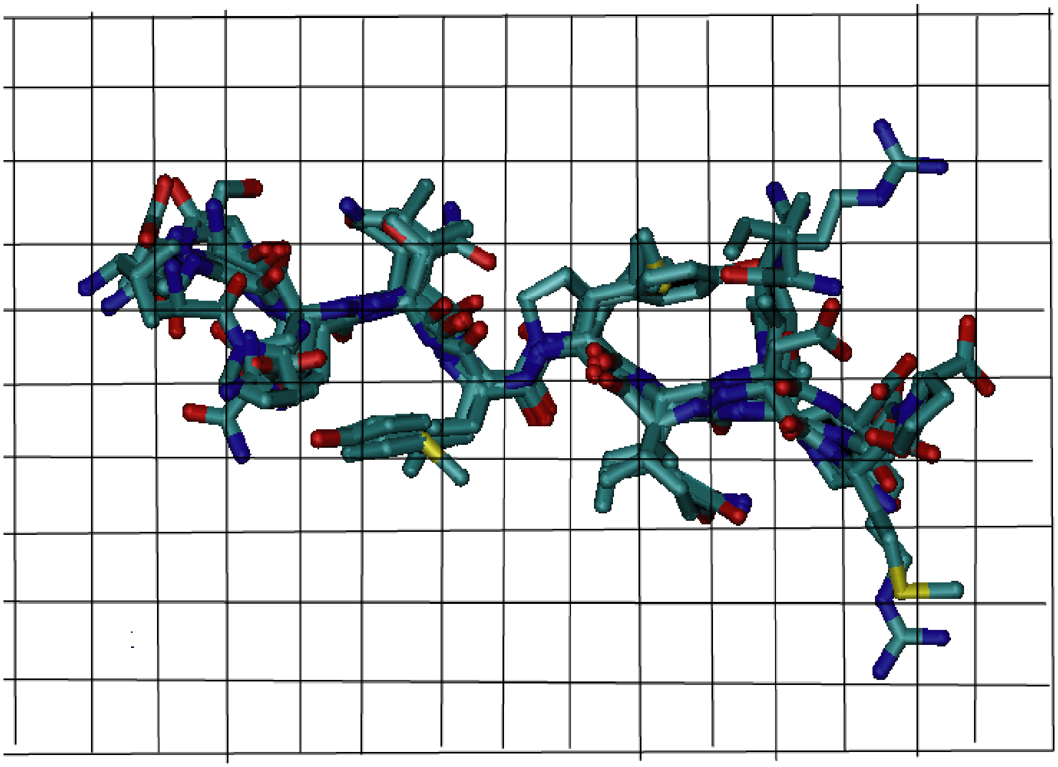


Figure 2.1: Alignment of HIV protease substrates on a 3D grid of substrate occupancy.

The fit of an inhibitor to the substrate envelope is computed as follows. A crystal structure of HIV protease with the bound inhibitor is aligned with the substrate-bound structures, as described above. Then the effective volume of the inhibitor outside the substrate envelope, V_{out}, is computed by summing the values of the grid points g_{ijk} that lie within the van der Waals volume of the inhibitor, normalizing the sum by 12, and converting to a volume by multiplying by the 0.008 Å^3 volume of a grid box:

$$V_{out} \equiv \frac{0.008}{12} \sum_{i,j,k}^{inside} (12 - g_{ijk})$$

Equation 2.1

Here "inside" implies that the sum runs only over grid points *ijk* that lie within the van der Waals volume of the inhibitor. As a control, the effective volume of the inhibitor that lies within the substrate envelope is computed as:

$$V_{in} \equiv \frac{0.008}{12} \sum_{i,j,k}^{inside} g_{ijk}$$

Equation 2.2

The total volume of an inhibitor, V_{tot}, is computed by adding these two quantities. The molecular weight and the number of non-hydrogen atoms were also included as alternative measures of molecular size.

The following crystal structures of HIV protease with bound inhibitors were drawn from the Protein Data Bank (33): 1HPV (102) (amprenavir; APV), 1HXB (103) (saquinavir; SQV), 1HSG (100) (indinavir; IDV), 1OHR (100) (nelfinavir; NFV) and 1HXW(21) (ritonavir; RTV). These structures were used to compute the values of V_{out}, V_{in} and V_{tot} (see above) of the respective inhibitors.

2.2.2 Binding data

The degree, to which an inhibitor's affinity declines when a mutant protease is substituted for wild-type, is quantified as $\log(K_d^{mut} / K_d^{wild-type})$, where K_d^{mut} and $K_d^{wild-type}$ represent the inhibitor’s dissociation constants for the mutant and wild-types, respectively. Dissociation constants from isothermal titration calorimetry (section 2.2.2.1) were obtained from our collaborators lab and from the literature (subsequent subsection). In each case, the ratio of mutant to wild-type is drawn from a single study to minimize noise due to experimental variations.

2.2.2.1 Literature Data

The new calorimetric data from the collaborators lab were supplemented with data drawn from the literature, including results for proteases with mutations only in the active site, only outside the active site, and both in and out of the active site. One study examines the consequences of mutants with a single mutation in the active site (I84V), multiple mutations outside the active site, (NAM10: L10I/M36I/S37D/M46I/R57K/L63P/A71V /G73/L90M/I93L) and their combination (ANAM11: L10I/M36I/S37D/M46I/R57K/

L63P/A71V/G73S/L90M/I93/I84V), upon resistance to a number of clinical inhibitors (53). A second study examines cooperativity among mutations V82A/I84V in the active site, M46I/I54V in the active site flaps, and L10I/L90M in the dimerization region away from the active site, as well as the combinations, HM (L10I/M46I/I54V/V82A/I84V/L90M) and QM (V82A/I84V/M46I/I54V) (104). A third study examines the active site mutation V82F/I84V against the background of viral strains A, B and C (105).

2.2.3 Clinical relevance of mutations

Treatment of a patient with HIV protease inhibitors selects for mutations that disrupt inhibitor binding while preserving enzyme function. If the substrate envelope hypothesis is valid, then inhibitors that fit the substrate envelope well should tend to retain affinity in the face of such clinically relevant mutations, but not necessarily to artificial mutations, which might disrupt the normal interactions of the enzyme with its substrates. The clinical relevance of the mutations studied here is assessed based upon their tendency to occur in patients treated with protease inhibitors, and in the absence of concurrent mutations known to be major resistance mutations. Thus, a combination of mutations is considered clinically relevant if clinical data suggest that it alone suffices to generate clinical resistance. Clinical data drawn from the HIV Drug Resistance Database (106) are used to define the clinical relevance C_i of a mutation set *i* as

$$C_i = 100 \frac{N_{i,only}}{N_{i,all}}$$

Equation 2.3

where $N_{i,only}$ is the number of isolates with mutation set *i* and no other major mutations, as defined at the Drug Resistance Database (http://hivdb.stanford.edu/cgi-bin/PRMut.cgi); and $N_{i,all}$ is the total number of isolates with mutation set *i*. That is, $N_{i,all}$ includes isolates with other major mutations.

2.3 Results

2.3.1 Fit of inhibitors to substrate density

Isodensity contours of the substrate density grid (Figure 2.2) show a rather smooth gradation of density, rather than a sharp drop from the maximal value of 12 to the minimal value of 0. The absence of an unambiguous substrate envelope motivates the present use of a smoothly varying measure of the volume of an inhibitor lying outside the substrate region (Equation 2.1), rather than a sharp cutoff. The consensus volume that is covered by all 12 substrate poses is rather constricted (red in Figure 2.2). It seems unlikely that an inhibitor could achieve high affinity without reaching outside this region. In fact, the inhibitors studied here all extend to some degree outside the level 8 contour, APV the least and RTV the most (Figure 2.3). This observation is quantified in the first row of Table 2.1, which lists the computed values of V_{out} (Equation 2.1) which range over a factor of 2. The computed volumes within the substrate envelope are more uniform, varying by only about 20%. Table 2.1 includes other measures of molecular size as well.

2.3.2 Binding affinities to wild-type protease and mutants

Table 2.2 lists the sensitivities ($\log(K_d^{mut} / K_d^{wild-type})$ of inhibitors to mutations MDR5, MDR3, and to 11 other mutant proteases for which data are drawn from the literature.

The affinity losses vary from 0.3 to nearly 4 logs. APV tends to lose least affinity to these mutations, while RTV tends to lose the most.

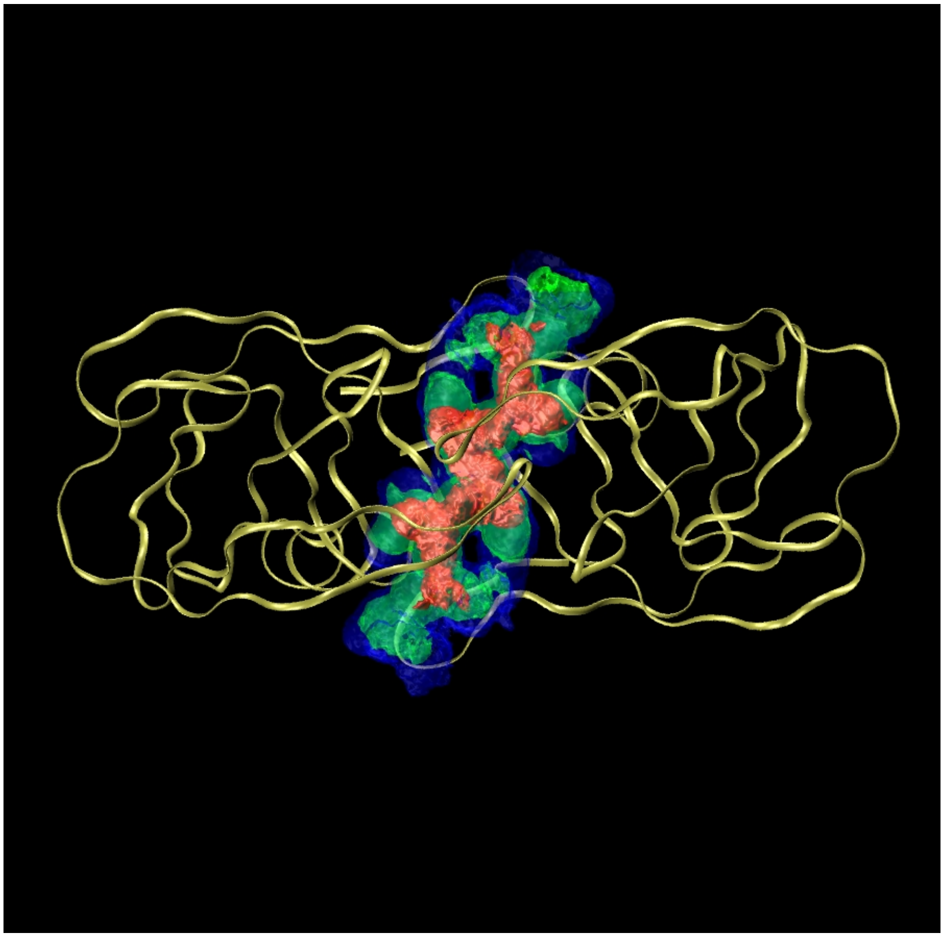


Figure 2.2: Isosurface contours of the substrate density.
Backbone trace of HIV-1 protease is also shown. Red: density 12. Green: density 8. Blue: density 4.

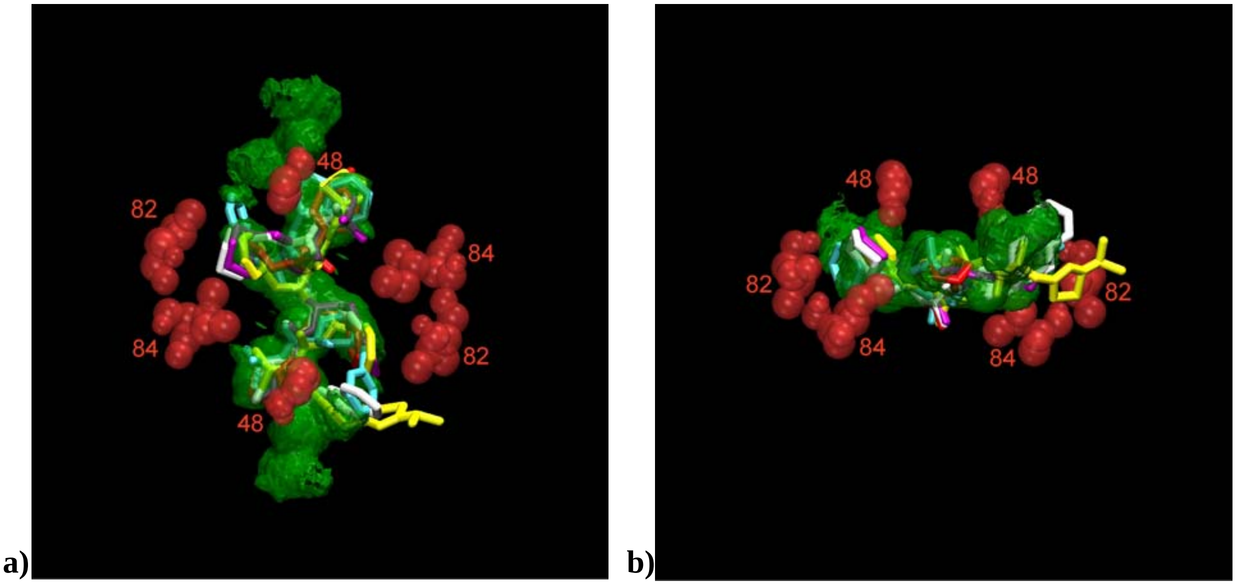


Figure 2.3: Two views of level 8 isodensity contours of the substrate density overlaid with crystal structures of clinical inhibitors.

The contours are shown in green color. Red: Amprenavir (APV). Cyan: Indinavir (IDV). White: Saquinavir (SQV). Purple: Nelfinavir (NFV) and Yellow: Ritonavir (RTV). The active site residues I82, V84, and G48 are shown with red spheres. The hydroxyl groups common to all five inhibitors protrude downward in (b).

	APV	IDV	SQV	NFV	RTV
$\boldsymbol{V}_{\mathbf{out}}$	128	180	213	166	256
$\boldsymbol{V}_{\mathbf{in}}$	267	315	319	288	308
$\boldsymbol{V_{tot}}$	395	495	531	454	564
Non-hydrogen atoms	35	45	49	44	50
Molecular weight (Da)	506	614	671	664	721

Table 2.1: Computed volumes (Å3) of inhibitors, with other measures of molecular size.

V_{out}: volume outside the substrate envelope. V_{in}: volume within the substrate envelope. V_{tot}: total volume.

MUTATION SETS	**APV**	**IDV**	**SQV**	**NFV**	**RTV**
MDR5 (L10I/G48V/I54V/L63P/V82A)	0.52	1.88	2.55	1.94	
MDR3 (L63P/V82T/I84V)	0.77	1.69	2.13	1.69	
NAM10 **(L10I/M36I/S37D/M46I/R57K/L63P/A71V/G73S/L90M/I93L)**		2.80	3.06	3.10	3.95
ANAM11 **(L10I/M36I/S37D/M46I/R57K/L63P/A71V/G73S/L90M/I93/I84V)**		2.98	3.29	3.13	4.56
I84V		0.57	0.60	0.54	1.73
V82F/I84V ***(Strain A)***		1.78	1.34	1.34	2.58
V82F/I84V ***(Strain B)***		1.83	1.32	1.30	2.57
V82F/I84V ***(Strain C)***		1.85	1.32	1.30	2.58
L10I/L90M	0.60	0.48	0.78	0.47	0.58
M46I/I54V	0.31	0.16	0.85	0.28	0.65
V82A/I84V	0.74	1.20	0.90	0.28	1.31
QM (V82A/I84V/M46I/I54V)	1.19	1.46	2.19	1.27	2.15
HM (L10I/M46I/I54V/V82A/I84V/L90M)	1.93	2.30	3.29	2.33	3.18

Table 2.2: Resistance values of five inhibitors and 13 HIV protease mutants.

Resistance: $\log(K_d^{mut} / K_d^{wild-type})$. K_d values for MDR3 and MDR5 were kindly provided by Dr Celia Schiffer. The other K_d values were obtained from the literature.

2.3.3 Clinical relevance of protease mutants

Clinical relevance of the studied mutations is given in Table 2.3. The mutants analyzed here appear to span a range of clinical relevance, as computed with Equation 2.3. The mutation sets of MDR5, QM and HM are found in a significant fraction (19-44%) of clinical isolates having no other major mutations; mutation sets MDR3, I84V, L10I/L90M and V82A/I84V appear in 2-5% of isolates with no other major mutations; and V82F/I84V, M46I/I54V appear in <1% of isolates without other major mutations. Mutation sets NAM10 and ANAM11 were not observed in any of the clinical isolates studied in Stanford database (48).

2.3.4 Correlation of V_{out} with sensitivity to clinically relevant mutations

Figure 2.4 examines the correlation of V_{out}, V_{in}, V_{tot}, number of nonhydrogen atoms, and molecular weight, with the loss of affinity of the various inhibitors on going from wild-type to the most clinically relevant protease mutants, MDR5, QM and HM. The data are drawn from Tables 2.1 and 2.2. The corresponding correlation coefficients for these mutants, and for the other, less clinically relevant mutants, are provided in Table 2.4. The volume of an inhibitor that lies outside the substrate envelope, V_{out}, correlates strongly with its susceptibility to the four most clinically relevant mutations, with correlation coefficients 0.94 – 0.97. Similar correlations are observed for many of the other mutants, but not all: the correlation coefficients range from 0.28 to 0.97. Interestingly, the other measures of molecular size show rather similar patterns. The correlations tend to be weakest for V_{in}, if only because this quantity has a rather small range of values.

MUTATION SETS	$N_{I,ONLY}$	$N_{I,ALL}$	**CLINICAL RELEVANCE**
MDR5 (L10I/G48V/I54V/L63P/V82A)	14	75	18.67
MDR3 (L63P/V82T/I84V)	1	53	1.89
NAM10 **(L10I/M36I/S37D/M46I /R57K/L63P/A71V/G73S/L90M/I93L)**	0	0	0
ANAM11 **(L10I/M36I/S37D/M46I/R57K/ L63P/A71V/G73S/L90M/I93/I84V)**	0	0	0
I84V	20	807	2.48
V82F/I84V *(Strain A)*	0	4	0.00
V82F/I84V *(Strain B)*	0	4	0.00
V82F/I84V *(Strain C)*	0	4	0.00
L10I/L90M	62	1264	4.91
M46I/I54V	1	390	0.26
V82A/I84V	4	166	2.41
QM (V82A/I84V/M46I/I54V)	7	28	25.00
HM **(L10I/M46I/I54V/V82A/I84V/L90M)**	4	9	44.4

Table 2.3: Clinical relevance of HIV protease mutants.

$N_{i,only}$: number of clinical isolates having the listed mutations and no other major mutations. $N_{i,all}$: total number of clinical isolates with the listed mutations. Clinical relevance is the percentage of $N_{i,only}$ to $N_{i,all}$.

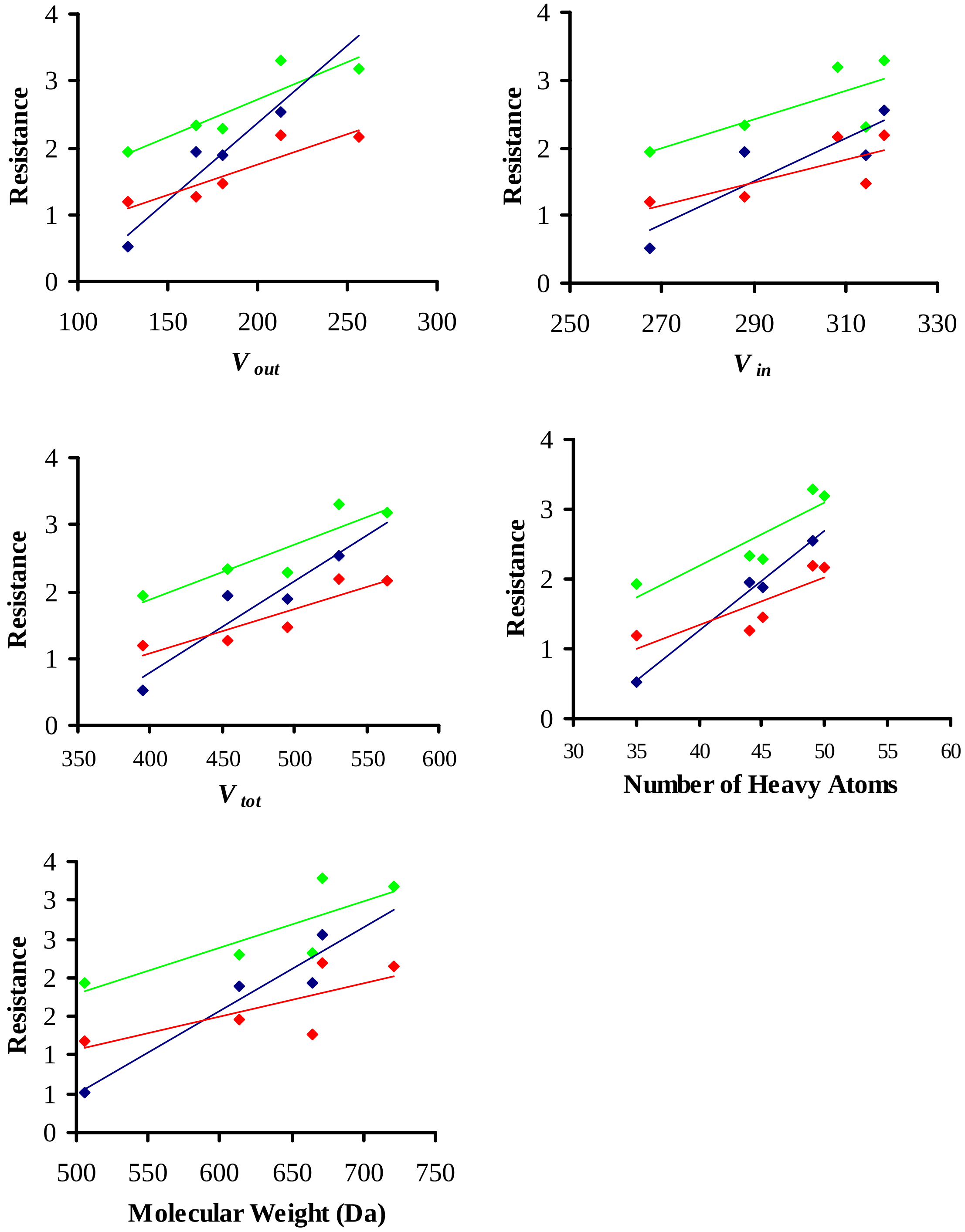


Figure 2.4: Correlation plots for the computed volume and other measures of size with the resistance.

Resistance is computed as the log loss in affinity on going from wild-type to mutant, for the three most clinically relevant mutants. Volume measures are given in cubic angstroms. Blue: MDR5. Green: HM. Red: QM.

MUTATION SET	V_{OUT}	V_{IN}	V_{TOT}	**NON-H ATOMS**	**MOL. WEIGHT**
MDR5 **(L10I/G48V/I54V/L63P/V82A)**	0.96	0.88	0.94	0.99	0.96
MDR3 **(L63P/V82T/I84V)**	0.97	0.89	0.95	0.99	0.95
NAM10 **(L10I/M36I/S37D/M46I/R57K/L63P/A71V/G73S/ L90M/I93L)**	0.74	0.20	0.57	0.53	0.84
ANAM11 **(L10I/M36I/S37D/M46I/R57K/L63P/A71V/G73S/ L90M/I93/I84V)**	0.94	0.17	0.84	0.79	0.83
I84V	0.89	0.10	0.77	0.71	0.81
V82F/I84V ***(Strain A)***	0.77	0.17	0.71	0.55	0.56
V82F/I84V ***(Strain B)***	0.75	0.20	0.69	0.52	0.50
V82F/I84V ***(Strain C)***	0.75	0.20	0.69	0.52	0.50
L10I/L90M	0.28	0.28	0.30	0.04	0.10
M46I/I54V	0.68	0.46	0.65	0.60	0.53
V82A/I84V	0.62	0.57	0.63	0.44	0.22
QM **(V82A/I84V/M46I/I54V)**	0.91	0.75	0.91	0.84	0.73
HM **(L10I/M46I/I54V/V82A/I84V/L90M)**	0.91	0.75	0.91	0.89	0.82

Table 2.4: Correlation coefficients of resistance with inhibitor properties.
Ligand characteristics and resistance are given in Table 2.1 and 2.2 respectively.

2.4 Discussion

The present results support the hypothesis that HIV protease inhibitors that conform better to the substrate envelope tend to be less susceptible to resistance mutations. The presumptive explanation is that a viable mutant must allow the protease to interact correctly with its substrates and so it will also tend to retain affinity for a substrate-like inhibitor. The present data do not definitively establish this mechanism, especially because nonspecific measures of inhibitor size, such as molecular weight, also are found to correlate with sensitivity to mutation. On the other hand, these additional correlations do not disprove the presumed mechanism; they may merely reflect the correlation of molecular weight, say, with V_{out}. Teasing apart the various correlations will require further studies. The ultimate aim of the present study, however, is to facilitate the design of new inhibitors that will resist mutation. It will therefore be of particular interest to observe the consequences of using the fit of candidate inhibitors to the substrate envelope as a figure of merit in computer-aided ligand-design.

The present analysis generalizes the notion of the substrate envelope to that of a substrate density, which falls rather gradually to zero from its maximum at the core of the substrate binding region. This approach provides more detailed information about the disposition of substrates in the binding site, as highlighted in Figure 2.2a, and avoids the need to set an arbitary level of substrate density at which to position a sharp substrate envelope. The substrate density is encoded on a 3D grid, allowing rapid calculation of the fit of a docked ligand to the substrate density. Analogous maps of ligand density, or of the density of a flexible receptor, could be useful in describing and modeling other molecular systems as well.

The present study evaluates the clinical relevance of protease mutations based upon the sequences of clinical isolates. For example, L10I/L90M occurs rarely in the absence of other major mutations (clinical relevance 4.9, Table 2.3), presumably because these two mutations alone confer less than one log of resistance to the clinical inhibitors that were studied here (Table 2.2). An alternative approach to assessing clinical relevance might have been to rely on in vitro vitality scores of the mutant proteases (45), which account for the enzymatic activity of the mutant against substrate. However, these data are unavailable for many mutants. In addition, the vitality may vary across substrates, whereas the clinical relevance score used here implicitly accounts for multiple substrates. This distinction may help explain why V82A/I84V (clinical relevance 2.4, Table 2.3) appears more frequently than V82F/I84V (clinical relevance 0, Table 2.3) in clinical isolates lacking other major mutations, despite the fact that the clinical inhibitors retain activity better against V82A/I84V than against V82F/I84V (Table 2.2), and both mutants affect the catalysis of a model substrate similarly (105). It is also worth noting that, although all the clinical inhibitors position the hydroxyethylene hydroxyl group outside the substrate envelope (Figure 2.3), this deviation should not provide a basis for resistance mutations because the hydroxyl contacts residues D25/D25', which are essential for catalysis and therefore are clinically irrelevant.

In summary, the failure of an HIV protease inhibitor to fit within the substrate envelope does appear to correlate with its susceptibility to mutational resistance. This is a low-resolution approach so exceptions will undoubtedly be found, and more rigorous approaches to identifying robust inhibitors are still needed. However, until such methods are available, the trend observed here suggests that designing ligands not only for tight-binding but also for fit to the substrate envelope could help accelerate the discovery of robust inhibitors of HIV protease.

Chapter 3. Combinatorial Library Design

3.1 Introduction

The compounds in a typical combinatorial library are built around a common structural scaffold possessing multiple connection points where substituents can be added by reliable synthetic steps (107). This format allows the efficient synthesis of many candidate inhibitors of a target protein. However, the number of compounds encompassed by such a combinatorial scheme frequently exceeds what can actually be synthesized and tested. This situation can be addressed by making and testing smaller sub-libraries where the compounds are selected based upon their similarity to known ligands (70) and/or their fit to the targeted binding site (77, 78). The present study focuses on the structure-based method, using ligand-protein docking. This method is expected to possess the advantages of yielding candidate ligands of diverse chemistries and of providing physical insight into interactions between ligand and protein. On the other hand, docking calculations tend to be more time-consuming than ligand-based methods, since one docking calculation typically requires seconds to minutes of CPU time. As a consequence, there is a particularly strong requirement for a library design algorithm that will make the best possible use of available computer resources.

One approach to the problem of library design is to apply a global optimization method, such as a genetic algorithm (108, 109). Successful applications of genetic algorithms in the design of both ligand-based combinatorial libraries (110-112) and the prediction of binding affinity have been previously reported (108, 109). A combinatorial sub-library designed via global optimization should perform significantly better than a randomly chosen sub-library. However, it is not clear whether global optimization methods can perform as well at picking chemical substituents as they do in more typical applications where the objective function depends upon better-defined degrees of freedom. A specific concern in the present application is that there is no guarantee that the GA will test every possible substituent at each position and, if it does not, the best substituents may well be missed.

Another approach to optimizing a combinatorial library using structure-based methods is to simplify the problem by assuming that the substituent sites can be optimized at least somewhat independently. This can be done, for example, by constructing an initial compound with an arbitrary set of initial substituents, and then making new compounds in which each position is converted to all other possible substituents, while the other substituents are held fixed. For a library with 4 substitution sites and 1000 candidate substituents at each site, this would yield almost 4000 compounds to be docked and scored, far fewer than the full virtual library of $1000^4=10^{12}$ compounds. The substituents which yield the best scores can then be selected and used to build a manageable set of compounds to be tested individually by docking and scoring. The approximation of independence provides for a dramatic acceleration in library evaluation, but if it is inaccurate, the best compounds may be missed. We are not aware of any systematic evaluation of the suitability of this approximation. Even though the idea of additivity has been employed in PRO_SELECT (78) and CombiDock (77), a comprehensive evaluation has not been discussed to date.

The present study evaluates the accuracy of additivity in docking and scoring of combinatorial libraries, tests its applicability to the design of targeted combinatorial libraries, and compares its productivity with that of a genetic algorithm method of library design, in applications to two model systems, HIV protease and cathepsin. For each system, three scenarios are considered: 1) design of sublibraries of a virtual library containing thousands of compounds with diverse substituents; 2) design of sublibraries of a virtual library containing thousands of compounds with substituents preselected to generate promising ligands; and 3) design of sublibraries of a virtual library containing millions of compounds with diverse substituents. The smaller virtual libraries allow more detailed characterization of the design methods because every compound in this library can be docked and scored. The larger libraries are more representative of real-world applications.

3.2 Methods

The methods section is organized as follows. The first three subsections describe the implementation of the design methods and their evaluation. The next two subsections describe the construction of combinatorial compounds, and the selection of substituent libraries for the studied test systems. The last two subsections provide details on the preparation of target protein structures and on the docking and scoring methodologies.

3.2.1 Genetic algorithm

A genetic algorithm (GA) is a stochastic optimization method that mimics the evolution of a population of chromosomes through a series of generations, where each chromosome represents a candidate solution for the optimization problem (108-112). For each generation, the objective function (fitness) is evaluated for each chromosome, and a new generation is formed by directly transferring a few top-scoring chromosomes (elites) from this generation to next one and by forming the rest of the population through processes mimicking crossing-over and single-site mutation. This process is repeated for a selected number of generations or until some criterion of convergence has been met. In this work, each chromosome represents a compound, each gene corresponds to one substitution site, and each allele of a gene corresponds to a candidate substituent at the site (Figure 3.1). The fitness of a chromosome is evaluated by constructing the corresponding compound and docking it to the target protein (section 3.2.7).

The GA used here is modified slightly for the sake of efficiency. The chief difference is that compounds in old generations are not discarded, but also are not brought forward into successive generations; i.e., no elite compounds are brought forward without modification. This change increases the number of compounds tested in one GA run. The second difference, based upon the first one, is that all chromosomes for the next generation are built by mutation and cross-over operations on parent chromosomes from not only the most recent generation but also from all prior generations. Parents are selected via the roulette-wheel method (113) where the probability of selecting a compound is proportional to its fitness rank. A predefined fraction of parents are used for cross-over operations; the rest are subjected to single-site random mutation. The cycle of evaluation, selection and modification is repeated for a user-defined number of generations, and the output is the ranked set of all compounds tried.

The GA includes several important operational parameters: the number of generations, the number of chromosomes in each generation, and the percentage of parent chromosomes subjected to cross-over. The total number of compounds tried, and hence the total number of docking calculations, equals the generation size multiplied by the number of generations. Except as otherwise noted, this total was kept as close as possible to the number of dockings used in the corresponding additivity-based calculations, so that the two methods could be compared on an equal footing. Preliminary calculations showed that the performance of the GA depends strongly on choices made regarding population size versus number of generations, and on the crossover versus the mutation rates. In particular, the mean fitness of compounds was found to improve only slowly after 8 generations, so all GA calculations used ~8 generations, and the desired number of dockings was set by adjusting the population size. The best overall results were obtained with a crossover rate of 25%, and this value was then used throughout.

The GA lends itself to parallelization on a loosely coupled computer cluster (114). In the present implementation, a "master" processor generates the chromosomes (compounds) for each generation, distributes *n* chromosomes apiece to *N* "servant" processors for docking and scoring, collects the results, and keeps a central list of all compounds tried and their scores. The servant processors decode the chromosomal representations into compounds, dock them, and return the scores to the master processor. The GA was implemented in parallel in the C++ programming language and using LAM/MPI libraries(115). For the smaller virtual libraries of several thousand compounds (Section 3.2.3), every compound in the virtual library was docked and scored. This allowed large numbers of GA runs to be carried out efficiently by construction of a lookup table of chromosomal fitness scores. This procedure is justified by confirmation of the consistency of the scores provided by the docking procedure used here (Section 3.3.1). For the larger libraries, *N*=60 processors were used, and each processor was sent *n*=4 chromosomes at a time. All results presented for the GA method are averages over 3 independent GA runs initiated with different random number seeds.

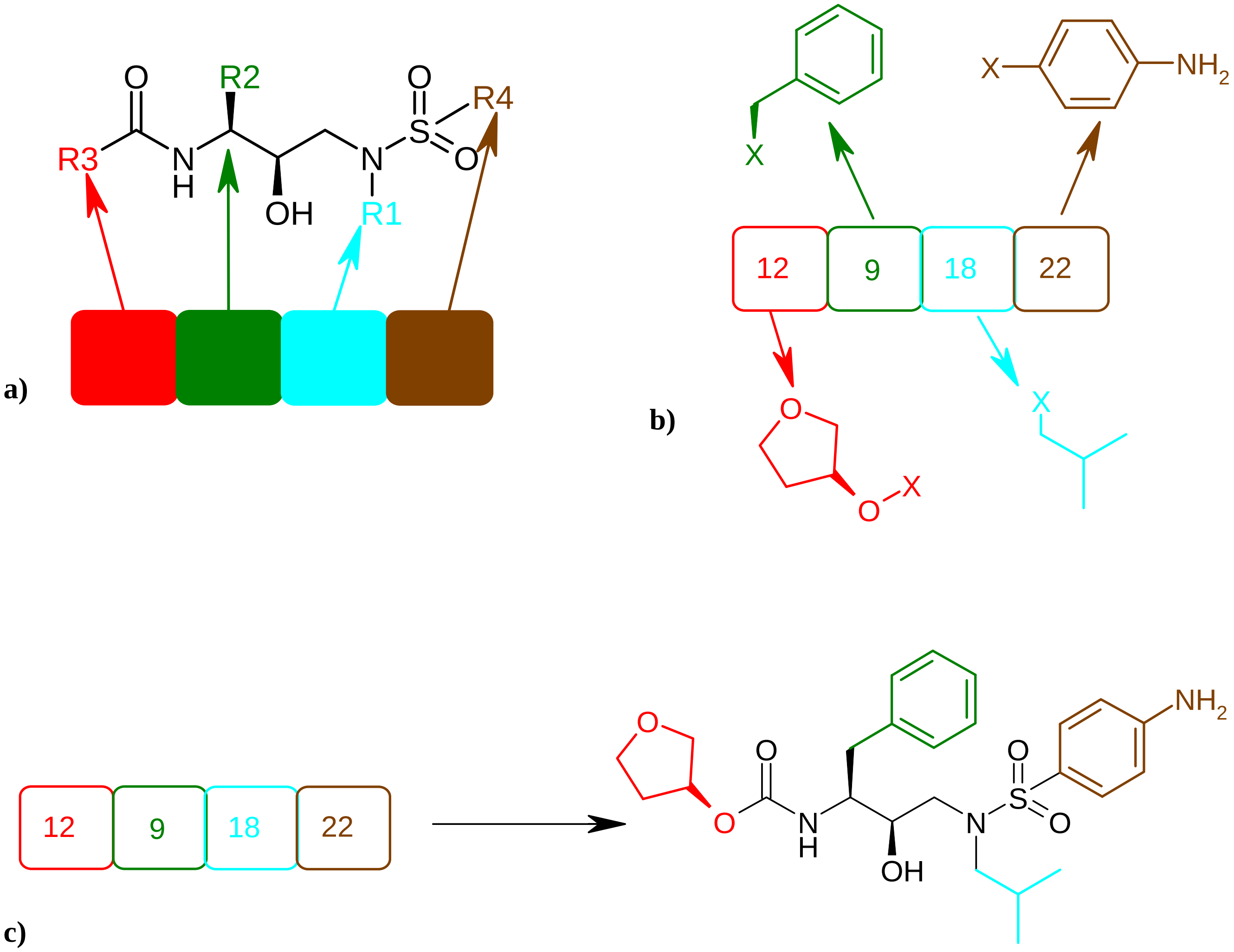


Figure 3.1: Chromosomal representation of a compound.
(a) Each gene in a chromosome represents a substituent position. (b) Each allele of a gene represents a unique substituent in the corresponding substituent library. (c) Each chromosome hence can be translated to a unique ligand

3.2.3 Additivity method

3.2.3.1 Additivity approximation

The additivity method is based upon the assumption that the difference in binding energy of two compounds in the combinatorial library can be approximated by a sum of the relative contributions from each substituent (116). Thus, for a scaffold with four substitution sites, the binding energy E_{ijkl} of a compound with substituents *i ,j ,k ,l* at each of the four sites, respectively, is estimated in terms of the binding energy of a reference compound E_{0000} with substituents *0,0,0,0*, and the change in binding energy when each substituent *0* is replaced independently by *i ,j , k* and *l*, respectively:

$$E_{ijkl} \approx E_{0000} + (E_{i000} - E_{000}) + (E_{0j00} - E_{0000}) + (E_{00k0} - E_{0000}) + (E_{000l} - E_{0000})$$

Equation 3.1

The success of the method depends upon the validity of this approximation.

Continuing with the present example of a 4-site scaffold, the additivity approximation is applied by carrying out docking calculations for a reference compound (0000) and for all compounds that can be made by replacing one substituent in the reference compounds with another substituent; that is, compounds (i000), (0j00), (00k0) and (000l), where *i, j, k* and *l* take on all possible values other than 0. This yields E_{0000} and all possible values of E_{i000}, E_{0j00}, E_{00k0} and E_{000l}. Thus, if there are 100 substituents for each of the four sites, then only 1+(4)(99)=397 docking calculations are needed to generate the quantities needed to estimate the binding energies of all 100^4 compounds in the full library, via Equation 3.1.

3.2.3.2 Choosing compounds with the additivity method

The top ranked compounds can be found by using Equation 3.1 to estimate the docking energy of all compounds, and then sorting them all, but this can lead to very large sorting calculations. A simple algorithm for identifying the top-ranked compounds was therefore devised. First, the substituents available at each site are sorted according to their energy scores. The top-ranked compound is simply the one with all of the top-ranked substituents. (See Figure 3.2, row 1, where the substituents' indices are simply their ranks.) The second-ranked compound, must be one of the following compounds: (2,1,1,1), (1,2,1,1), (1,1,2,1), or (1,1,1,2), (Figure 3.2, row 2), which can be identified by computing these compounds' energies with Equation 3.1. Here, the second compound is taken to be (1,1,2,1) (Figure 3.2, row 3). The third-ranked compound is found similarly: four new compounds are formed by replacing each site of the second-best compound with the next best substituent (Figure 3.2, row 4), grouping the new compounds with all others that have previously been generated (Figure 3.2, rows 4-5), eliminating compounds that are clearly not candidates (Figure 3.2, red boxes) to form a reduced set (Figure 3.2, row 5), and choosing the best compound (here taken to be (1,2,1,1)) according to Equation 3.1 (Figure 3.2, row 6). This process is repeated until the desired number of top-ranked compounds has been identified. Selected compounds from this method are then scored by docking and sorted to find the top-scoring compounds among them.

3.2.3.3 Pair wise additivity

The additivity approximation described in Section 3.2.2.1 uses a single-site decomposition of the binding energy. Other decompositions also are possible and may be useful when there is reason to believe that substitution sites do not affect the binding energy independently. For example, the substituents in the present four-site example can be grouped into two pairs, sites 1 and 2, and sites 3 and 4, and the energy of molecule (ijkl) can be estimated as

$$E_{ijkl} \approx E_{0000} + (E_{ij00} - E_{000}) + (E_{00kl} - E_{0000})$$

Equation 3.2

This pair-wise approximation requires calculating the energies of all pair-wise substitutions at sites 1 and 2 (E_{ij00}) and at sites 3 and 4 (E_{00kl}). If 100 substitutions are possible at each site, then evaluating these pair-wise energies requires (2)(99^2) or about 20,000 dockings. This is more demanding than the single-site method, but may increase the accuracy of the energy predictions.

3.2.4.4 Reference compounds

The additivity approximation requires a reference compound, compound (0,0,0,0) in the 4-site example. Amprenavir (117) is used as the reference compound for amprenavir-like combinatorial libraries targeting HIV-protease. For the cathepsin system, no crystal structure is available with any compound based upon the combinatorial scaffold used here. Therefore, three different reference compounds were tried; all were drawn arbitrarily from the small virtual combinatorial library. For the larger virtual combinatorial library, only one arbitrarily chosen reference compound was used.

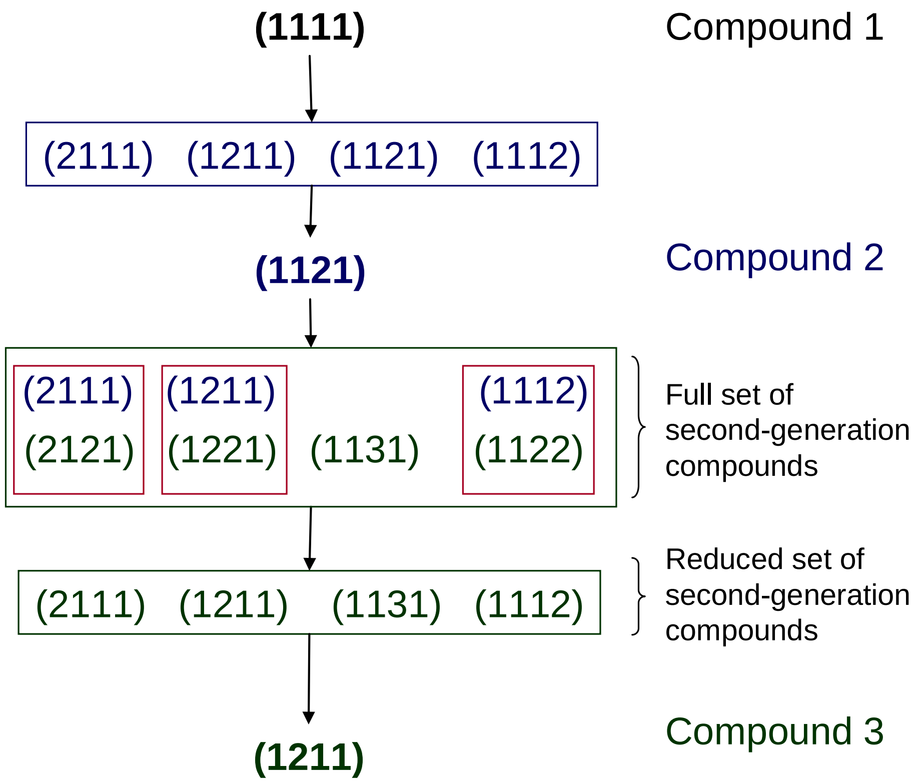


Figure 3.2: Combination of optimal substituents by single-site additivity method. Compound 1 is the combination of the best substituents for all sites. A first generation of compounds is generated by substituting each site with next best substituent, one at a time. In the present example, Compound 2 is taken to be the best of this generation. This iteration of substitution and selection continues until a predetermined number of compounds have been generated. The ranks of the substituents are represented by their indices, and the best compound from each generation is shown in bold font. The red blocks highlight pairs of compounds of which one (bottom) can trivially be eliminated in favor of the other (top).

3.2.3 Evaluation of additivity and GA methods

In each comparison, the GA and Additivity methods were run for a preselected number of docking calculations. The number of dockings for the GA, N_{GA}, is the product of the number of generations and the generation size. For the single-site Additivity method, the total number of dockings, N_{add}, is given by $N_{add} = 1 + \sum_{i=1}^{N_{sites}} (n_i - 1) + N_{combinations}$. Here the 1 accounts for docking the reference compound, N_{sites} is the number of substitution sites on the scaffold, n_i is the number of candidate substituents at site i, and the summation represents the number of dockings required to generate the parenthesized quantities in Equation 3.1. Finally, $N_{combinations}$ reflects the additional work of computing the actual docking energies of the specific compounds predicted by Equation 3.1 to be top-scorers. The value of $N_{combinations}$ is chosen to make the total number of dockings for the Additivity method equal to N_{GA}; hence $N_{combinations} = N_{GA} - \left(1 + \sum_{i=1}^{N_{sites}} (n_i - 1)\right)$.

For the smaller virtual libraries (Section 3.2.5), the compounds generated by the GA and Additivity methods are assessed by two measures. The first measure is the fractional recovery of the top-scoring 5% of compounds ("computational binders") in the full virtual library. This quantity can be evaluated only if all compounds in the virtual library have been docked and scored, and therefore can be used only for the smaller virtual libraries (Section 3.2.5). The second measure is the docking energies of the top 5% of the compounds in the designed library; this measure can be applied to the larger virtual libraries as well. For the large virtual libraries (Section 3.2.4), 6000 dockings were done for both the GA and single-site Additivity method, and the distribution of the scores of the top 1000 compounds generated by the two methods were analyzed. The pair-wise additivity method was not tested for the larger virtual libraries, as it would have required more than the allotted 6000 dockings to generate the pairwise terms found on the right hand side of Equation 3.2.

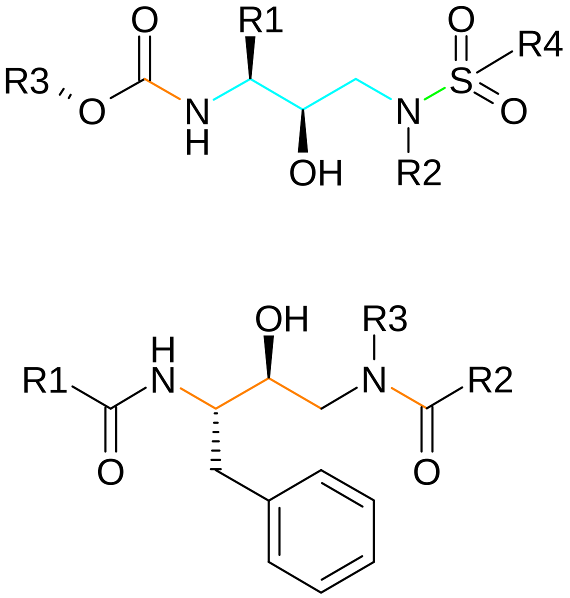


Figure 3.3: Combinatorial scaffolds.
Amprenavir-like (top) and pepstatin-like (bottom) scaffolds were used in the construction of combinatorial libraries. Torsional freedom along the light brown, cyan and bright green colored bonds was restrained to ranges of 40, 60 and 180 degrees respectively.

3.2.4 Construction of compounds

The docking calculations (Section 3.2.7) vary only the rotatable bonds and therefore require reasonable initial 3D structures of each candidate ligand as input. They also require molecular "topologies" comprising bonding information and force-field energy parameters, such as the Lennard-Jones parameters used in computing van der Waals interactions. These structures and topologies were prepared as follows. The initial 3D structure of the amprenavir-like and pepstatin-like scaffolds were drawn from the crystal structures 1HPV (117) and 1LYB (118), respectively. The initial 3D structures of all substituents (Section 3.2.5) drawn from the ZINC database (119) were taken as-is from Zinc. The initial 3D structures of the preoptimized substituents for the cathepsin system were prepared with Quanta (99). The initial 3D conformation of a compound in the virtual library was assembled when required by overlaying linking bonds and setting to ideal bond-length based on the atom types of connected atoms. "Ab initio"-like partial charges of the assembled compounds were generated with VCharge, an electronegativity equalization method parameterized to reproduce electrostatic potential fields computed at the 6-31G* level (120). The resulting partial charges closely resemble those of CHARMM and AMBER for amino acids and nucleic acids (120). Dreiding force-field parameters were used for bond-torsions, and CHARMM parameters were used for Lennard-Jones interactions.

3.2.5 Substituent libraries

3.2.5.1 HIV protease system

Three sets of substituents were prepared, one with 100 candidate substituents at each site, and two with 15 candidate substituents at each site. All carboxylic acids and primary amines were drawn from the building blocks collection of Zinc, and the carboxylic and amine groups were deleted and replaced with dummy linker atoms. Thus, other functional groups could have been drawn from Zinc; the present choice was one convenient option. A diverse set of 100 candidate substituents for each substituent site was drawn from these initial Zinc fragments, as follows. The program Dragon (42) was used to compute, for each candidate substituent, the molecular weight, number of hydrogen bond donors, number of hydrogen bond acceptors, number of rotatable bonds, aromatic ring density, logP, and surface area. The Euclidean distance, in the resulting descriptor space, was computed between each substituent of amprenavir and each candidate substituent and, for each substitution site, the 30 closest and 30 most distant candidates were chosen, along with 40 candidates of intermediate distance. This procedure produces four different substituent libraries of 100 compounds, one library for each site of the amprenavir-like scaffold.

Substituents for the small diverse library of size 15 fragments were drawn from this initial library of 100 diverse substituents, by choosing compounds at regular intervals of distance from the baseline amprenavir substituent. A set of 15 preoptimized substituents for each scaffold site was formed by computing all 397 values of E_{i000}, E_{0j00}, E_{00k0}, and E_{000l} (Section 3.2.2) for the 100 substituents, using amprenavir as the reference compound (0000), and choosing the top-scoring 15 substituents at each site for further study.

3.2.5.2 Cathepsin system

A set of 25 candidate substituents preoptimized for the cathepsin target was drawn from a prior study (109); note that the same substituents were used at each position. A small diverse set of 25 candidate substituents was constructed from random compounds in the R1 fragment library of the amprenavir system (Section 3.2.5.1), along with 5 substituents from, the above mentioned, preoptimized set. A larger set of diverse substituents was constructed by supplementing the preoptimized 25 substituents with the 100 candidate R1 substituents prepared for the HIV protease system, for a total of 125 possibilities at each site.

3.2.6 Protein structures

Candidate ligands for HIV protease and cathepsin were docked to PDB (33) structures 1HPV (117) and 1LYB (118), respectively. In both cases, the ligand and other nonprotein atoms were removed from the binding pockets, except for the flap water in the case of 1HPV. The program Quanta (99) was then used to add all polar hydrogen atoms and their positions were optimized by energy minimization with only hydrogen atoms free to move.

The structure of HIV protease and its substrates specificity has already been described in section 1.6. A brief description on cathepsin, its structure and its specificity is as follows.

Cathepsin D is a mammalian aspartic protease found primarily in lysosomes and suspected of involvement in a variety of diseases such as cancer, Alzheimer's (121) and muscular dystrophies (118). At 366 residues, it is somewhat larger than HIVP. Its crystallographic structure reveals a fold similar to that of other mammalian aspartic proteases, such as renin, chymosin and pepsin (118). It also shares a close-packed core of conserved, hydrogen-bonded polar residues. It structure can be divided into an N-terminal domain, a C-terminal domain, and an interdomain connecting both terminals. The N- and C-terminal domains each contribute one aspartic acid to the active site, which lies in the deep cleft formed by both the domains (Figure 3.4) (122). This cleft is wider than that of renin, so cathepsin can bind larger substrates and inhibitors (118).

Cathepsin D can bind to substrates 9 amino acids long, which occupy sites from S_5 to S_3' in the active site. Its substrate specificity is not yet fully characterized, as its role in human physiology is not completely understood. Like HIVP, it prefers hydrophobic residues around the scissile bond (122). It also has a strong preference for a hydrophobic residue in the P_2 site of the substrate, but can still accommodate hydrogen bonding residues at this location. However, the P_2 site does not accept cationic amino acids. The presence of Met in the S_2 subsite may help explain its especially strong preference for hydrophobic substrate residues, relative to other aspartic proteases.

Binding of the inhibitor pepstatin to cathepsin D induces small structural changes in the "flap region" (residues 72-87, Figure 3.5) and in a proline-rich loop (Figure 3.5) (118). In contrast, the flap regions of both monomers of HIVP are believed to move substantially on binding (section 1.1.6.2). Numerous hydrogen bonds between the main chain atoms of bound pepstatin with the active site residues stabilize the complex. In addition, the hydroxyl of the central statine group of pepstatin forms hydrogen bonds with the catalytic aspartates much as the core hydroxyl of most HIVP inhibitors interact with the catalytic aspartates of HIVP.

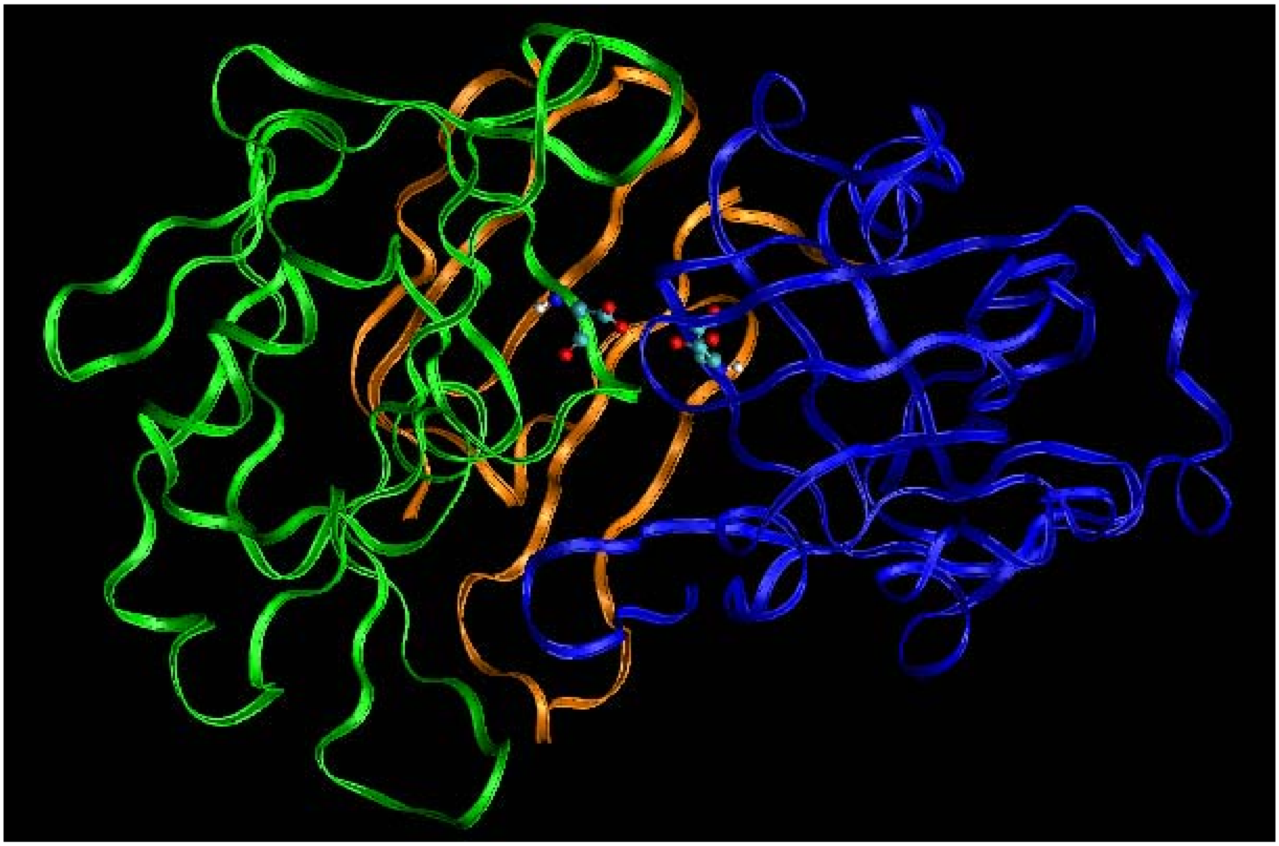


Figure 3.4: Cartoon representation of Cathepsin D

Blue: N-terminal domain. Green: C- terminal domain. Orange: Interdomain. Catalytic aspartates are shown in ball and stick model

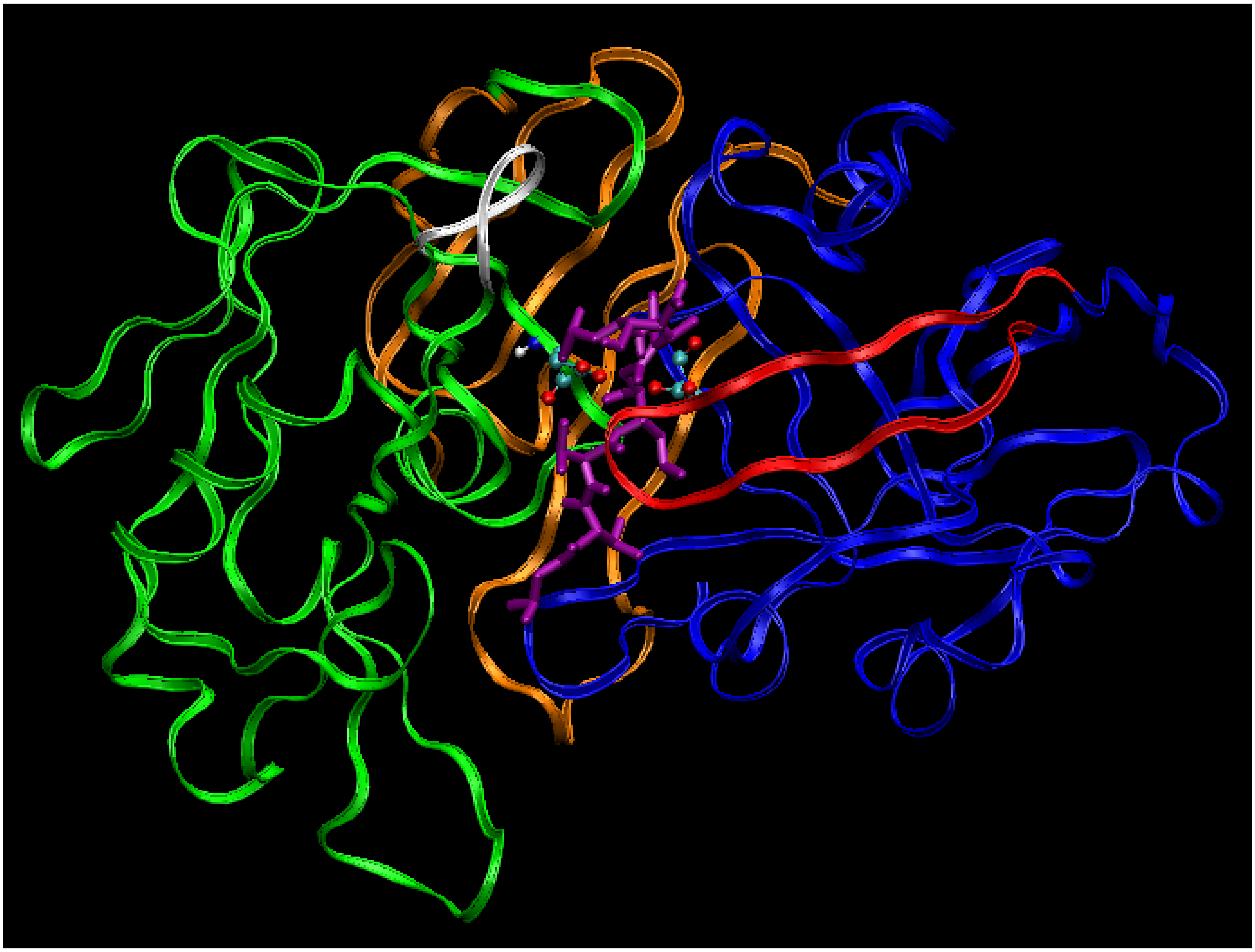


Figure 3.5: Flap region and proline loop of Cathepsin D

Blue: N-terminal domain. Green: C- terminal domain. Orange: Interdomain. Red: Flap region (residues 72 to 87). White: Proline loop (residues 312 to 317). Catalytic aspartates are shown in ball and stick model. Inhibitor, pepstatin is shown in licorice model.

3.2.7 Docking and scoring

Candidate ligands were docked and scored with VDock (85, 86), using a distance-dependent dielectric constant with a coefficient of 4. In order to accelerate convergence, the scaffold of each compound was restricted to lie near the pose observed for amprenavir in the HIV protease tests and pepstatin in the cathepsin tests. The translational box had dimensions of 1 and 5 angstroms for HIV protease and cathepsin systems respectively. Rotation of amprenavir scaffold was restricted to 30 degrees, where as the cathepsin scaffold was not rotationally restricted. Torsional freedom was restrained for 6 and 4 dihedrals for amprenavir and cathepsin systems respectively (see Figure 3.5 for torsional restraints). One thousand ligand conformations were tested during both of the "hunt" and "fine tune" phases of the protocol (85). Each docking run generated 20 minimized conformations and ten independent docking runs were performed for each ligand. The lowest-energy conformation of the resulting 200 conformations was chosen to be the predicted binding pose and its energy was recorded as the compound's score. Scoring of a compound with this protocol takes approximately 6 minutes on a commodity computer.

3.3 Results

The results section is organized as follows. The first subsection deals with the reproducibility of the docking energy scores. The second subsection analyzes the distribution of predicted binding affinities of smaller virtual libraries, which may help in the understanding of results in the following subsections. The third subsection evaluates the ability of Additivity methods to predict the docking energy scores. The last two subsections compare the Additivity method with the GA, based upon the retrieval of computational binders and on the docking energies of the top-scoring compounds that each method provides.

3.3.1 Convergence of VDock

The reproducibility of the VDock scores was assessed for both the HIV protease and cathepsin systems by docking and scoring 1000 compounds with randomly picked substituents, using the docking protocol described in Section 3.2.7. Figure 3.6 compares the score from one scoring run of 10x20=200 dockings with the score from a second, equivalent run started with different random number seeds. The two scores agree reasonably well for the lowest energy compounds (lower curve in each graph), while somewhat greater scatter is evident for higher-energy compounds (upper curve in each graph). Presumably the tightest binding compounds are the easiest to fit repeatedly into a low-energy conformation.

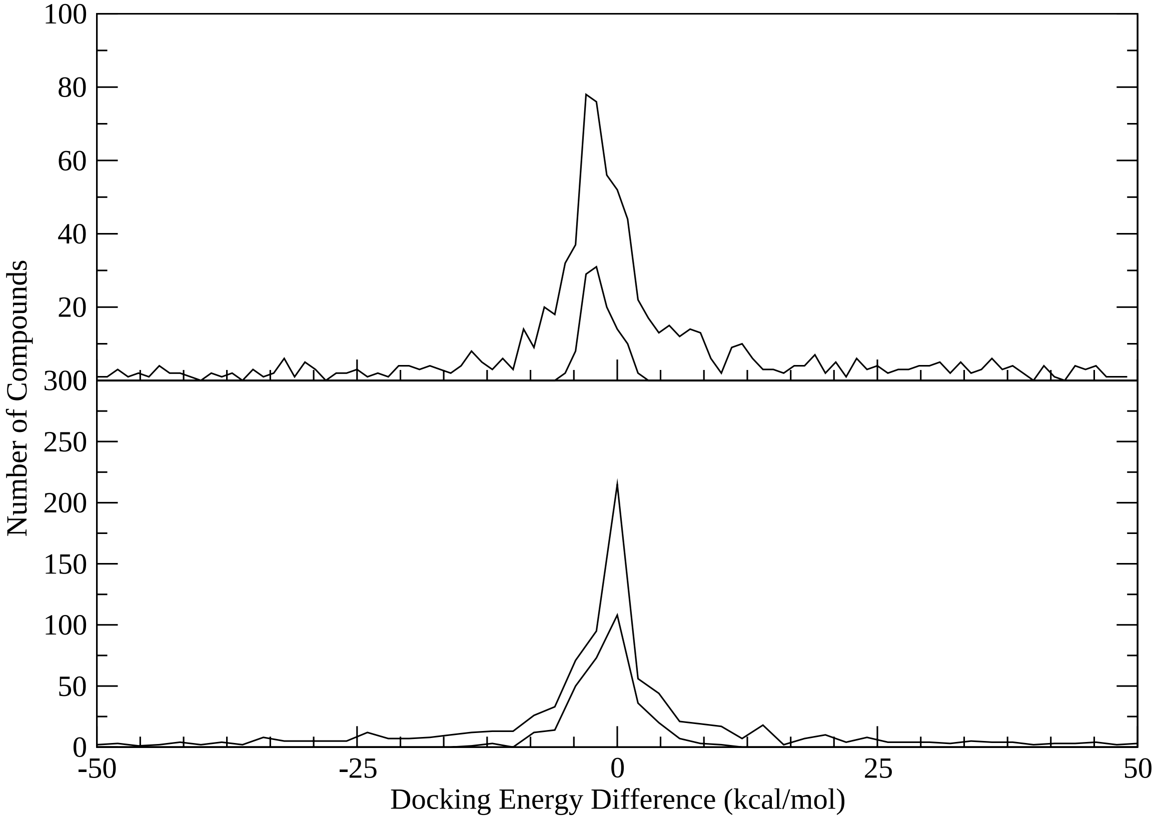


Figure 3.6: Evaluation of the consistency of the docking calculations. Histograms of the difference in docking energy score from two independent calculations on the same compound, for HIV protease with 1000 diverse compounds (top) and Cathepsin with 1000 preoptimized compounds (bottom). Upper curve in each graph is for all 1000 compounds; lower curve in each graph is for compounds with both energy evaluations less than -50 kcal/mol.

3.3.2 Characterization of compound libraries

The distribution of energy scores of the four smaller libraries is presented in Figure 3.7. The two libraries constructed of preoptimized substituents (left-most peaks in the figure) extend to scores as low as about -75 kcal/mole, with a peak at about -65 and tails to about 0 kcal/mol. The random cathepsin library extends to about -70 kcal/mol and peaks at about -48 kcal/mol, while the energies of the random amprenavir-based library are very broadly distributed, extending to about -65 kcal/mol and with a very wide peak at about -20 kcal/mol. Analysis of these distributions help in the explanation of some of the results obtained in tests of the Additivity and GA methods.

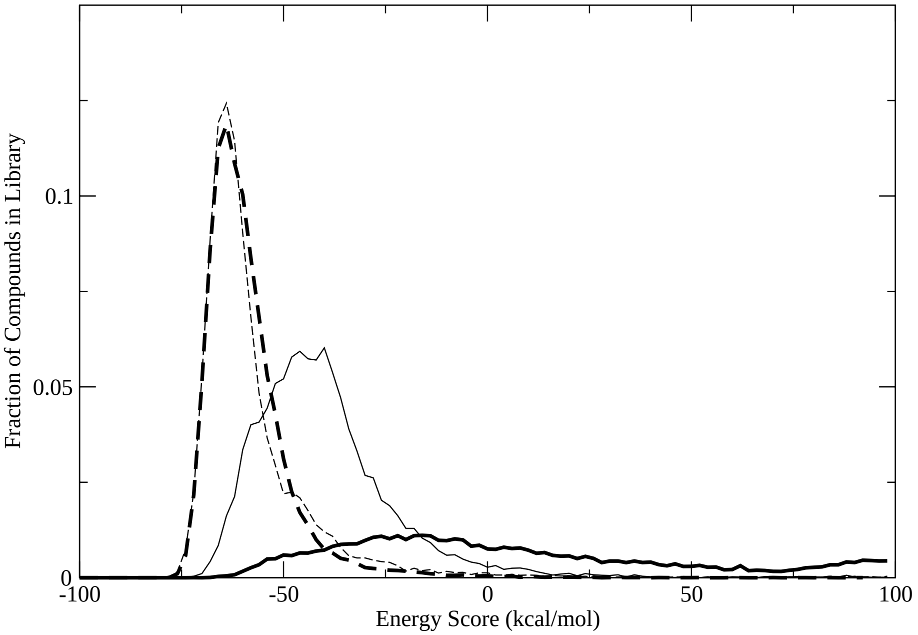


Figure 3.7: Distribution of docking energy scores for the smaller libraries. Fraction of library compounds in energy bins of width 2 kcal/mol. HIVP system: heavy lines. Cathepsin system: thin lines. Random libraries: solid lines. Preoptimized libraries: dashed lines.

3.3.3 Assessment of additivity

The numerical accuracy of the additivity approximation was assessed here by comparing the docking energies predicted with the Additivity approximation (E_{pred}) with those calculated for the same compounds by docking and scoring them (E_{calc}). The analysis was done for the smaller compound libraries (Section 3.2.5), which are amenable to exhaustive evaluation. Figures 3.8, 3.9 and 3.10 summarize the results with histograms of E_{calc} -E_{pred}. Perfect additivity would correspond to histograms sharply peaked at zero; in fact, all of the histograms peak near zero, but the width of the histograms indicates substantial deviations from perfect additivity. These deviations in the additivity could be due to the differences in the positioning of the substituents or even the scaffold itself between the predicted and calculated poses. Examples of such cases, where the Additivity could fail are shown in Figure 3.11. It is also important to recognize that some of the deviations from additivity undoubtedly result from the imperfect reproducibility of the docking scores themselves, for compounds with poor predicted binding affinity: the additivity graphs may be compared with the reproducibility graphs in Figure 3.6

Interestingly, the Additivity approximation tends to become more accurate (sharper peaks) for compounds with better energy scores. This is clear from two types of comparison within the graphs. First, the distributions are sharper for the preoptimized libraries (bottom of Figure 3.8, and all of Figure 3.10), than for the diverse libraries (top of Figure 3.8 and all of Figure 3.9); Second, the lower family of curves in each graph, which shows the distribution of errors for the lowest-energy compounds in each library, tends to be sharper and centered more precisely at zero. This observation may result, in part, from the great reproducibility of the docking energies of lower-energy compounds (Figure 3.6).

The accuracy of single-site (heavy solid lines) versus pairwise additivity (thin lines) can be assessed by comparing the curves within each family. It is evident that including pairwise interactions explicitly with the pairwise additivity method does tend to yield narrow error distributions that are centered more closely on zero. Moreover, pairing nearest-neighbor substituents (solid thin lines) tends to yield the greater improvement than other pairings (thin dashed lines).

Finally, Figures 3.9 and 3.10 provide information on the sensitivity of the results to the choice of reference compound; i.e., compound (0000) in Equations 3.1 and 3.2. Although the overall results are similar for different reference compounds, based upon comparison of the top, middle and bottom graphs, it can also be seen that different baseline compounds cause the distributions to skew differently. For example, the calculated energies tend to be more negative than predicted in the top graph of Figure 3.10, but the opposite trend is observed in the bottom graph of Figure 3.10. Potential mechanisms by which the choice of reference compound may shift these distributions are considered in the Discussion (section 3.4).

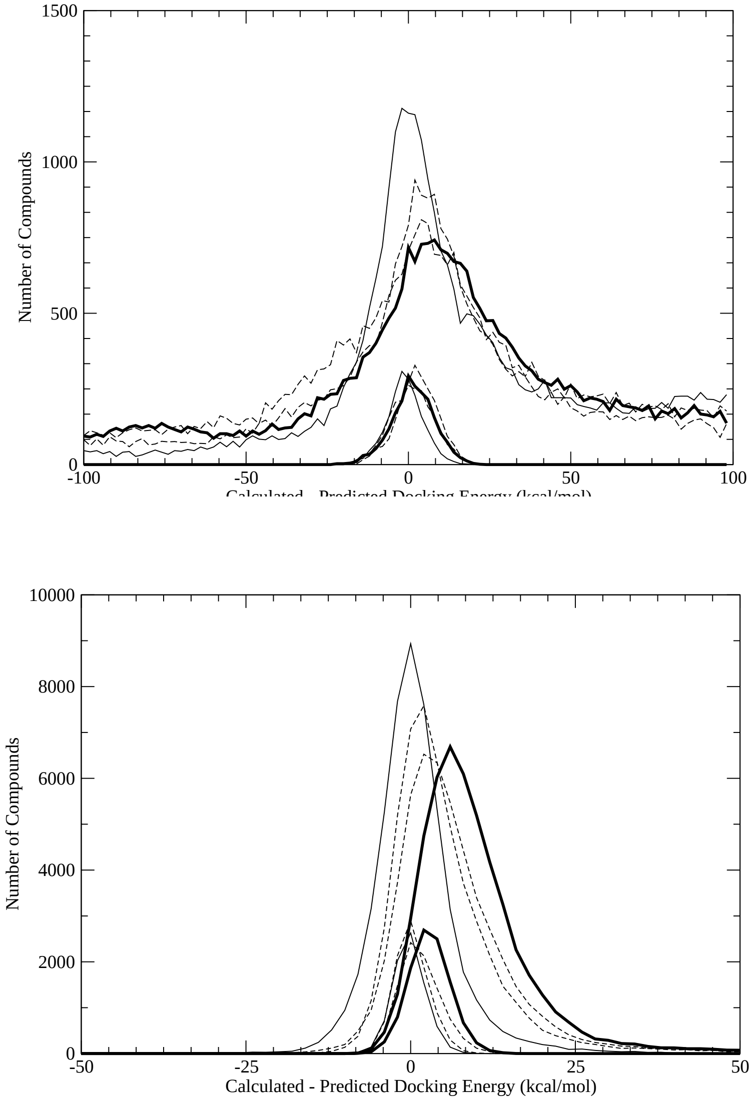


Figure 3.8: Accuracy of Additivity approximations for HIV protease system. Histograms of differences between calculated and predicted energies scores of compounds, with 2 kcal/mole bins, for single-site Additivity model (heavy solid), and Pairwise Addivitity models based upon sites R_1R_2-R_3R_4 (dashed), R_1R_4-R_2R_3 (dashed), and R_1R_3-R_2R_4 (thin solid). Diverse fragment libraries (top), showing histograms for all compounds (upper family of curves) and for compounds with predicted and calculated energies less than -50 kcal/mol. Preoptimized fragment libraries (bottom), showing histograms for all compounds (upper family of curves) and for compounds with predicated and calculated energies less than -70 kcal/mol.

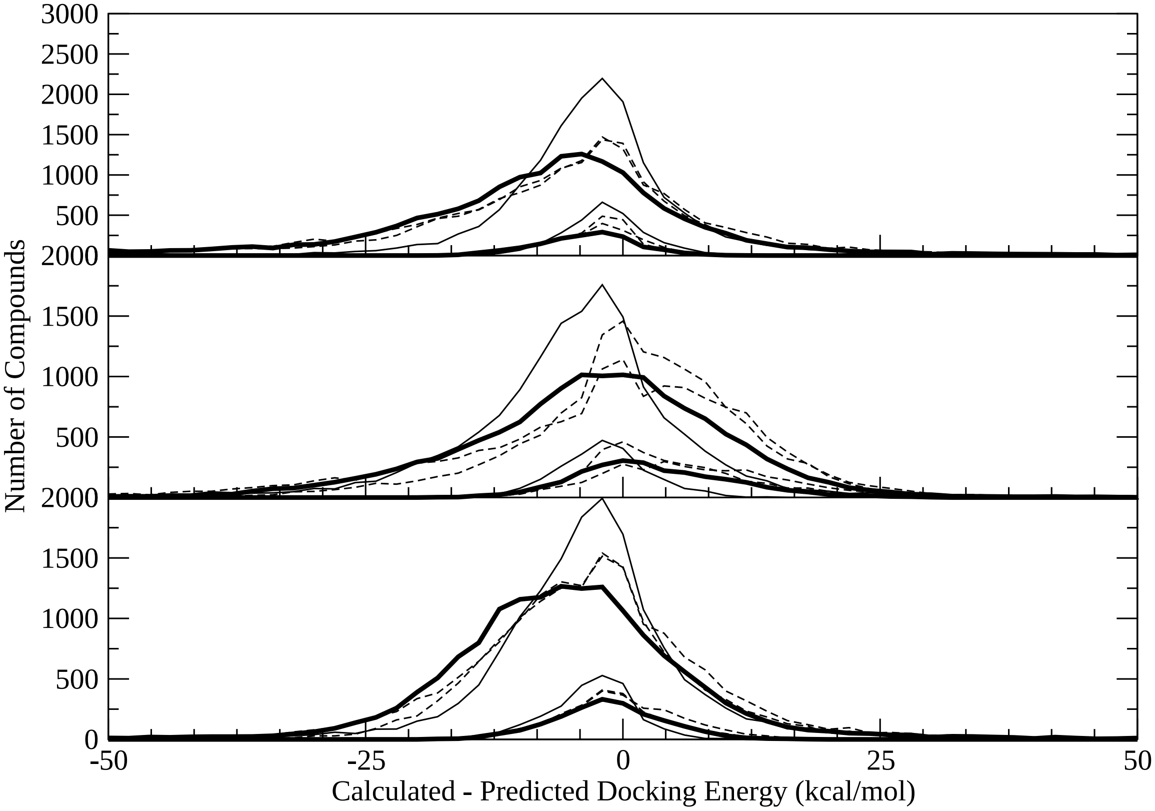


Figure 3.9: Accuracy of Additivity for Cathepsin system with diverse library. Histograms of differences between calculated and predicted energies scores of compounds, with 2 kcal/mole bins, for single-site Additivity model (heavy solid), and Pairwise Addivitity models based upon sites R_1-R_2R_3 (dashed), R_1R_3-R_2 (dashed), and R_1-R_2R_3 (thin solid), for three different reference compounds (top, middle, bottom). In each graph, the upper curves are for all compounds, and the lower curves are for compounds with docking scores less than -50 kcal/mol.

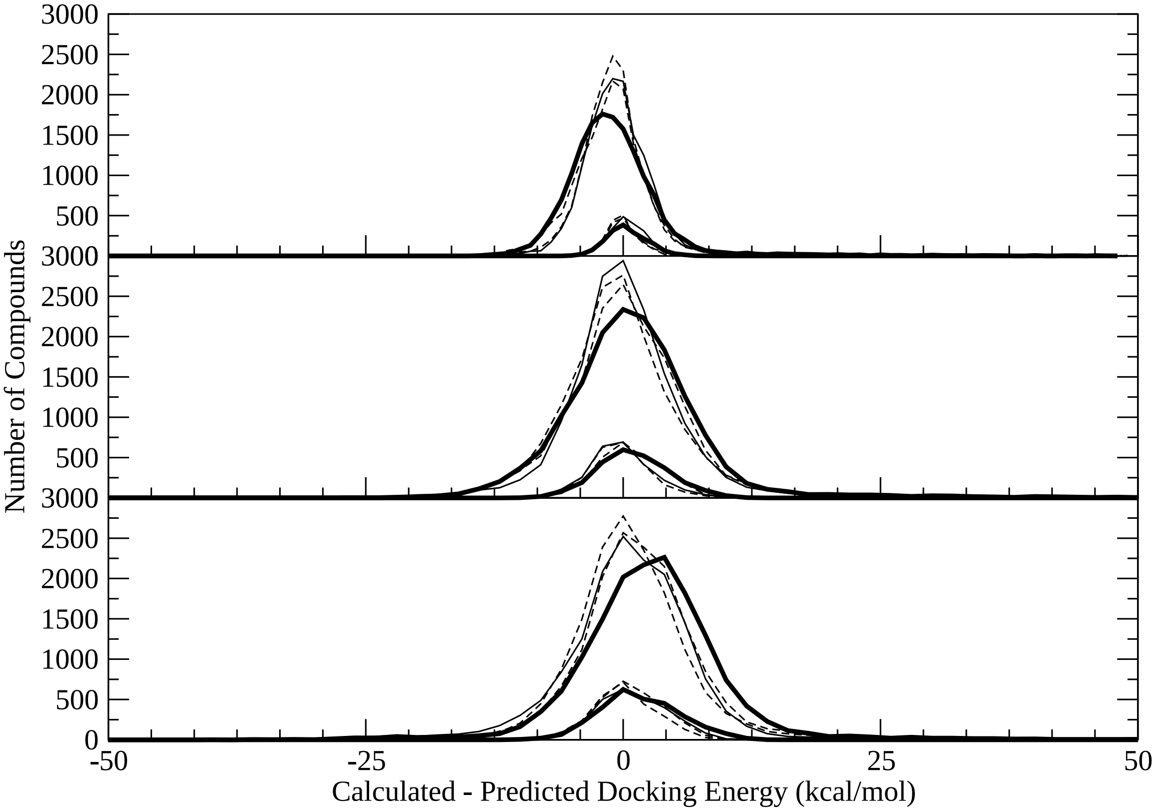


Figure 3.10: Accuracy of Additivity for Cathepsin system with preoptimized library. Histograms of differences between calculated and predicted energies scores of compounds, with 2 kcal/mole bins, for single-site Additivity model (heavy solid), and Pairwise Addivity models based upon sites R_1-R_2R_3 (dashed), R_1R_3-R_2 (dashed), and R_1-R_2R_3 (thin solid), for three different reference compounds (top, middle, bottom). In each graph, the upper curves are for all compounds, and the lower curves are for compounds with docking scores less than -65 kcal/mol.

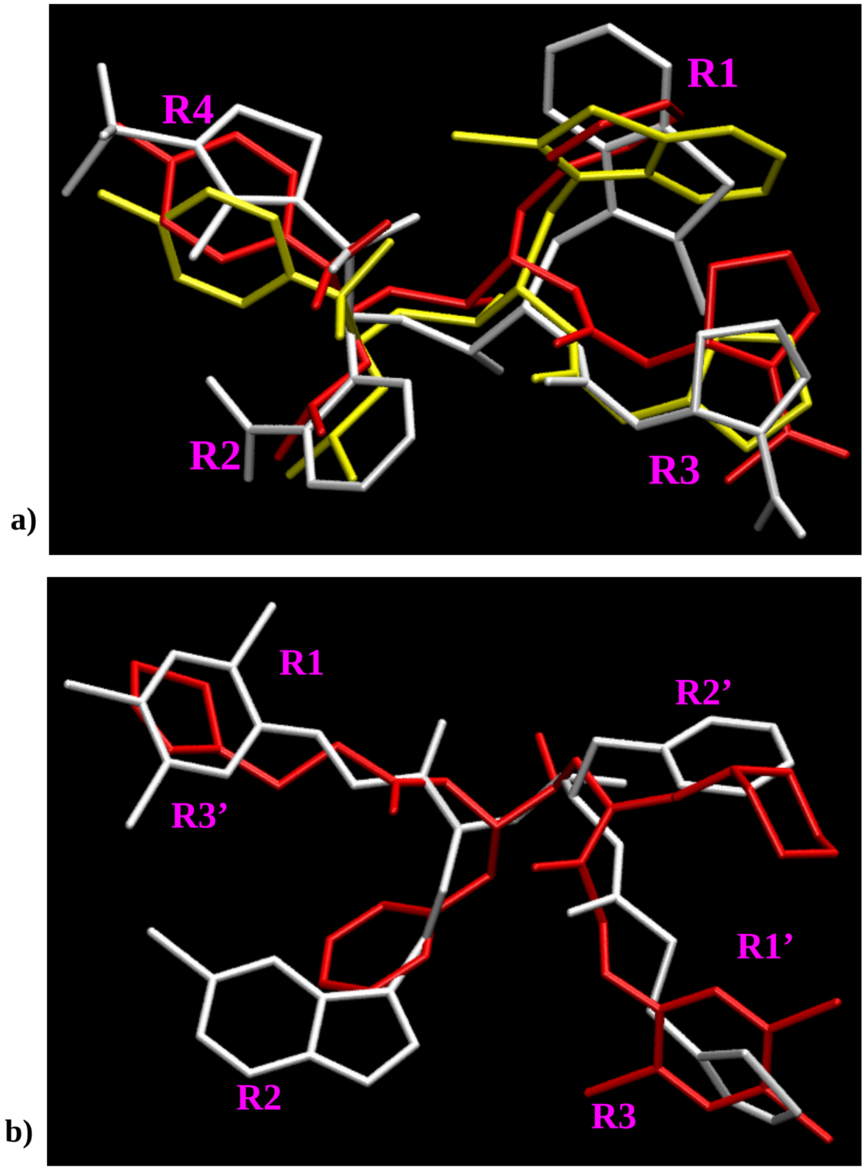


Figure 3.11: Examples of failure for the Additivity method.

(a) Differences in the orientation of substituents at R_1 and R_3 positions in the HIVP system. (b) Rotation of combinatorial scaffold itself by 180^{o} for cathepsin system.

3.3.4 Retrieval of computational binders

The Additivity and GA design methods were evaluated according to their success in recovering the computational binders in the smaller virtual libraries; i.e., the top-scoring 5% of compounds in each library. Figures 3.12 and 3.13 graph the fraction of computational binders recovered as a function of the number of docking runs. The results of hypothetical ideal and random design methods are included for comparison.

As expected, both the Additivity (red and dashed lines) and GA methods (green) consistently outperform random compound selection (cyan) and underperform ideal (black). The performance of both methods is better for the diverse libraries (Figures 3.12a and 3.13a-c), than for the preoptimized libraries (Figures 3.12b and 3.13d-f). This probably is a consequence of the fact that the diverse libraries have far fewer high-scoring compounds than the preoptimized libraries (Figure 3.6), so the computational binders are easier to distinguish from the others compounds. This also helps explain why better results are obtained for the diverse HIVP library than for the diverse Cathepsin library: the latter contains a higher concentration of tight-binding compounds.

Interestingly, the single-site Additivity method consistently outperforms the GA, although the degree to which it is favored is case-dependent. Its advantage is most marked for lowest-scoring libraries; i.e., for the random Cathepsin libraries and especially for the diverse HIVP libraries, as reflected in the energy histograms in Figure 3.6. The pairwise Additivity calculations (dashed lines) give mixed results. For the HIV protease system, they tend to outperform single-site additivity. Particularly good results are obtained with the pairing that accounts for interactions between the nearest-neighbor substituents, R_1R_3-R_2R_4 (gold dashed line). This is consistent with the enhanced predictivity of this pairing evident in Figure 3.8. It is worth noting that the pairwise calculations require more docking calculations as overhead. For example, evaluating all the component terms E_{000}, $E_{i,00}$ and $E_{0,jk}$ for the R_1-R_2R_3 pairing in the Cathepsin series, with 25 candidate substituents at each site, requires $1 + 24 + 24^2 = 601$ dockings at minimum. Single-site Additivity for the same library requires only $1 + 3(24)=73$ dockings. This difference accounts for the fact that the pairwise curves start further to the right than the single-site and GA curves. This higher overhead may also account for the fact that the yield of computational binders for a given number of dockings tends to start low. The greater accuracy of the pairwise predictions (Figures 3.8-3.10) then accounts for the fact that the pairwise curves then tend to rise above the single-site curves.

For Cathepsin, the consequences of changing from one reference compound to another can be assessed by comparing across Figures 3.13 a,b,c for the diverse library, and across d,e,f for the preoptimized library. The results are fairly uniform, except that the single-site and the two pairings other than R_1-R_2R_3 pairing (gold dashed) perform worse for the diverse library with reference compound one (Figure 3.13a) than for the other reference compounds (Figure 3.13b,c). The basis for this difference is not yet clear.

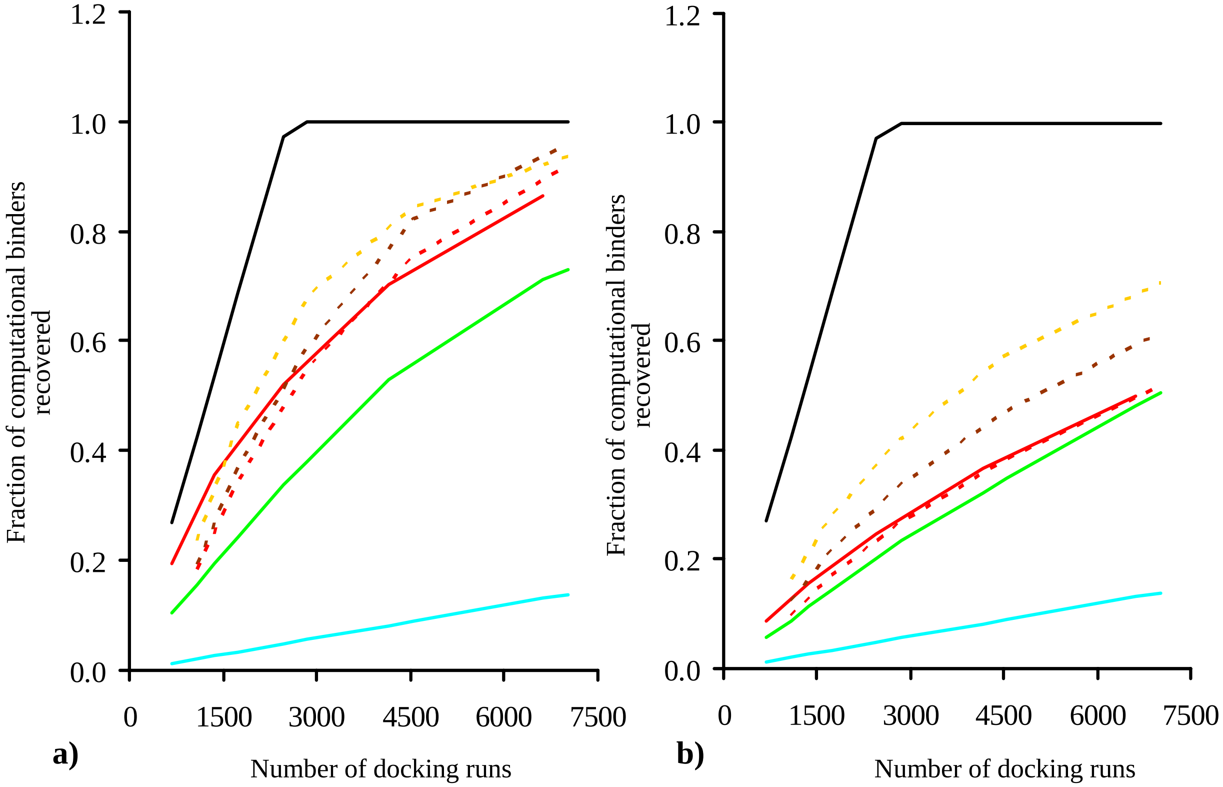
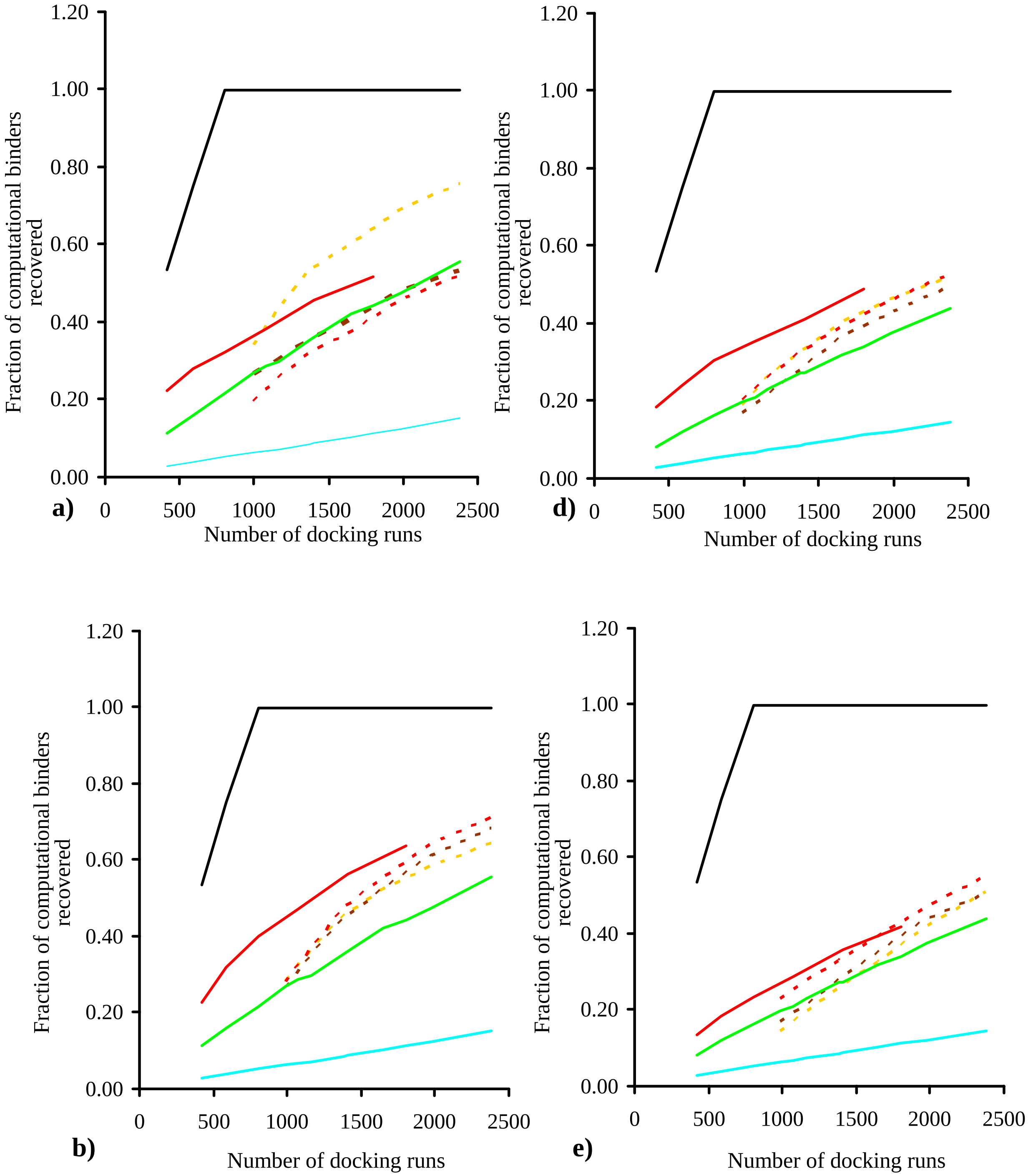


Figure 3.12: Retrieval of computational binders for HIVP system

Fraction of computational binders retrieved as a function of the number of dockings, for HIV protease system, with small diverse (a) and preoptimized (b) substituent libraries. Ideal: black. Random: cyan. GA: green. Single-site Additivity: red. Pairwise additivity, dashed, with R_1R_3-R_2R_4 gold; R_1R_4-R_2R_3 red; R_1R_2-R_3R_4 brown.


1.20
1.00
0.80
0.60
0.40
0.20
0.00
Fraction of computational binders recovered
a)
0
500
1000
1500
2000
2500
Number of docking runs
1.20
1.00
0.80
0.60
0.40
0.20
0.00
Fraction of computational binders recovered
d)
0
500
1000
1500
2000
2500
Number of docking runs
1.20
1.00
0.80
0.60
0.40
0.20
0.00
Fraction of computational binders recovered
b)
0
500
1000
1500
2000
2500
Number of docking runs
1.20
1.00
0.80
0.60
0.40
0.20
0.00
Fraction of computational binders recovered
e)
0
500
1000
1500
2000
2500
Number of docking runs

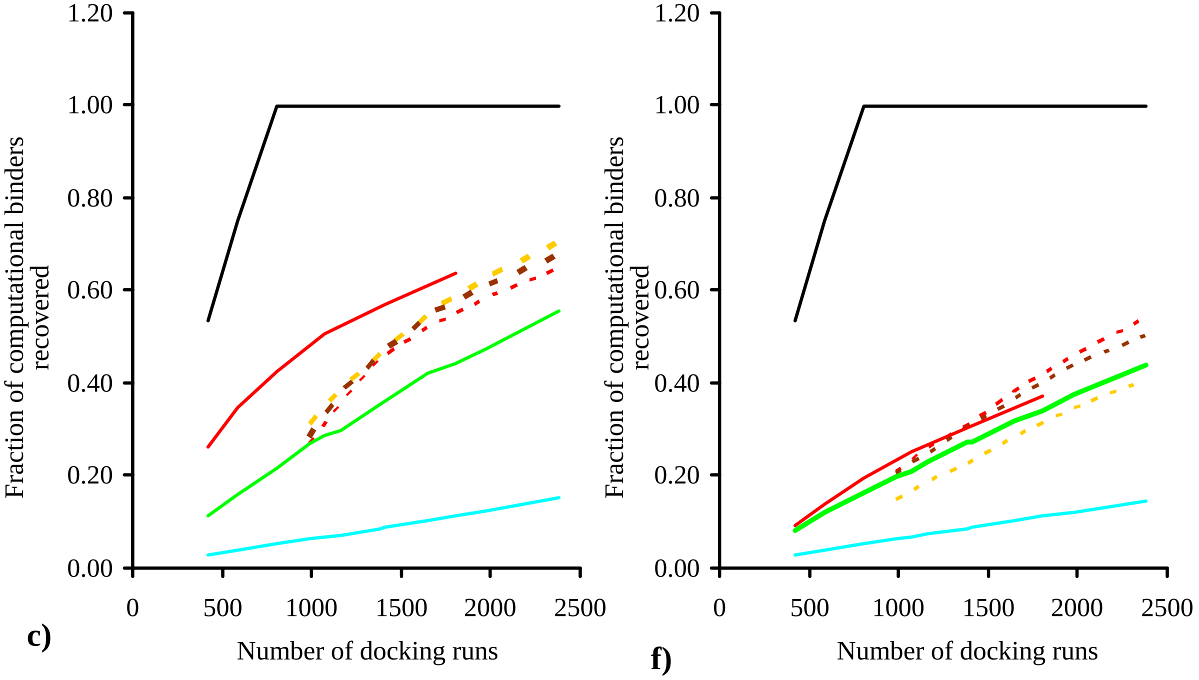


Figure 3.13: Retrieval of computational binders for Cathepsin system

Fraction of computational binders retrieved as a function of the number of dockings, for Cathepsin protease system, with small diverse (a, b, c) and preoptimized (d, e, f) substituent libraries, with different reference compounds. Ideal: black. Random: cyan. GA: green. Single-site Additivity: red. Pairwise additivity, dashed, with R_1-R_2R_3 gold; R_1R_2-R_3 red; R_1R_3-R_2 brown.

3.3.5 Docking energy scores of top compounds

The performance of the Additivity and GA methods was also evaluated based upon their ability to generate compounds with good energy scores. This measure does not require exhaustive docking of the virtual library and so can be used for the larger virtual libraries as well. Results are presented here for both the smaller and larger libraries.

3.3.5.1 Smaller combinatorial libraries

The average docking energy of the top-scoring 5% of compounds found by each method are graphed against the number of dockings for the HIVP (Figure 3.14) and Cathepsin (Figure 3.15) systems. Graphs for theoretically ideal and random compound selection are again included, for comparison. Data are provided for both the diverse (left) and preoptimized (right) small libraries. As expected, the gap between ideal and random is larger for the diverse than for the preoptimized libraries, and the best energies (ideal graphs) are more favorable in the preoptimized libraries. As expected, both the Additivity and GA methods yield much better results than random selection, yielding curves that lie fairly close to ideal, more so, apparently, than for recovery of computational binders (Section 3.3.4), presumably because the libraries contain a significant number of compounds that are not computational binders but that have energies similar to those of the computational binders (Figure 3.7).

Overall, the single-site Additivity method again tends to outperform the GA especially for the more diverse libraries, yielding sub-libraries with generally better (more negative) binding scores for a given number of dockings. The pairwise Additivity calculations again yield a mixed picture. It is worth noting that the energy ranges in the graphs are rather small, especially for the preoptimzed libraries, making fine distinctions uncertain.

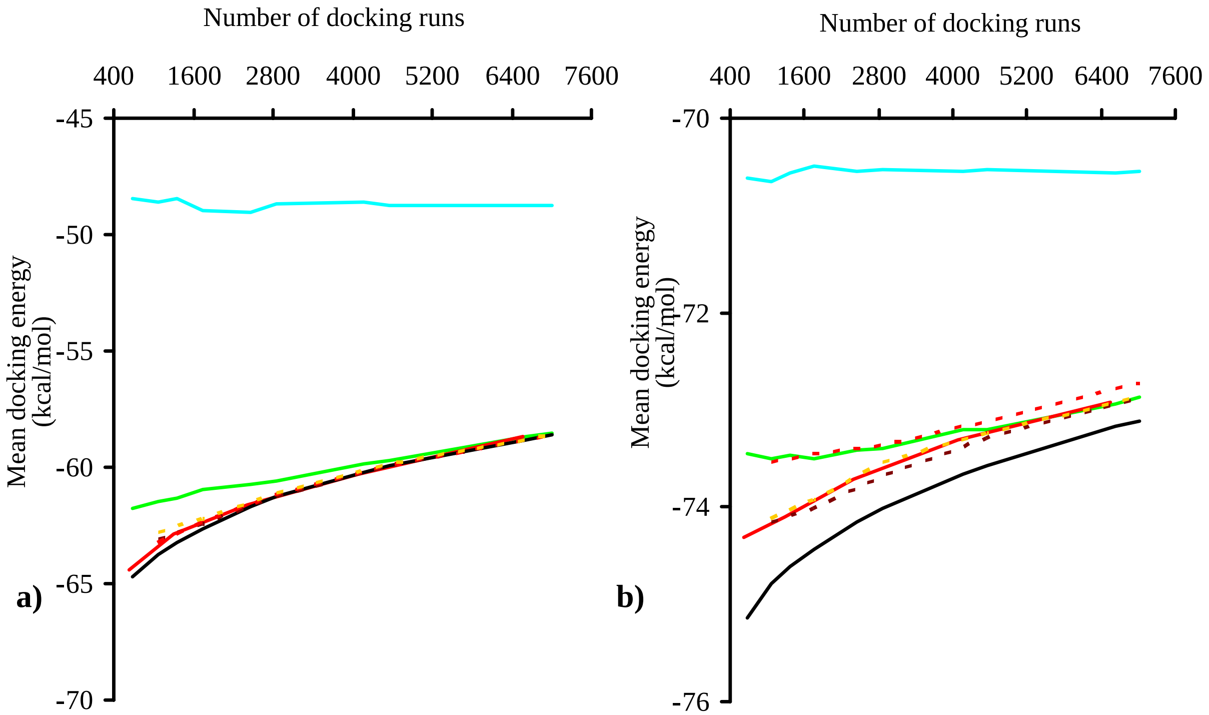
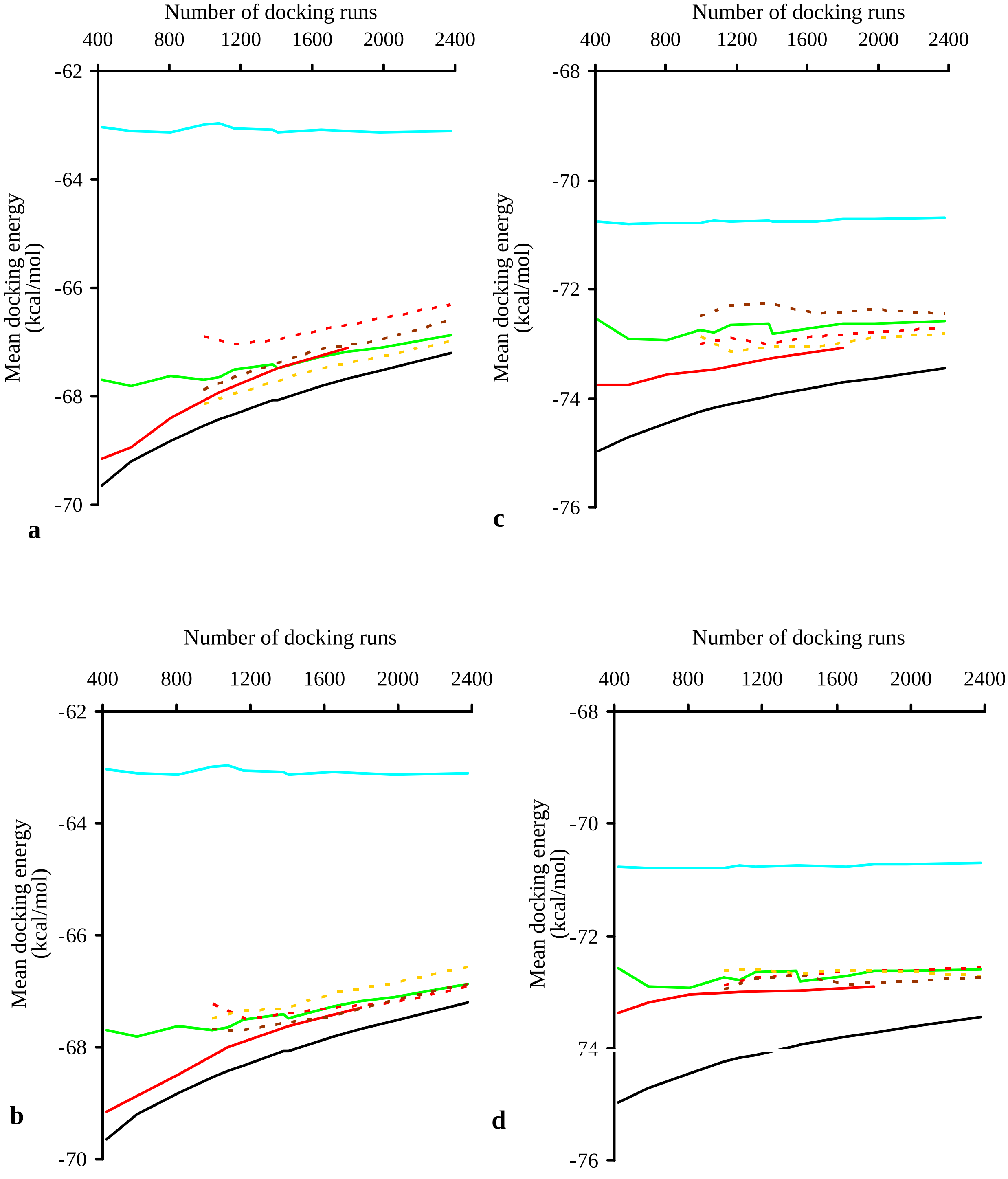


Figure 3.14: Mean docking energy of the top 5% of designed compounds, for HIVP system

(a) Diverse fragment libraries. (b)Preoptimized fragment libraries. See previous legends for symbols.



Number of docking runs
400
800
1200
1600
2000
2400
-62
-64
-66
-68
-70
Mean docking energy (kcal/mol)
a
Number of docking runs
400
800
1200
1600
2000
2400
-68
-70
-72
-74
-76
Mean docking energy (kcal/mol)
c
Number of docking runs
400
800
1200
1600
2000
2400
-62
-64
-66
-68
-70
Mean docking energy (kcal/mol)
b
Number of docking runs
400
800
1200
1600
2000
2400
-68
-70
-72
-74
-76
Mean docking energy (kcal/mol)
d

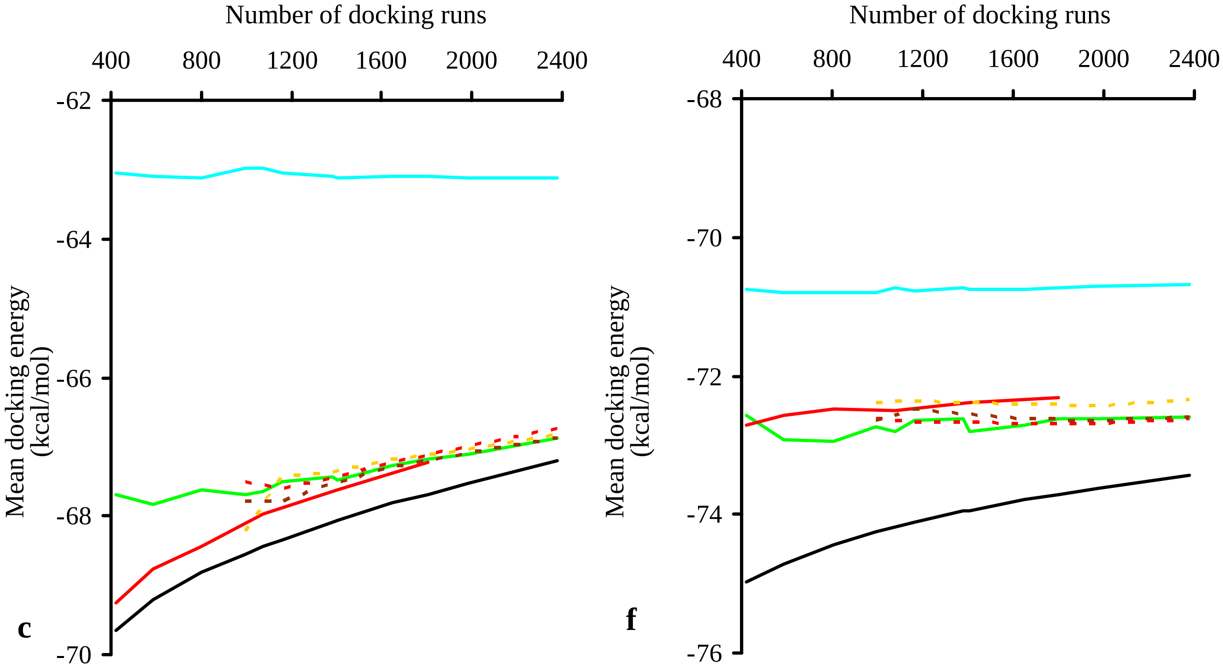


Figure 3.15: Mean docking energy of the top 5% of designed compounds, for Cathepsin system

(a) Random fragment libraries. (b)Preoptimized fragment libraries. See previous legends for symbols.

3.3.5.2 Larger combinatorial libraries

The distribution of docking energy scores of the top 1000 compounds found by each design method for the larger virtual combinatorial libraries are given in Figure 3.16. In spite of the differences in the systems, the two sets of graphs are remarkably similar. The Additivity method yields very few poor compounds, but the peaks of the GA distributions are shifted to lower energies than those of the Additivity distributions. Over the entire distribution, the Additivity method seems to yield a higher number of compounds with better docking energy scores than the GA. However, the docking energy scores of top compounds from GA are better than from the Additivity method. The differences in the patterns of the two systems could be due to a difference in the distribution of compound energies.

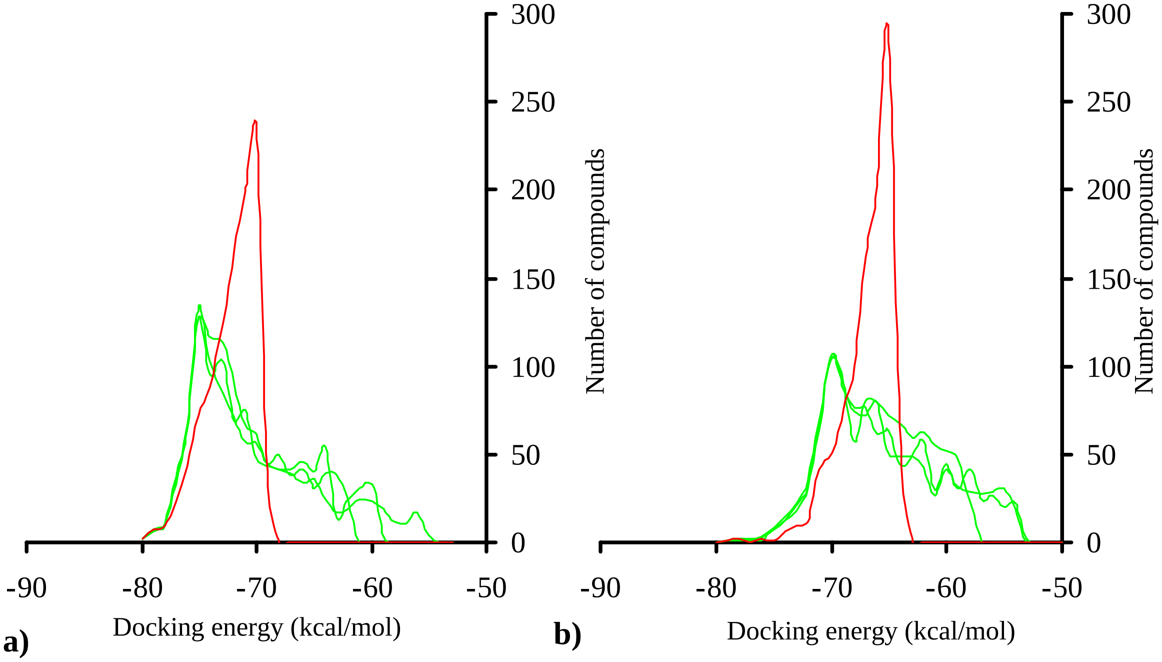


Figure 3.16: Distribution of docking energy scores of top 1000 compounds
(a) HIV protease system. (b) cathepsin system. Docking energy scores are given in kcal/mol. Bright green: GA and Red: Single site additivity method

3.4 Discussion

The present study demonstrates that both the Additivity approximation and a GA can strongly enhance the yield of high-scoring compounds in a combinatorial sublibrary, relative to random selection, when a structure-based docking method is used to score compounds. Thus, either method may be used in real-world library design. Overall, the Additivity method is found to perform as well as or better than the GA method, when both are allowed the same number of dockings to optimize a combinatorial sublibrary. One reason for this may be that, whatever its weaknesses, the Additivity method has the strength of being guaranteed to try every candidate substituent at each site; in contrast, it is highly unlikely that the GA will try every substituent at each site, when a large library is studied. Therefore, the GA risks missing good substituents. On the other hand, the GA has the potential advantage of identifying specific combinations of substituents that work particularly well because of their sizes and charges, for example, are mutually complementary. Another advantage of the Additivity method is that it is essentially trivial to implement, whereas the GA is relatively complex piece of software.

All of the methods tend to yield better results for the HIVP system than for the Cathepsin system. One reason may be that the HIVP substituents are more diverse and the compounds' energies therefore are spread more widely, making it easier to pick out the top scorers. Another reason undoubtedly is the unanticipated flipping of some compounds in the Cathepsin library. Such flips cause a given substituent to make a different energy contribution depending upon the part of the binding site with which it interacts. Note that the GA results also depend to some degree upon additivity: the crossing-over two compounds will yield a new and better compound only if the good substituents of the two compounds are still good when they are brought together in one compound. In general, both the Additivity and GA methods are expected to perform best when the combinatorial scaffold adopts a rather uniform pose within the binding site.

The pairwise approach is logical and does improve results in a number of cases, especially for the HIVP system, and especially when nearest-neighbor substituent sites are paired. Its weaker performance for the Cathepsin system may be a consequence of the flipping of some compounds, which limits the overall accuracy of additivity. The pairwise method does impose a higher computational overhead than the single-site method because it requires calculation of energy contributions from all pairs of interest, not just all single substituents. In a real-world application, knowledge that two substituents are likely to contact each other in the binding site would argue in favor of pairing these substituents. The likely benefit of the pairing could be evaluated by test-calculations of the sort described in Section 3.2.3.3.

The chief exception to the rule that the Additivity method is superior to the GA is for the large, diverse Cathepsin library, where the GA does somewhat better than single-site Additivity at identifying the highest scoring compounds. The reasons for this are still uncertain, but may have to do with the tendency of some Cathepsin ligands to flip in the binding site. On the other hand, flipping should also pose problems for the GA. It will be interesting to repeat some of these calculations with a restraint on the scaffold that will prevent flipping. Another possible explanation is that there may simply be more nonadditivity in the Cathepsin library, even in the absence of flipping, due to interactions between substituents. The pairwise Additivity approach is one way to alleviate this problem, but it would be time-consuming to run the >10,000 dockings need to run a full additivity analysis of one pair of substituents the large Cathepsin library. The time could be cut constructing and docking pairwise combinations of only the most promising substituents from a single-site analysis.

Previous applications of the Additivity concept have used single-site substitutions to pick the most promising substituents at each site, and then formed libraries by combining them. The present implementation goes beyond these methods by using Equation 3.1 and 3.2 to predict the binding energies of the various combinations. This quantitative approach should be better in the case where the energy contributions of the various substitutions are not uniformly distributed. For example, if the next-ranked two substituents at R_1 both give excellent binding energies, but neither of the next-ranked two substituents at R_2 gives good energies, and then expanding the library with the two substituents at R_1 should be better than expanding it by adding one at R_1 and one at R_2. The quantitative approach taken here also leads directly to the pairwise Additivity method, which often yields better predictions than the single-site approach.

It is worth noting that using different docking software could influence the results of this analysis. For one thing, VDock uses a pairwise additive energy function, and this clearly favors the Additivity method, as well as the GA. Incorporation of a solvation model that is not pairwise additive, such as the Poisson-Boltzmann (123, 124) or Generalized Born (124) model, might degrade the performance of these methods. On the other hand, the non-additivity of these solvent models may not be severe for a ligand in a highly desolvated binding site. A second issue is the imperfect reproducibility of the docking calculations, which limits the potential accuracy of the Additivity approximation. It would be interest to explore both of these issues through tests with different docking algorithms and scoring functions.

Ultimately, a more fundamental and interesting question is how well the Additivity approximation holds in reality; that is, how well the measured affinity of a new compound can be predicted based upon additive contributions of its substituents. Somewhat surprisingly, this issue does not appear to have been addressed in the literature; it is therefore being addressed in a separate project, to be described elsewhere.

3.5 Conclusions

The present study introduces a quantitative Additivity method for the efficient structure-based design of combinatorial libraries. The method provides for a purely single-site additivity approximation, as well as a more sophisticated pairwise approach that allows for interactions between substituents, but comes at a greater computational cost. These Additivity methods are compared with a GA design method in various situations. Both approaches are much better than random compound selection, and the Additivity method tends to outperform the GA. Not surprisingly, it is harder to identify the top-scoring compounds in a library of compounds that all score well; and when the combinatorial scaffold flips orientation in the binding site, predicted affinities become less reliable.

Chapter 4. Design, Synthesis, and Biological Evaluation of HIV-1 Protease Inhibitors with Broad Specificity

4.1. Introduction

The use of HIV protease inhibitors significantly reduced the mortality and morbidity rate in AIDS patients (25, 92, 93), but the emergence of resistance and cross resistance to the existing protease inhibitors has become a major threat in AIDS therapy. There is a thus a need for inhibitors with broad specificity against existing treatment-resistant strains, and without vulnerability to potential future mutations. One approach to developing such inhibitors may be to design compounds that make only substrate-like interactions with the binding site, so any mutation that weakens binding of the inhibitor should also weaken binding of the substrate, and thus lead to reduced enzymatic activity and a less viable virus.

Crystallographic studies of complexes of HIV-1 protease with its substrates have shown that, despite the differences in their amino acid sequences, the various substrates fill a rather uniform volume, with a toroidal shaped component on the unprimed side of the cleavage site and an extended shape on the primed side, (95, 125). This consensus substrate volume differs significantly from that occupied by existing clinical inhibitors in the binding site, and residues near locations where inhibitors extend outside the substrate volume were observed to be loci of resistance mutations. These observations led to the hypothesis that the selective recognition of substrates by treatment-resistant variants of

HIVP is attributable to the differences in the shape of bound substrate versus inhibitors., and that inhibitors that fit within the border of the consensus volume, termed the “substrate envelope”, may be less susceptible to resistance mutations (96). Chapter 2 provides retrospective data that support the validity of this substrate envelope hypothesis.

The present chapter describes a prospective evaluation of the hypothesis through the design and experimental characterization of a combinatorial library of HIV-1 protease inhibitors that aim to achieve high binding affinity while fitting the substrate envelope. The library is based upon a hydroxylethylamine scaffold (Figure 4.1), because it is a part of amprenavir, an existing HIVP inhibitor which has been shown to fit well within the substrate envelope. The ability of this scaffold to make key hydrogen bonds with HIV protease through the flap water and the catalytic aspartates, along with considerations of synthetic feasibility, also favored its selection. The present study also examines the consequences of the two accessible inversion geometries of the sulfonamide nitrogen (126) upon the computational results and predictions. Two of the designed compounds are found to bind to wild type protease with nanomolar affinity, and to retain substantial affinity against a panel of clinically relevant resistance mutations. These results support the validity of the present design strategy.

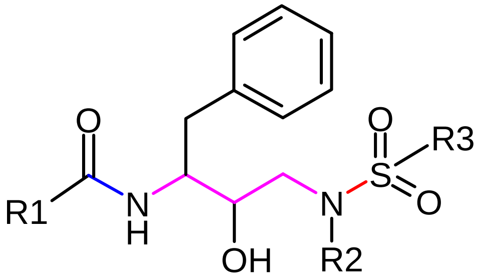


Figure 4.1: Combinatorial scaffold showing the restrained torsions. Torsional freedom along the blue, magenta and red colored bonds is restrained by ±20, ±30 and ±90 degrees respectively.

4.2 Methods

This section is organized as follows. The first subsection details the design methodology, and the next two subsections give a brief overview of screening and crystallographic studies, which were kindly provided by our collaborators. The last subsection describes the analysis of the effect of sulfonamide inversion on the docking.

4.2.1 Computational methods

4.2.1.1 Scaffold selection

The ideal characteristics of a combinatorial scaffold are as follows: 1) Synthetic feasibility and ability to provide diverse chemistry through a number of attachment points. 2) Ability to establish key interactions with the target protein. The second feature not only improves the affinity of compounds in the combinatorial library but also helps limit the movement of scaffold in the active site. This restriction can accelerate the structure-based virtual screening process by reducing the computational time involved in conformational search (Section 4.2.1.4)(78). We selected the hydroxylethylamine scaffold shown in Figure 4.1, based on the above criteria, as well as the scaffold's ability to position the three variable substituents, R_1, R_2 and R_3, within the substrate envelope. Synthetically, R_1 substituents require a carboxylic group, R_2 substituents require a primary amine, and R_3 substituents require a sulfonylhalide. The initial 3D conformation of the scaffold was prepared from the crystal structure of amprenavir in complex with HIV protease (1HPV)(127) in structure definition file (SDF) format. The sulfonamide geometry of the scaffold thus was that found in 1HPV.

4.2.1.2 Substituent libraries

Initial designs used functional groups from the "all purchasable" subset of the Zinc database (119). This subset was sorted by functional group, carboxylic acids and primary amines were extracted for the R_1 and R_2 positions, respectively. Sulfonyl halides were collected from the Sigma-Aldrich and Maybridge catalogs. All candidate substituents were restricted to have fewer than 12 non hydrogen atoms, in order to favor the construction of small compounds that are likely to fit within the substrate envelope. This restriction yielded approximately 7000, 1200 and 350 compounds for the R_1, R_2 and R_3 positions respectively. Starting 3D conformations of the substituents available in Zinc were used as-is; the rest were constructed with the program Quanta. The chemical components (carboxylic acids, amines and sulfonyl halide) were tagged as belonging to R_1, R_2 and R_3, respectively, and the functional groups were then removed and replaced with linker atoms connected by pseudobonds, to facilitate compound construction, as described below.

4.2.1.3 Construction of compounds

A new program, termed “Dovetail”, was used to build *in-silico* combinatorial libraries in 3D conformations in SDF format from initial 3D structures of the scaffold and substituents. Dovetail builds a combinatorial compound by matching the substituent sites and the corresponding functional groups for the desired compound, overlaying a pseudobond with the terminal substitution bond in the scaffold, deleting the redundant linker atoms, forming new bonds joining the substituents to the scaffold, and assigning force-field parameters to the resulting compound. “Ab initio”-like partial charges of the resulting compounds are generated with VCharge, an electronegativity equalization method parameterized to reproduce electrostatic potential fields computed at the 6-31G* level (120). Lennard-Jones parameters are assigned from CHARMm (128) atom types, and bond-torsions parameters are drawn from the Dreiding force field (129).

4.2.1.4 Docking and scoring

The program VDock was used to dock and score the combinatorial compounds, using a distance-dependent dielectric constant with a coefficient of 4. This docking program uses three pre-computed grids: an electrostatic grid, an attractive Lennard-Jones grid and a repulsive Lennard-Jones grid, which allow the rapid calculation of the interaction of the ligand with a fixed conformation of the protein (85). Combinatorial library compounds were docked to 1HPV (127). This protein structure was superimposed on 1HSG (100), to place the docked conformations in the same reference frame as that of the substrate envelope grid (section 4.2.1.5). The receptor structure was prepared by removing ligand

and other non protein atoms from the binding pockets, except for the flap water. The program Quanta (99) was then used to add all polar hydrogen atoms and their positions were optimized by energy minimization with only hydrogen atoms free to move.

In order to compute the interaction energy between a receptor and a ligand, VDock should identify the most stable bound conformation of the ligand and compute the energy for this bound conformation. Identification of the low energy conformations from a large conformational search space can be computationally challenging. We accelerated the search by restricting the movement of the combinatorial scaffold, so that it lies close to the binding pose observed in 1HPV. As the scaffold forms key interactions with the catalytic aspartates and the flap water, which beautifully anchor it in the active site, restraints on its movements in our docking and scoring calculations can be justified. The translational movement of the scaffold was restricted to ±1 Å in each axis, and the rotation was restricted to ±30 degrees. Movement along six dihedrals in the scaffold was also restrained, as shown in Figure 4.1. Ten independent docking runs were performed for each ligand, with each docking run resulting in an output of 20 docked conformations. The lowest energy conformation among the 200 generated conformations was taken as the predicted binding pose and the energy as the predicted binding affinity. Scoring of a compound with this protocol takes approximately 6 minutes.

4.2.1.5 Evaluation of substrate envelope fit

Candidate inhibitors were also scored for their fit within the substrate envelope using a grid based method described in Chapter 2. Briefly, a cubic three-dimensional grid with side-length 10 Å and grid spacing 0.2 Å was centered on the active site of HIV protease, 1HSG (100). Six HIV protease substrates (1F7A, 1KJ4, 1KJ7, 1KJF, 1KJG, 1KJH (37)) and their symmetry-operated structures, for a total of 12 substrate structures, were superimposed on 1HSG, based upon the coordinates of backbone atoms. Each grid point was assigned a value of 0 and then incremented by 1 for every substrate structure containing the grid point. A grid point is considered to be contained by a substrate if it lies within the CHARMm (128, 130) van der Waals radius of any non-hydrogen atom of the substrate.

The fit of a ligand to the substrate envelope is computed as follows. The docked ligand conformation is overlaid laid on the substrate envelope grid, and the effective volume of the ligand outside the substrate envelope, V_{out}, is computed by summing the values of the grid points g_{ijk} that lie within the van der Waals volume of the inhibitor, normalizing the sum by 12, and converting to a volume by multiplying by the 0.008 Angstrom^3 volume of a grid box:

$$V_{out} \equiv \frac{0.008}{12} \sum_{i,j,k}^{inside} (12 - g_{ijk})$$

Equation 4.1

Here “inside” implies that the sum runs only over grid points *ijk* that lie within the van der Waals volume of the inhibitor.

4.2.1.6 Selection of compounds

Our two design criteria, high predicted binding affinity and the fit to the substrate envelope were combined to give one single score, to prioritize our compounds. This was done with Z-scores, since energy and V_{out} scores have different units (131). For a given distribution, the Z-score of an individual with observed value X_i is

$$Z_i = (X_i - \bar{X}) / \sigma$$

Equation 4.2

where $\bar{X}$ and σ are the mean and standard deviation of the distribution respectively. For each candidate inhibitor, the Z-scores of its VDock energy and V_{out} were computed in the context of the distribution of energies and substrate volumes of all candidate inhibitors studied. The two Z-scores were then simply averaged to arrive at a composite figure of merit for the compound.

4.2.1.7 Combinatorial library design

An exhaustive virtual combinatorial library would comprise approximately 3 billion compounds, based upon all the candidate substituents described in section 4.2.1.2. Docking and scoring of all these compounds is not practical, as it takes approximately 6 minutes per compound. This problem was addressed as follows. First, single-site Additivity analysis was carried out, with Amprenavir as the reference compound; see Chapter 3 for methodological details and validation of this approach. The candidate substituents at each site, R1, R2 and R3, were ranked by the combined Z-scores of VDock energy and V_{out}, according to the single-site results, and the top 150 candidates at each site were identified. Multiple Genetic Algorithm calculations (Chapter 3) were carried out with these 150 candidates at each position, with the combined Z-score as a fitness function, and substituents at each position that occurred repeatedly in the top 100 compounds were identified and discussed. Unsuitable substituents, such as symmetric amines, and ones that proved to be unavailable, were eliminated. Further substituents were eliminated in order to narrow attention to fully combinatorial sublibraries. Ultimately, a fully combinatorial library of 27 compounds was proposed, and 26 of these compounds were synthesized and tested.

4.2.2 Enzymatic assays

The dissociation constant of an enzyme inhibitor, K_i, can be measured by the change in the rate of catalysis in its presence. The change in the consumption of substrate or in the generation of product over time gives the change in rate of catalysis. There are several ways to measure this change in the velocity of the reaction. One such method is the fluorometric assay in which a difference in the fluorescence of substrates or products is detected. Here, K_i values were measured with fluorescence resonance energy transfer assays. A protease substrate was terminally labeled with a florescence energy transfer donor and acceptor. On proteolysis, the fluorescence of the fluorophore is recovered, and can be monitored at suitable excitation and emission wavelengths. The proteolysis rate is thus reflected by rate of evolution of fluorescence emission, andK_i is obtained by nonlinear regression fitting to the plot of enzyme velocity as a function of inhibitor concentration.

The designed compounds were screened for their binding affinity not only to the wild-type HIV protease but also for a panel of proteases that have clinically relevant sets of mutations: M1 (D30N/L63P/N88D), M2 (L10I/G48V/I54V/L63P/V82A) and M3 (L10I/L63P/A71V/G73S/I84V/L90M)).

4.2.3 X-ray crystallography

X-ray crystallography was used to determine the structures of selected protein-ligand complexes. Crystals of protein-ligand complexes were prepared by the hanging drop vapor diffusion method. In this method, a droplet of concentrated solution of protein and precipitating agent is applied to a glass cover slip, which is then inverted so as to suspend the droplet above a larger reservoir of solution with higher concentration of precipitating agent. Over time, water in the droplet evaporates and then condenses in the reservoir, leading to a gradual increase in concentration of precipitant in the suspended droplet, and hence to crystallization of the protein-ligand complex. The protein crystals are harvested, cooled with liquid nitrogen, and used to generate X-ray diffraction patterns which are reocorded and analyzed with commercially available software to provide the 3D structure of the complex to atomic resolution.

4.2.4 Analysis of the effect of sulfonamide geometry on the predictions

The combinatorial scaffold has a sulfonamide nitrogen which is capable of inversion, resulting in two different conformations. As our docking program does not try alternate conformations of this nitrogen during the docking calculations, the sulfonamide geometry is held fixed throughout the calculations. The choice of conformation was uncertain because there were only ten PDB structures of HIVP with ligands containing such a sulfonamide moiety, three of them with 1HPV-like geometry, and 7 with the alternate geometry. As we used the receptor structure from 1HPV, we assigned the corresponding sulfonamide geometry to the combinatorial scaffold during the design process. However, the crystal structures of the designed ligands showed them to adopt the other sulfonamide geometry. This observation led to retrospective analysis of the effect of the sulfonamide geometry on the docking calculations.

The analysis was carried out by docking the designed ligands with both sulfonamide geometries into two crystal structures of HIV protease, 1HPV and KB60 (Schiffer and coworkers, unpublished) which were solved with sulfonamide-containing ligands having opposite nitrogen geometries. In the present paper, the sulfonamide geometry found in KB60 will be referred to as the "inverted" geometry. These calculations provide information regarding the direct influence of the sulfonamide geometry on ligand confrormation, and also regarding any indirect influence of the sulfonamide geometry that may result from its effect on the conformation of the protein. The docking calculations were performed as detailed in section 4.2.1.4, and the root mean square distance (RMSD) between the non-hydrogen atoms in the predicted and crystallographic structures were used to compare the docking results.

4.3 Results

This section presents results of the computational design process, provides the binding affinities of the designed compounds, then analyzes the ligand-protein interactions of the two highest-affinity inhibitors according to the computational predictions. These predicted interactions then are compared with those observed in crystal structures of HIVP with the two inhibitors. Finally, the consequences of sulfonamide geometry on the docking predictions are analyzed.

4.3.1 Computational design of combinatorial libraries

Figure 4.2 shows how the distributions of the energy and V_{out} scores of the compounds under consideration changed during the design process, and elucidates the trade-off between the energy and V_{out} figures of merit. The initial virtual library of ~3x10^9 compounds, based upon all candidate substituents, has a mean docking energy of ~-30 kcal/mol and a mean V_{out} of ~250 Å^3 (solid black curve). Both of these values are approximations based upon reconstruction of the energies and volumes via the single-site Additivity approximation (Chapter 3) for 10^8 compounds randomly picked from the full virtual library. Not surprisingly, better average docking and V_{out} scores, and narrower distributions, are observed for a smaller virtual library constructed from the 150 substituents at each position that yielded the best combined docking and V_{out} Z-scores (dashed black curve).

The red and green distributions highlight the trade-off between selecting compounds according to docking score versus V_{out}. In both graphs, the red graph represents the distributions for the 100 compounds with the most favorable predicted docking scores: as expected, their docking scores are very low (top graph); but these compounds tend to be worse than the average compound in the 150x150x150 library (bottom graph). Thus, choosing compounds based purely on docking scores would not yield compounds that fit well into the substrate envelope. Conversely, the green curves represent the distributions for the 100 compounds with the most favorable predicted values of V_{out}. These compounds fit the substrate envelope well (bottom graph), but have relatively poor docking scores (top graph).

Finally, the blue curves in both graphs show the distributions of docking scores and V_{out} computed for the 26 compounds that were actually synthesized and tested. These data were obtained by docking each compound individually, rather than by applying the Additivity approximation. It is evident that these compounds represent a compromise between optimization of docking scores alone (red graphs) and V_{out} alone (green) graphs, as the blue distributions peak between the red and green distributions. The docking scores of these compounds tend to be similar to that of Amprenavir (solid vertical line), while their values of V_{out} are a little larger than that of Amprenavir.

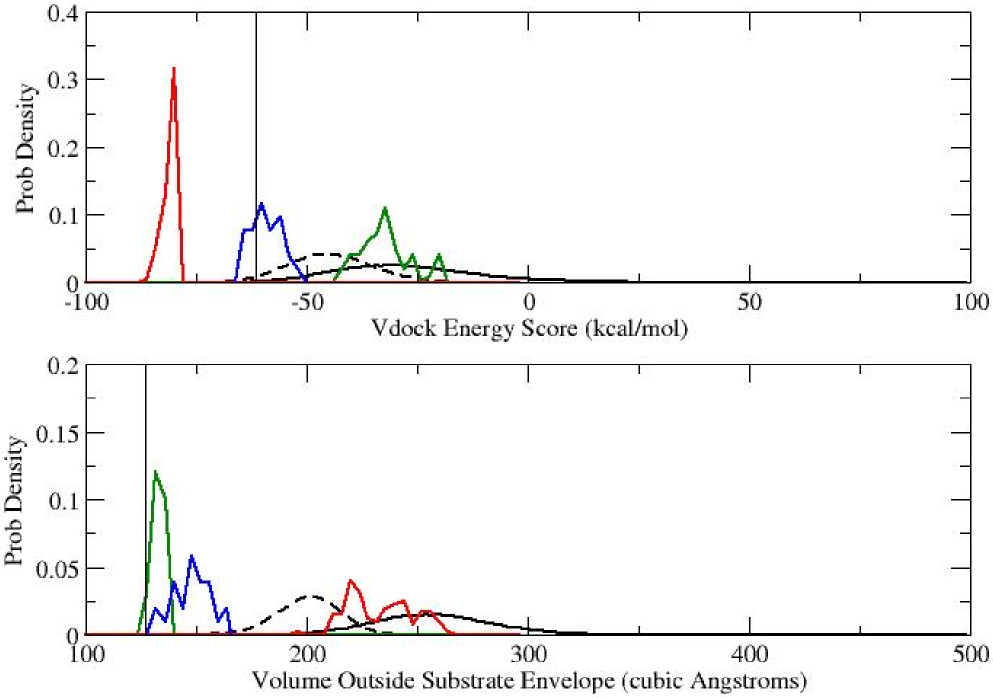
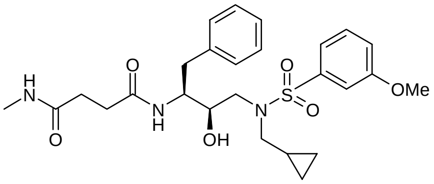
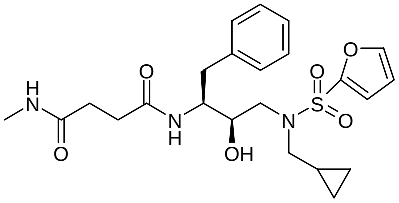
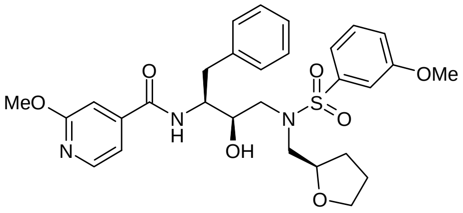
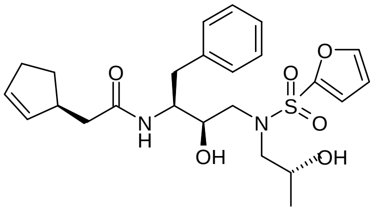
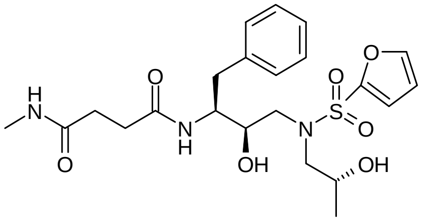
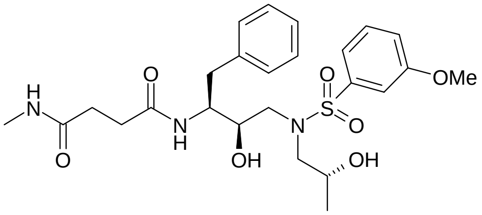


Figure 4.2: Distributions of docking energies (top) and volumes outside substrate envelope (*V_{out}*; bottom) computed for various compound sets.
Solid black: 10^8 compounds drawn randomly from the full virtual library of ~$3x10^9$ compounds; data estimated by single-site addtivity approximation. Dashed black: all compounds constructed from the 150 candidate substituents at each position that gave optimal combined Z-scores (see 4.2.1.7); data estimated by single-site additivity approximation. Red: 100 compounds with the best (lowest) docking scores estimated by single-site additivity. Green: 100 compounds with the best (lowest) values of V_{out} estimated by single-site additivity. Blue: 26 synthesized compounds; data computed by docking and scoring each of the 26 compounds. Computed results for Amprenavir are indicated by vertical black lines.

4.3.2 Binding affinity and resistance to mutation

This work resulted in seven compounds with dissociation constants for wild-type HIVP in the nanomolar range, as shown in Table 4.1. For comparison, the table also provides the structures of amprenavir, the smallest of the first-generation clinical inhibitors, and ritonavir, one of the largest, along with their measured dissociation constants. The new compounds are similar in size to amprenavir. This is consistent with the use of fit to the substrate envelope as a design criterion, because smaller compounds tend to fit the substrate envelope better. The two designed compounds of highest affinity, AD-37 and KB-45, have low-nanomolar dissociation constants, but do not bind quite as tightly as amprenavir.

Compounds AD-37 and KB-45 were further tested against the panel of three mutant proteases, M1, M2 and M3, which are highly clinically relevant (as defined in Section 2.2.3) according to the clinical relevance values listed in Table 4.2. The affinities of these two compounds for the three mutants are presented in Table 4.3, along with comparison data for a set of first-generation HIVP inhibitors in clinical use. A graphical representation of the robustness of all the compounds is presented in Figure 4.3, using the definition provided in Section 2.2.2. Each line in Figure 4.3 represents one compound; and a lower, more level line indicates greater robustness to mutation. By this measure, the two new compounds, AD-37 and KB-45, are more robust than any of the inhibitors except for amprenavir. They also fit the substrate envelope better than any of the other inhibitors except for amprenavir, according to the values of V_{out} in Table 4.3. However, it is important to note that the affinities of the new compounds are not especially good in absolute terms, as also evident from the data in Table 4.3.

INHIBITORS	CHEMICAL STRUCTURE	K_i (nM)
AD037		23.9
KB045		58
AD008		113.5
KB051		258.8
KB049		542.9
KB032		764.2

(Table 4.1 is continued on the next page.)

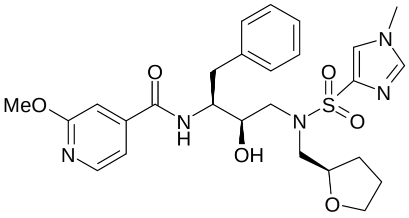
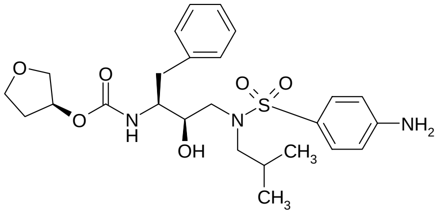
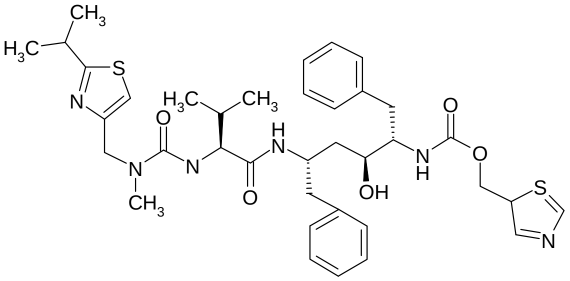
AC097		909.7
APV		0.13
RTV		0.06

Table 4.1: Structures and inhibition constants, K_i ,for selected compounds. APV: amprenavir; RTV: ritonavir. The designed compounds were synthesized by Dr Akbar Ali and Dr Kiran Reddy. The enzyme inhibition experiments were conducted by Dr Hong Cao.

MUTATION SETS	**$N_{I,ONLY}$**	**$N_{I,ALL}$**	**CLINICAL RELEVANCE**
M1 (D30N/L63P/N88D)	223	322	69.3
M2 (L10I/G48V/I54V/L63P/ V82A)	14	75	18.7
M3 (L10I/L63P/A71V/G73S/I84V/L90M)	39	102	37.5

Table 4.2: Clinical relevance analysis of mutant proteases M1, M2 and M3. See Section 2.2.3 and Table 2.4 for details.

INHIBITORS	***V_{OUT} (Å^3)***	**K_i(nM)**			
		Wild-type	M1	M2	M3
AD37	155	23.90	62.90	358.4	371.70
KB45	139	58.00	129.30	1288	2882.00
Amprenavir	128	0.13	0.21	0.15	1.40
Indinavir	180	0.18	0.73	33.58	21.15
Saquinavir	213	0.07	1.03	89.53	78.44
Nelfinavir	166	0.28	3.49	14.58	18.73
Ritonavir	256	0.06	0.46	3.03	2.81
Lopinavir	170	0.005	0.040	6.1	0.90

Table 4.3: Measured inhibition constants K_i of designed compounds and first-generation HIVP inhibitors.

V_{out}: volume outside the substrate envelope, based upon crystal structures of the respective complexes. The enzyme inhibition experiments were conducted by Dr Hong Cao.

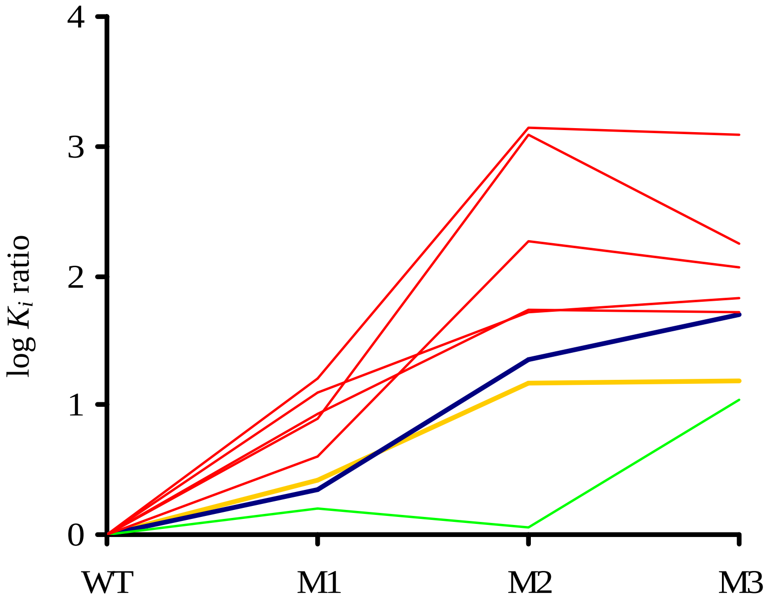


Figure 4.3: Resistance profiles graphs.
Gold: AD37. Blue: KB45. Bright green: Amprenavir. Red: Saquinavir, Indinavir, Nelfinavir, Ritonavir and Lopinavir.

4.3.3 Docked structures of inhibitor complexes

Close interactions between HIVP and inhibitors AD37 and KB45 in their predicted poses are shown in Figures 4.4 and 4.5 respectively. The two compounds differ only at their R_3 substituent position, so it is not surprising that their interactions with HIV protease through the scaffold and other substituent sites are similar. The combinatorial scaffolds of both AD37 and KB45 form hydrogen bonds with the flap water and with residue Asp25, one of the catalytic aspartates (Figure 4.4a and 4.5a). The phenyl group of the scaffold and the cyclopropyl groups of both ligands are predicted to form nonpolar interactions at S1' and S1 subsites of HIVP, respectively. The residues involved in these interactions are Pro81' and Val82' at the S1' subsite, and Leu23 and Val82 at the S1 subsite (Figures 4.4b, 4.4d, 4.5b and 4.5d).

Crystal structures show that residues in the S_2 and S_2' subsites of HIVP form both hydrophobic and hydrogen bonding interactions with existing inhibitors (31), and the docking calculations place the R_1 and R_3 substituent groups at the S_2 and S_2' subsites, respectively. Even though the amide group of the R_1 substituent is within hydrogen bonding distance of residues Asp29, Asp30 and Gly48 (Figures 4.4c and 4.5c), the angle is not optimal, so the interactions of the R_1 substituent with the residues in S_2 subsite are interpreted as mainly electrostatic in nature. Residue Asp30' in the S_2' subsite is predicted to form a hydrogen bond with the R_3 substituent of AD37. Otherwise, the R_3 substituents of both the ligands are predicted to form mainly nonpolar interactions with the protein. In particular, Ala28 is predicted to be in hydrophobic contact with the m-methoxy benzyl group of AD37 and the tetrahydrofuryl group of KB45.

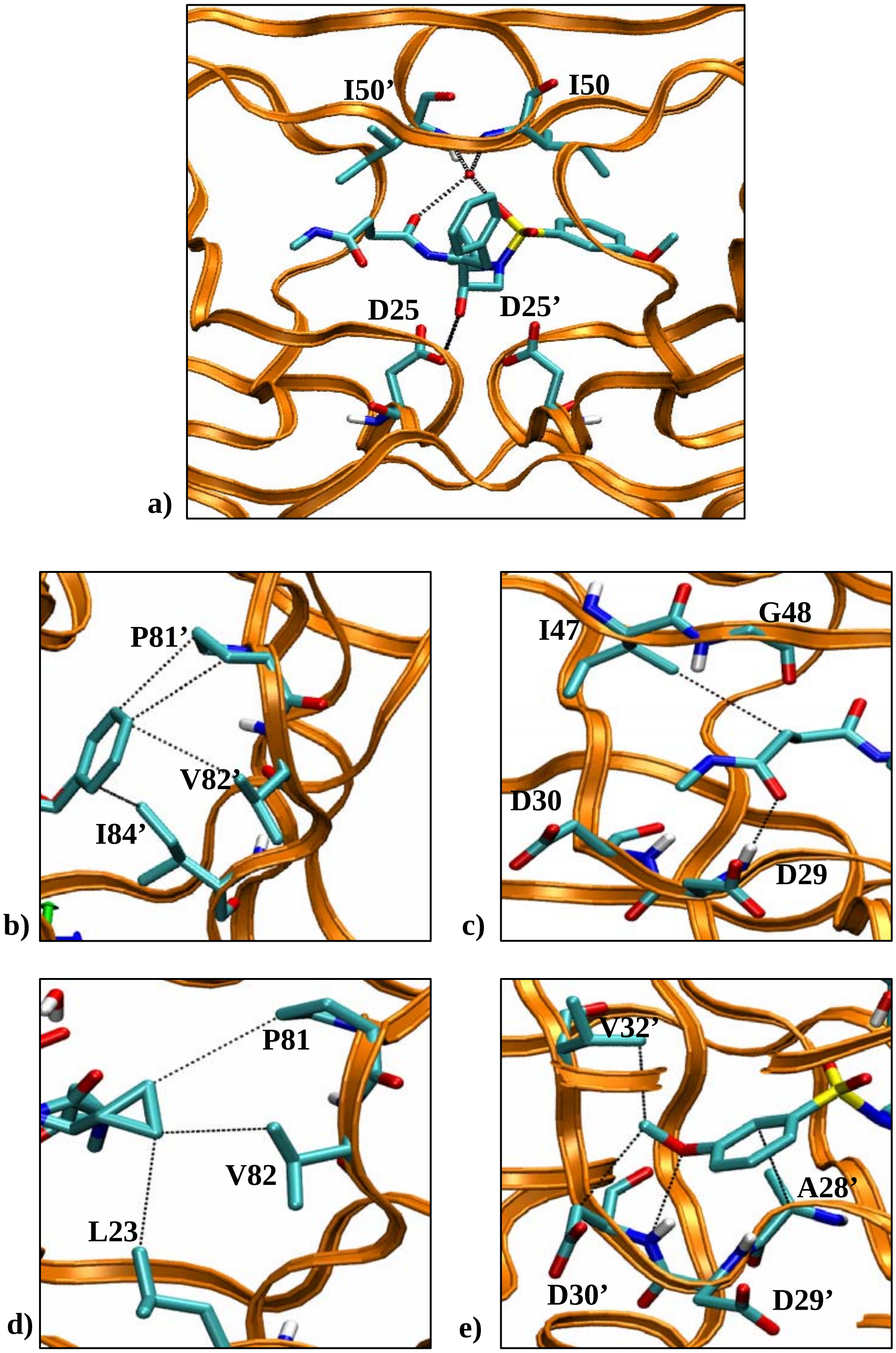


Figure 4.4: Docked structure of AD37 with HIVP.

(a) Hydrogen bonding interactions of the combinatorial scaffold. (b-e) Interactions of phenyl, R_1, R_2 and R_3 substituents with the protease. Ligand and neighboring residues are shown in licorice model. The dotted lines represent non-bonded interactions between ligand and residues within 4.1 Å

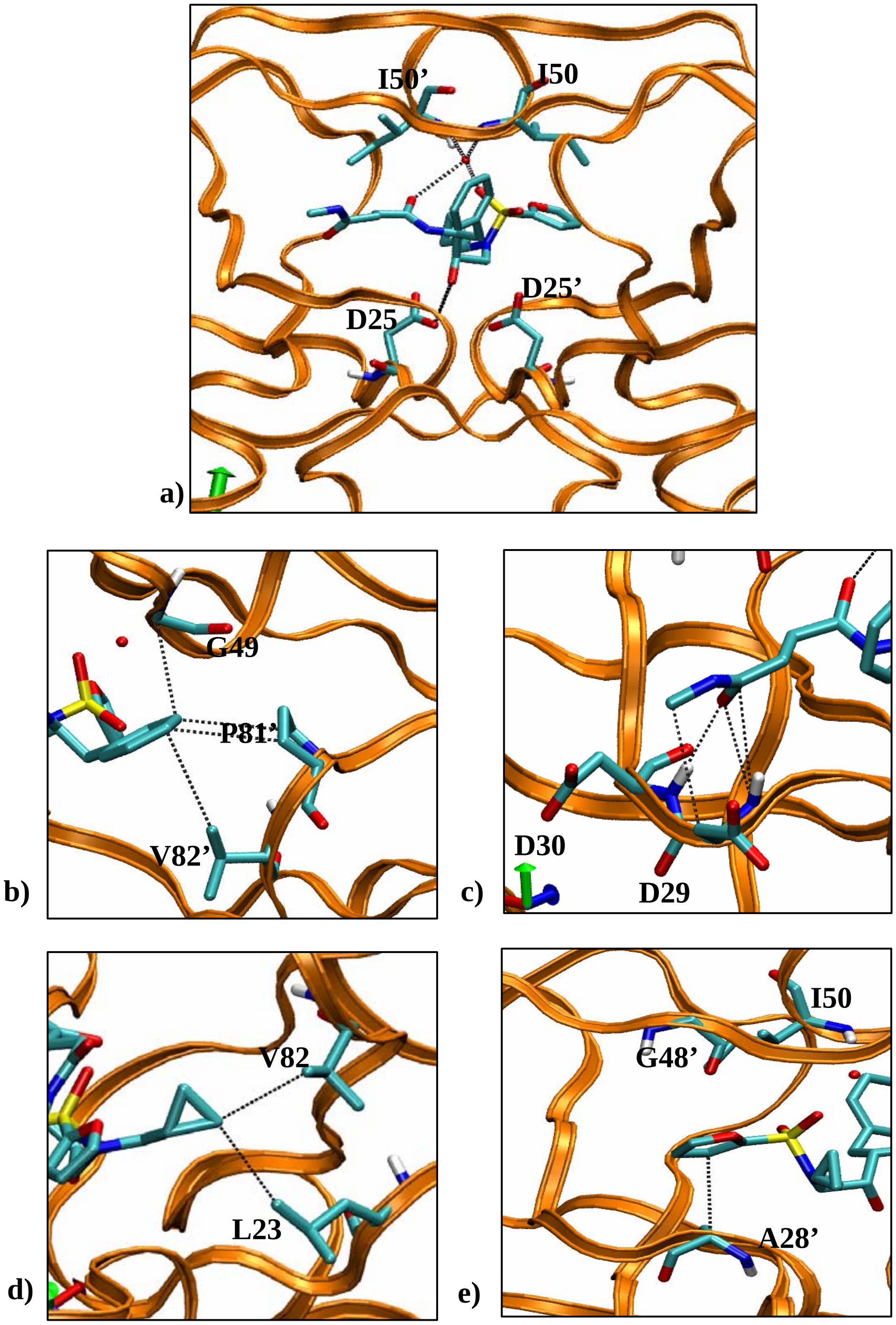


Figure 4.5: Docked structure of KB45 with HIVP.

(a) Hydrogen bonding interactions of the combinatorial scaffold. (b-e) Interactions of phenyl, R_1, R_2 and R_3 substituents with the protease.

4.3.4 Crystal structures of HIV-1 protease complexes

The crystallographic poses of AD-37 and KB-45 agree rather well with the predictions; the RMSD deviations of non-hydrogen atoms are 1.32 Å and 1.08 Å, respectively. The observed interactions of these inhibitors with HIV-1 protease also are similar to the predictions, but some differences can be discerned by comparing Figure 4.4 with Figure 4.6 and Figure 4.5 with Figure 4.7. More schematic comparisons are presented in Figures 4.8 and 4.9.

The observed hydrogen bonding interactions between the scaffold and the flap water are as predicted, but the scaffold in the crystal structures forms one more hydrogen bond with Asp25’ than predicted: compare Figure 4.4a with Figure 4.6a and Figure 4.5a with Figure 4.7a. The hydrophobic interactions of the scaffold’s phenyl group also are close to those observed in the docked poses, involving Pro81’, Val82’ and Gly49 in the S1’ subsite (Figures 4.4b vs. 4.6b and 4.5b vs. 4.7b). However, the orientation and the interactions of the cyclopropyl group at R_2 differ significantly from the predictions. In particular, the predicted nonpolar interaction with Leu23 at the S_1 subsite is absent. The R_1 substituents of both ligands form a hydrogen bond with backbone nitrogen of Asp29 at the S_2 subsite, although this interaction was not observed in the most stable computed poses (Figures 4.4c vs. 4.6c and 4.5c vs. 4.7c).

The methoxy methyl moiety of the R_3 substituent of AD-37 was predicted to form a nonpolar interaction with Val32 in the S_2’ subsite, but the crystal structure instead shows multiple interactions between the benzene moiety of R_3 and Val32. Similarly, more nonpolar interactions were observed than predicted between the R3 substituent of KB-45 and nearby nonpolar groups (Figure 4.7e versus Figure 4.5e).

Finally, the sulfonamide nitrogen in the crystal structures is inverted relative to that used in the docking calculations. It was conjectured that this difference might account for differences between the predicted and observed positions of the various substituents. The following subsection analyzes the consequences of the sulfonamide geometry for the docked conformations.

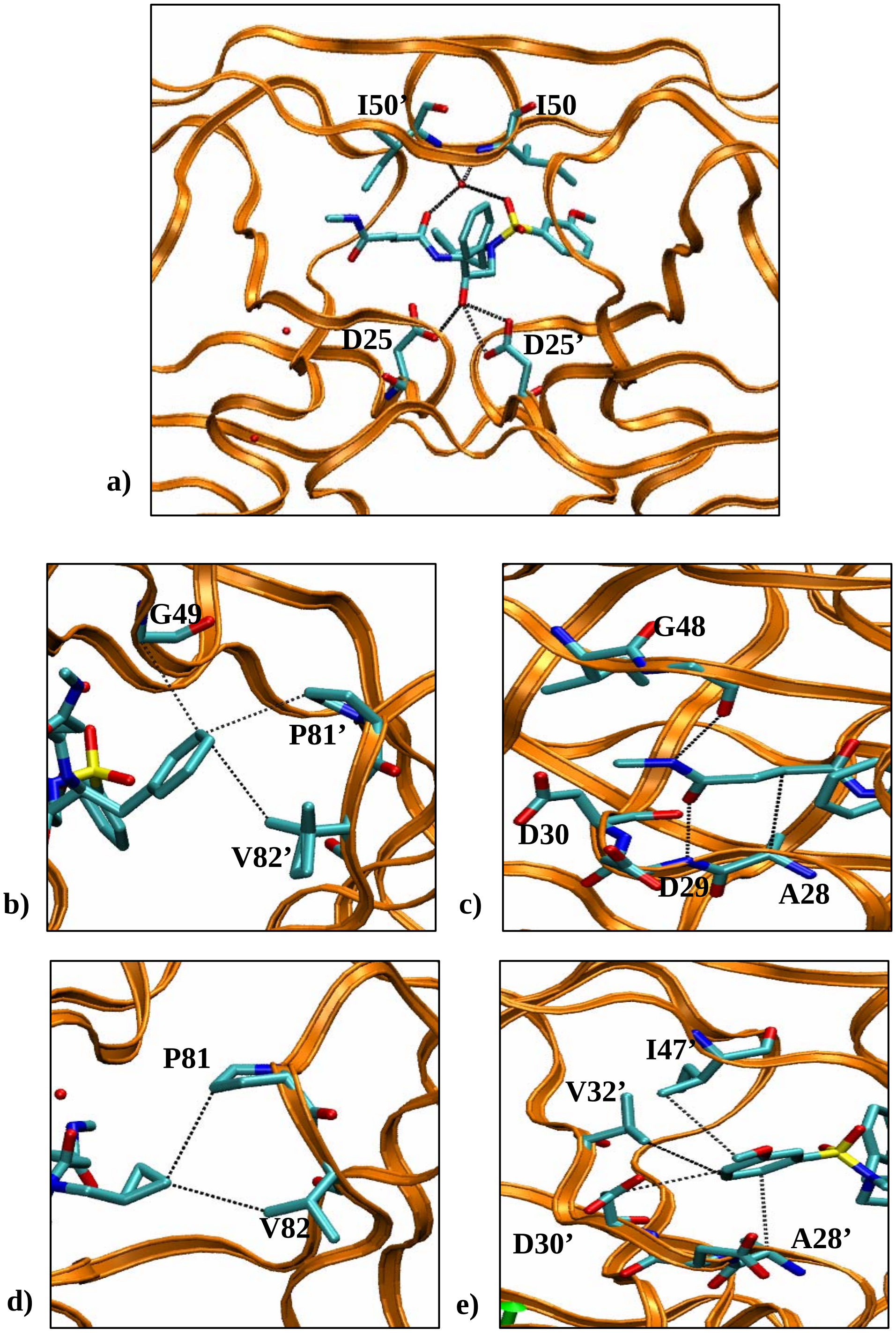


Figure 4.6: Crystal structure of AD37 in complex with HIV-1 protease.
(a) Hydrogen bonding interactions of the combinatorial scaffold. (b-e) Interactions of phenyl, R1, R2 and R3 substituents with the protease. The crystallographic analysis was performed by Dr Madhavi Nalam.

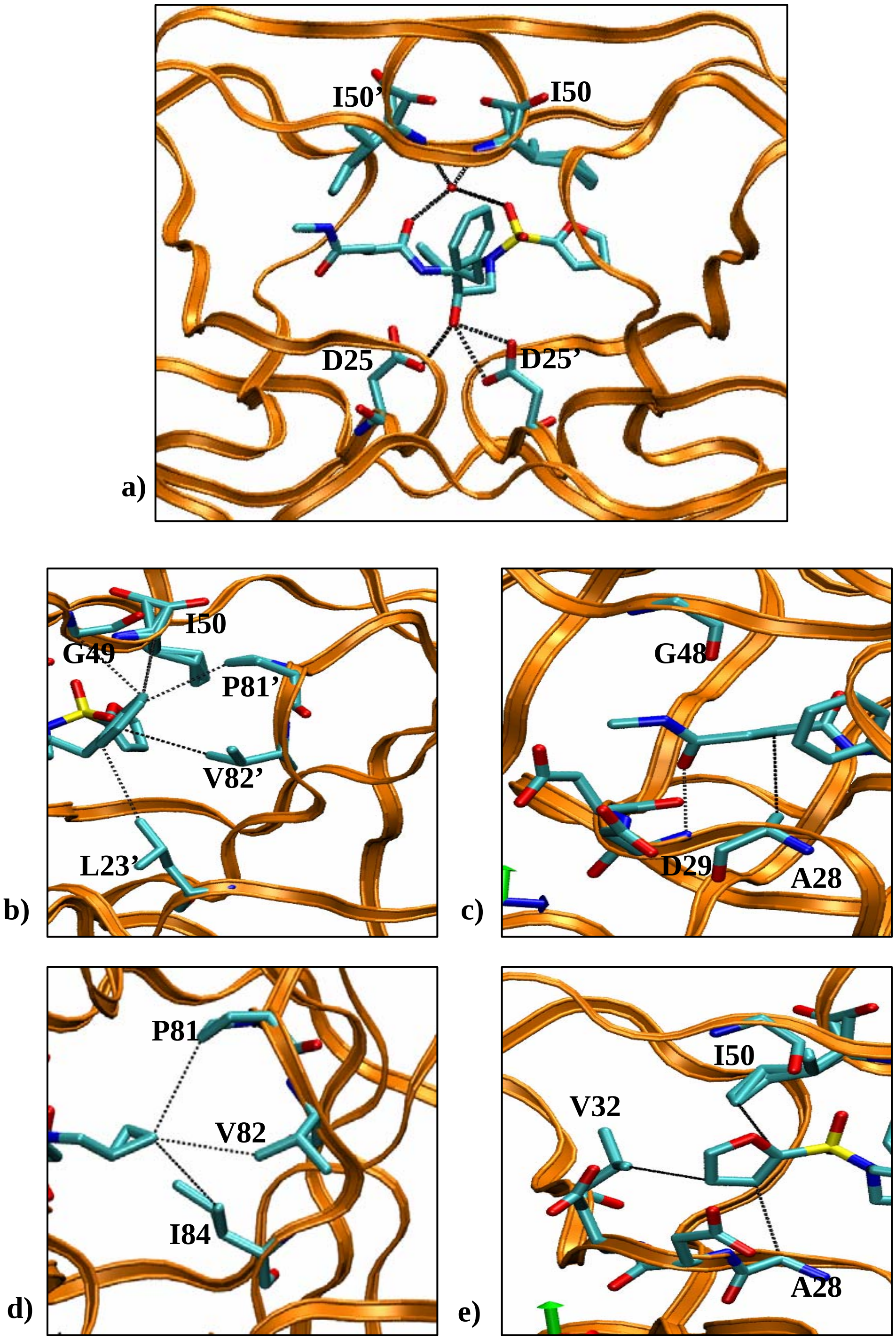


Figure 4.7: Crystal structure of KB45 in complex with HIV-1 protease.
(a) Hydrogen bonding interactions of the combinatorial scaffold. (b-e) Interactions of phenyl, R1, R2 and R3 substituents with the protease. The crystallographic analysis was performed by Dr Madhavi Nalam.

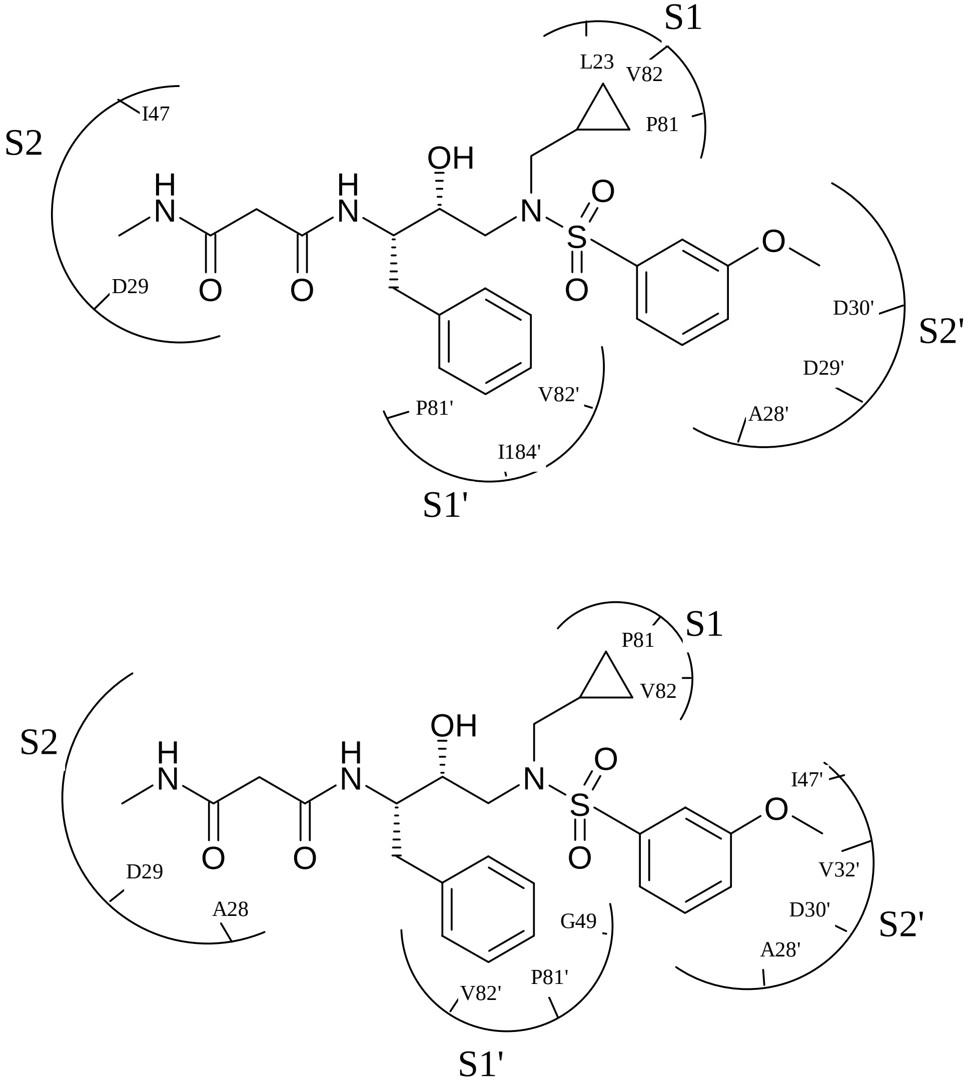


Figure 4.8: Interactions of AD-37 with the active site of HIV protease
As predicted (a) and crystallographic (b) Residues within 4.1Å of the inhibitor are shown.

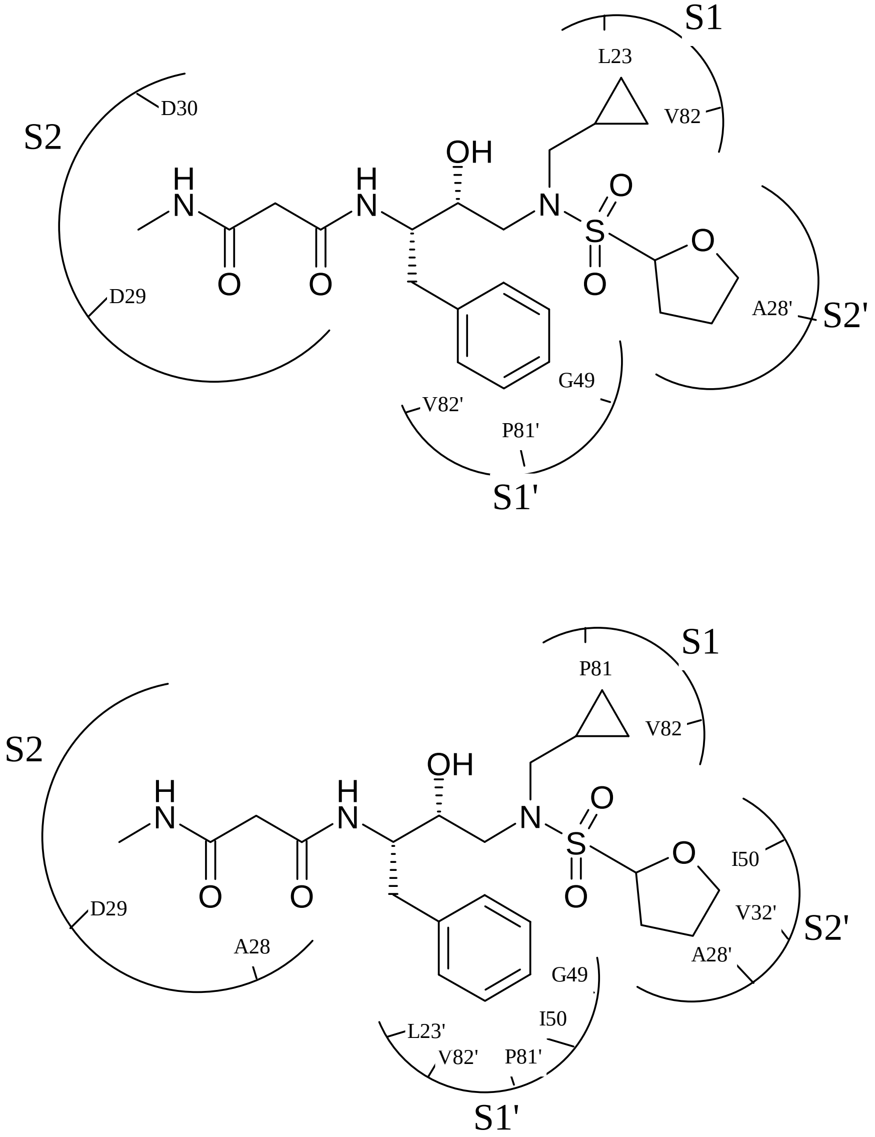


Figure 4.9: Interactions of KB-45 with the active site of HIV protease
As predicted (a) and crystallographic (b) Residues within 4.1Å of the inhibitor are shown.

4.3.5 Consequences of sulfonamide geometry

The geometry of the sulfonamide nitrogen could affect the VDock predictions directly by affecting the positioning of the substituents, and indirectly by affecting the protein conformation. The direct effect of the sulfonamide geometry was analyzed by additional docking calculations using AD-37 and KB-45 with both possible sulfonamide geometries. Possible indirect effects were examined by docking the ligands into two receptor structures, crystallized with ligands having different sulfonamide geometries, 1HPV and KB60. The four resulting docked conformations were superimposed and compared. Tables 4.4 and 4.5 provide the RMSD of non-hydrogen atoms in the four docked conformations, relative to the crystal structure. In all cases, docking with the inverted sulfonamide geometry of the crystal structures yielded lower RMSDs. However, the choice of target protein structure has an inconsistent influence on the RMSD values: docking into KB60 gives lower RMSDs for AD-37, but higher RMSDs for KB-45.

As shown in Figure 4.10, the conformation of the scaffold is similar across all four docked conformations of AD-37, so this aspect of the docked conformation is insensitive to sulfonamide geometry. The scaffold of KB-45 is somewhat more sensitive to the sulfonamide geometry, as shown in Figure 4.11.

The positioning of the cyclopropyl substituent at R_2 overlays well on the crystal conformation only when the inverted sulfonamide geometry is used (two blue conformations in red regions of Figures 4.10 and 4.11); a significant discrepancy is observed when the 1HPV sulfonamide geometry is used (red and orange in Figures 4.10

and 4.11). It is not surprising that the cyclopropyl group should be particularly sensitive to the sulfonamide geometry because it is linked directly to the nitrogen in question. The R_3 substituent also is bonded to the sulfonamide group and, like the cyclopropyl group, is best positioned when the inverted sulfonamide geometry of the crystal conformations is used during docking (two blue conformations in the red shaded regions of Figures 4.10 and 4.11). The error in the predicted position of R_3 of KB45 may result not only from the incorrect sulfonamide geometry, but also from its lack of strong interactions at the S_2' subsite, relative to AD-37 (Figures 4.10: red box vs. 4.11: red box), and hence greater mobility.

Finally, the conformation of the R_1 substituents of both AD37 and KB45 vary significantly among the docked poses, but no clear correlation is observed between the sulfonamide geometry options and the agreement with the crystal structures (Figures 4.10 and 4.11: blue shaded region).

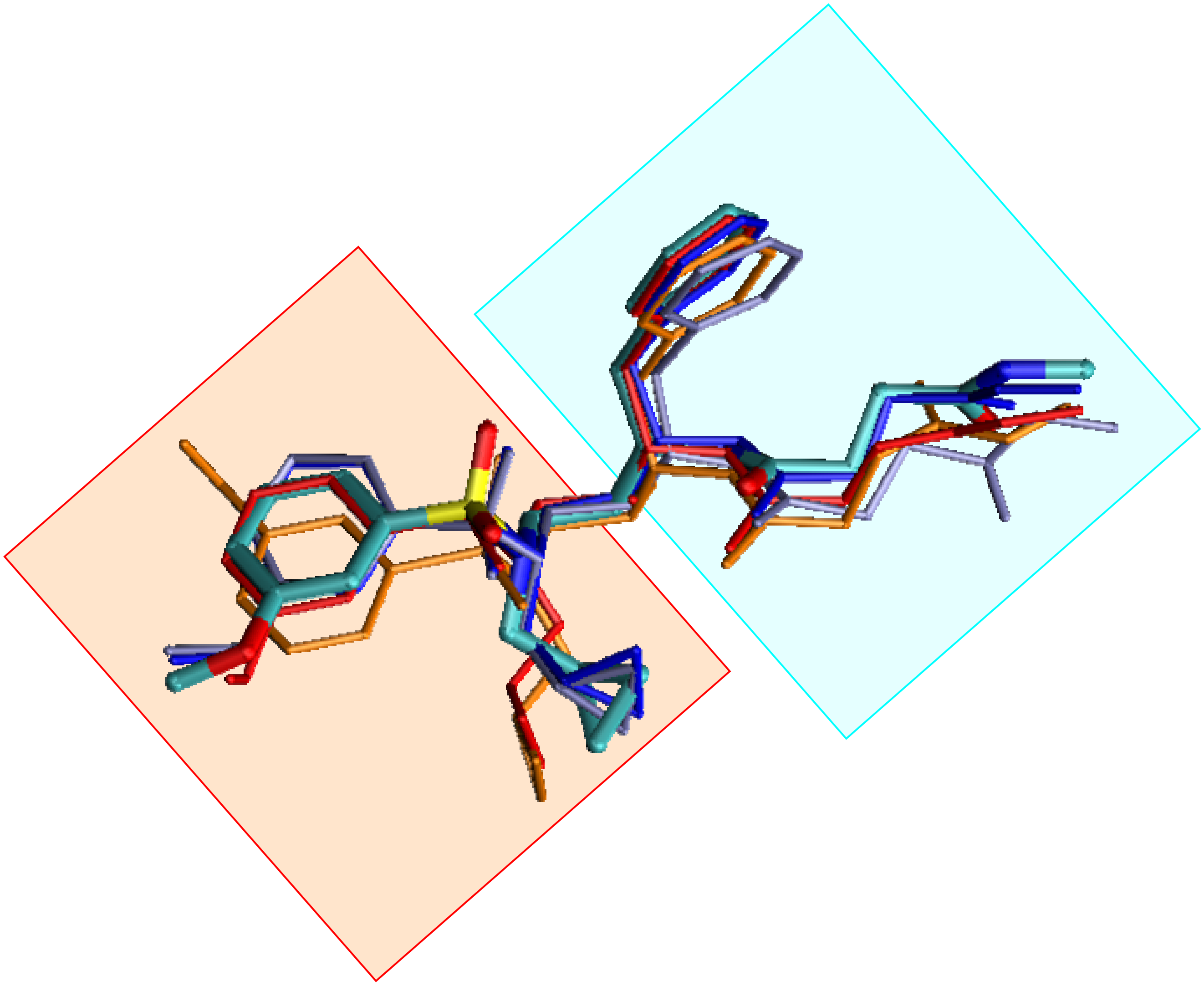


Figure 4.10: Superimposed docked structures of AD37.

The R1 and the phenyl group of the scaffold are grouped together by a blue box and the R2 and R3 substituents by a red box, to highlight the variability within the docked poses. Predicted binding poses of AD37 with inverted sulfonamide geometry when docked into KB60 (blue) and 1HPV (ice blue); and AD37 with 1HPV like sulfonamide geometry when docked into KB60 (red) and 1HPV (orange) are shown.

PROTEIN STRUCTURE	**SULFONAMIDE GEOMETRY**	**RMSD(Å)**
1HPV	1HPV-like	1.32
1HPV	Inverted	0.90
KB60	1HPV-like	0.73
KB60	Inverted	0.65

Table 4.4: RMSD between the corresponding non-hydrogen atoms in the docked and the crystal structure of AD-37

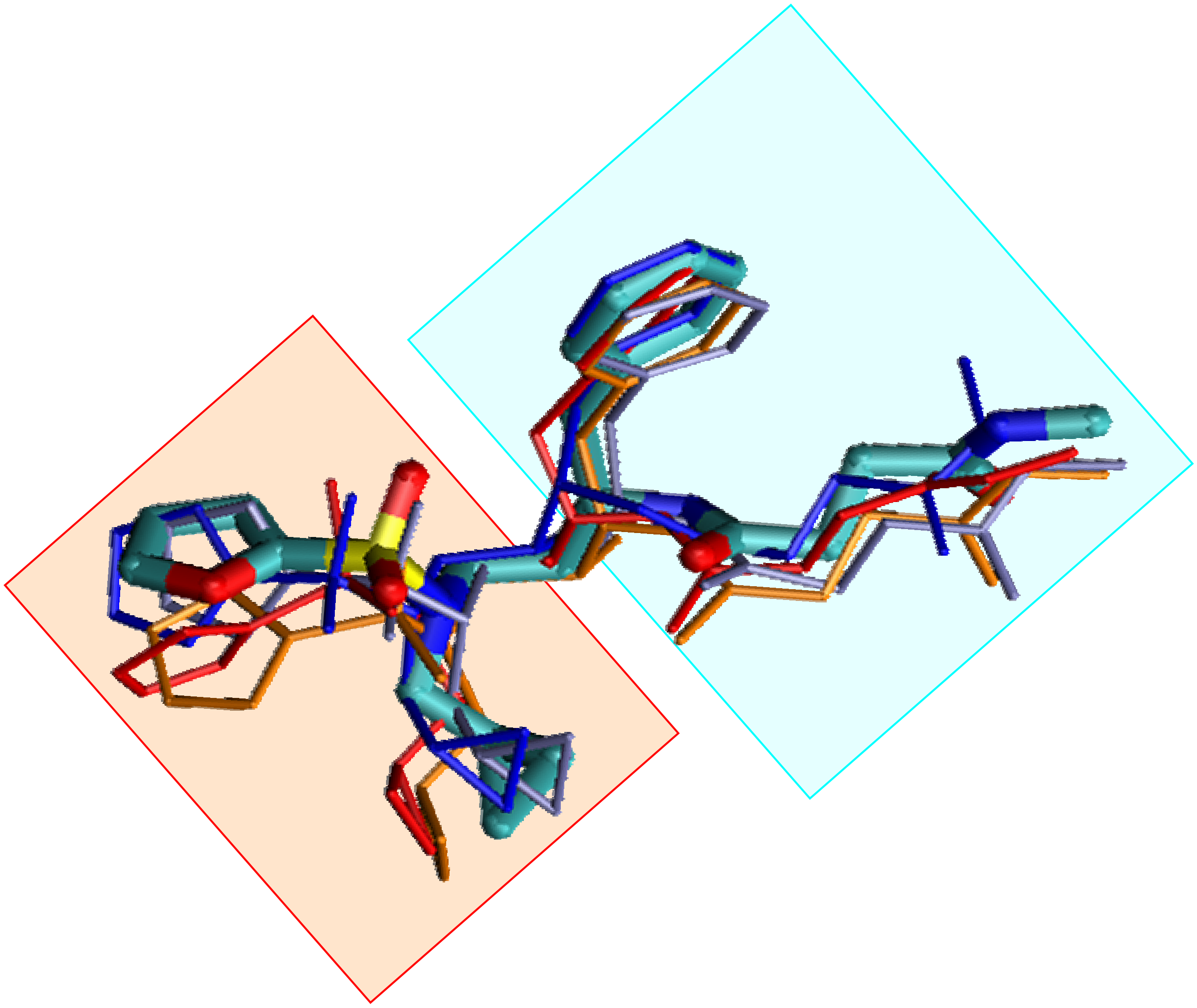


Figure 4.11: Superimposed docked structures of KB45.
The R1 and the phenyl group of the scaffold are grouped together by a blue box and the R2 and R3 substituents by a red box, to highlight the variability within the docked poses .at different substituent sites Predicted binding poses of KB45 with inverted sulfonamide geometry when docked into KB60 (blue) and 1HPV (ice blue); and KB45 with 1HPV like sulfonamide geometry when docked into KB60 (red) and 1HPV (orange) are shown.

PROTEIN	**SULFONAMIDE GEOMETRY**	**RMSD(Å)**
1HPV	1HPV-like	1.08
1HPV	Inverted	0.83
KB60	1HPV-like	1.31
KB60	Inverted	1.18

Table 4.5: RMSD between the corresponding non-hydrogen atoms in the docked and the crystal structure of KB45.

4.4 Discussion

The present study provides a prospective evaluation of the substrate envelope hypothesis as a basis for the design of HIVP inhibitors with broad specificity against clinically relevant variants of HIV protease. Incorporation of fit to the substrate envelope as a design criterion led to two new HIVP inhibitors of small size with relatively flat affinity profiles against a panel of clinically relevant mutants. The volumes of the ligands lying outside the substrate envelope were computed from crystal structures of their complexes with HIVP, and were found to be less than those of all but one of the clinical inhibitors. These results support the validity of the hypothesis that compounds which fit within the envelope will resist mutations. It is worth mentioning that higher affinity compounds, with greater susceptibility to mutation, might have been chosen if the goal of achieving high affinity had not been partly balanced by the goal of fitting within the substrate envelope.

Crystallographic studies of the new compounds AD37 and KB45 show generally good agreement between the predicted and crystal structures. However, the observed sulfonamide nitrogens are inverted relative to the conformation used in the design calculations. Further docking calculations with the inverted geometry indicate that the R_2 and R_3 substituents would have been more accurately positioned had the correct geometry been known in advance. It is possible that repeating the full design procedure with the inverted geometry would lead to compounds of greater affinity than those identified here. More generally, the present results indicate that there might be considerable value in a docking procedure that would automatically sample alternative geometries of invertible nitrogen atoms.

Although the present results are encouraging, it is not expected that a relatively blunt instrument like the substrate envelope criterion will be fully reliable. Utlimately, the subtleties of specific ligand-protein interactions will need to be considered. Nonetheless, the substrate envelope method may be useful, especially given its convenient simplicity and the inexactness of current ligand scoring functions.

4.5 Conclusions

Inhibitors of HIVP that were computationally designed to stay within the consensus substrate volume were found to have favorable resistance profiles when tested against a panel of protease variants with clinically relevant mutations, although the affinities are not as great as those of current clinical inhibitors. This result supports the validity of the substrate envelope hypothesis.

Chapter 5. General Discussion

Within approximately 15 years of the recognition of HIV protease as a viable target for AIDS, eight HIV protease inhibitors have been approved for the clinical use. The advent of these inhibitors greatly reduced the morbidity and mortality rate in AIDS patients. But unfortunately, the prevalence of treatment resistant strains has been observed to quickly rise within few years of treatment initiation(132). This rapid development of resistant strains is a major challenge in the AIDS therapy.

The main goal of our work was to develop and test an approach to the computational design of HIV protease inhibitors with minimal susceptibility to treatment resistant mutations. This project poses three major challenges: devising a computable quantity that might correlate with the robustness of an inhibitor against mutations; development of a efficient method of handling the combinatorial problem of library design; and selection of a scoring function or energy model that might be predictive of ligand affinity. The following paragraphs present a brief discussion of these challenges, our experience with them, and their possible or partial solution.

5.1 Design strategy

There are several workable strategies for the design of mutation resistant inhibitors. One is to design inhibitors that remain within the consensus substrate volume. This strategy is based on the observation that the primary active site resistance mutations occur at sites which are essential for inhibitor, but not substrate, recognition. We devised a quantitative measure of the fit of a candidate ligand to the substrate envelope and observed a correlation between it the observed clinical resistance. This result justified using it in our design of a combinatorial library of inhibitors, which yielded inhibitors with nanomolar affinity and good resistance profiles. As this simple method is not expected to be infallible, designed compounds could also be evaluated against other criteria and selected based upon a consensus scoring scheme.

Another approach would be to design compounds with high predicted affinity for the wild type protease and also for a panel of specific protease variants with clinically observed mutations (60). HIV-2 protease also could be used in place of a mutant HIV-1 protease, because it differs from HIV-1 protease at the residues that are prone to mutate (133). This approach requires docking and scoring of compounds into multiple similar protein structures. Because an efficient method of serial docking of ligands into multiple target structures has already been implemented and tested in our lab (134), it would fairly straightforward to use this approach to seek compounds with broad specificity. Other serial docking methods, such as the one studied by Lamb *etal* can also be used for this purpose (135).

Inhibitors could also be designed to interact only with main chain atoms and the conserved residues of the HIV protease. This approach is based on the observation that the overall shape of this protein remains constant in spite of differences in the substrates and ligands that bind it (31). This design method also would be easy to implement with our grid-based scoring function. Grids that store interaction potentials of only the main chain atoms and conserved residues can be readily generated by using a modified HIV protease, in which residues other than the conserved ones are replaced by alanine residues.

Finally, inhibitors that target nonstandard sites, such as the dimerization region or the open conformation of the active site, are also potential drug candidates to avoid resistance (132). Compound databases and combinatorial libraries can readily be screened against such targets.

5.2 Combinatorial library design

With the advent of automation technologies and high throughput screening, today's medicinal chemistry has enormous potential to yield drug leads. However, it still is not possible to synthesize and screen the billions and trillions of compounds that could in principle be built from a combinatorial scaffold and ever-expanding libraries of building blocks (77). Hence computational techniques are needed to guide selecting of a sublibrary of compounds for synthesis and testing. There are several virtual screening methods available for this purpose. Among them, structure-based drug design has been shown to have higher predictability and more efficiency (65), but such methods are computationally more expensive than other virtual screening methods. Even if it takes less than a minute to screen a compound, it could take years to screen a huge combinatorial library of compounds. This problem can be addressed by several optimization methods, such as simulated annealing and genetic algorithms. Alternatively, selection of substituents independently for each substituent position can make the combinatorial problem a linear one and circumvent the combinatorial explosion. This approach is based on the assumption that there will be little or no interaction between the substituents at different substituent positions.

We have evaluated this simple additivity method and used it in the selection of sublibraries of compounds from virtual combinatorial libraries of 3 billion compounds. Interestingly, the additivity method in general worked as well as the GA method. However, the assumption of additivity can break down under certain conditions, such as when a combinatorial scaffold has excessive conformational freedom, or when substituents contact each other. In our HIVP test system, the combinatorial scaffold has key interactions with the target protein which anchor it in the active site, and the crystal structure of a compound with this scaffold bound to HIVP confirms the validity of the positioning of the scaffold. This situation allowed reasonable restrictions to be placed on movement of the scaffold during the docking calculations. These restrictions also helped speed the calculations.

In some cases, one may not have prior information on the scaffold position. The literature suggests several computational approaches to this situation. The binding pose can be determined by docking compounds that share the same combinatorial scaffold (136), which can be obtained from the literature (136) or by building a few combinatorial compounds using methyl groups or diverse functional groups as substituents, or by using substituents that are common to the class of compounds that was studied (137). The consensus binding pose of the scaffold in all the docked poses of the screened compounds can then be selected for use in the additivity method. There is also one report that mentions the use of a bare scaffold (77), to obtain the binding pose. It would also be possible to use a GA for this purpose, with the expectation that the top scoring compounds from the final generation will have only a few distinct binding poses. The efficiency of the GA for this purpose can be further enhanced by limiting cross-over to the segments that are proximal in the active site.

The GA is a versatile optimization method which could be further tuned for combinatorial library design. For example, mutation operations could be biased in favor of choosing substituents that are chemically similar to moieties of a known ligand, or in favor of choosing substituents with chemotypes that have not yet been tested during the current GA run. Mutations could also be biased to choose substituents that score highly in single-site Additivity calculations, thus blending the Additivity and GA approaches. The fitness function can also be tuned for specific purposes, much as we combined the docking energy with the substrate envelope criterion by using Z-scores to combine fitness measures with different units. Other methods such as Pareto ranking can also be used to combine multiple objectives (70).

There are some preliminary filters that can also be used to increase the efficiency of combinatorial library design. For example, substituent libraries can be preselected to eliminate compounds with more than one functional group that can react in the selected combinatorial synthetic scheme, in order to generate designed compounds that avoid synthetic pitfalls. Subtituent libraries can also be sorted or filtered based on cost and availability, or enriched with the bio-isosteres or with the substituents from other compounds known to bind the protein target (137, 138). As the availability and binding affinity of these substituents to a similar target protein are already known, this approach in the selection of substituent library will appeal to the medicinal chemists tasked with synthesizing the designed libraries.

Lipinski's "rule of five" (139) , which seeks to differentiate drug-like compounds from others based on simple physical properties (molecular weight, number of hydrogen bond donors, acceptors and partition coefficient) can also be used as a preliminary filter to eliminate non-drug-like compounds from the combinatorial library before screening them with more computationally intensive docking and scoring functions. Other ligand-based virtual screening methods can also be used for pruning the huge combinatorial search space.

5.3 Docking and scoring functions

Even though docking and scoring functions are considered more reliable than ligand-based virtual screening methods, they have their own limitations. For example, they make gross approximations regarding the flexibility of the protein, or lack thereof, and in the treatment of solvent effects. These approximations severely limit the predictivity of affinity calculations, and a high level of accuracy is not routinely achievable (140). It is important for the user to be aware of the limitations of current docking and scoring methods, and to interpret their results in the light of the approximations they make. In spite of all these limitations, these methods can still contribute significantly to drug discovery.

Our in-house docking and scoring function, VDock led to several low nanomolar inhibitors with good resistance profiles. It is worth noting that the compounds were optimized not only for their docking energy scores, but also for their fit to the substrate envelope; compounds with higher affinity for the wild-type protease might have been discovered if the substrate envelope criterion had not been applied. Nonetheless, our energy scores correlated poorly with measured affinity (data not shown), so there is clearly much room for improvement. Prediction of binding energies is still a daunting task in the field of virtual screening.

As mentioned above, one source of error probably is the imprecise treatment of solvent effects. For example, VDock does not impose an energy penalty for removing a polar moiety from solvent upon binding. It might be possible to overcome this limitation, at least in part, by penalizing bound conformations with unsatisfied hydrogen bonding donors and acceptors groups. This would avoid the unfavorable placement of nonpolar groups next to polar groups. It might also be possible to include a limited number of explicit water molecules while docking candidate ligands. These would probably need to have rather limited freedom of movement to avoid computationally intensive calculations. Such water molecules could be restricted to locations where there is sufficient space to accommodate a water molecule and where there are unsatisfied hydrogen bonds in both the ligand and the receptor. Identification of such spots may require preliminary docking runs without water molecules, followed by a second set of docking runs with localized water molecules..

Another major missing element in scoring functions is change in configurational entropy on binding. This is often approximated based upon the number of rotatable bonds in the ligand, but recent calculations in our group suggest that such approaches are not wellfounded physically and markedly underestimate the entropy penalty. Ongoing research the group may lead to better approximations for this missing term.

It seems likely that improving the treatment of solvent and of configurational entropy will significantly improve the accuracy of docking and scoring calculations. Further improvement may be gained by combining multiple, complementary scoring functions, rather than relying upon just one. Such “consensus scoring” methods have been shown to significantly improve the yield of ligands in structure-based drug design (141).

Chapter 6. Conclusions

This thesis has described the design of HIV protease inhibitors with broad specificity. The main contributions are as follows: (1) A method to quantify the fit of a compound to the substrate envelope has been developed and evaluated, both retrospectively and prospectively. The method can be easily used in virtual high throughput docking and the design of combinatorial libraries. (2) A novel criterion for the clinical relevance of HIV protease mutations has been put forward. (3) A fast, simple Additivity method for the structure-based design of combinatorial libraries has been implemented, evaluated, employed in a real-world design project. (4) A Genetic Algorithm has also been developed for combinatorial library design. This can be useful for systems for which the Additivity method is not expected to be applicable. (5) Our work resulted in two low nanomolar compounds with favorable resistance profiles against a panel of clinically relevant resistance mutations.

References

1. Wang, W. K., M. Y. Chen, C. Y. Chuang, K. T. Jeang, and L. M. Huang. 2000. Molecular biology of human immunodeficiency virus type 1. Journal of microbiology, immunology, and infection = Wei mian yu gan ran za zhi 33:131-140.

2. UNAIDS. 2005. Global summary of the HIV/AIDS epidemic.

3. Gallo, R. C., S. Z. Salahuddin, M. Popovic, G. M. Shearer, M. Kaplan, B. F. Haynes, T. J. Palker, R. Redfield, J. Oleske, B. Safai, and et al. 1984. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. Science 224:500-503.

4. Dalgleish, A. G., P. C. Beverley, P. R. Clapham, D. H. Crawford, M. F. Greaves, and R. A. Weiss. 1984. The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. Nature 312:763-767.

5. Deng, H., R. Liu, W. Ellmeier, S. Choe, D. Unutmaz, M. Burkhart, P. Di Marzio, S. Marmon, R. E. Sutton, C. M. Hill, C. B. Davis, S. C. Peiper, T. J. Schall, D. R. Littman, and N. R. Landau. 1996. Identification of a major co-receptor for primary isolates of HIV-1. Nature 381:661-666.

6. Lang, W., H. Perkins, R. E. Anderson, R. Royce, N. Jewell, and W. Winkelstein, Jr. 1989. Patterns of T lymphocyte changes with human immunodeficiency virus infection: from seroconversion to the development of AIDS. Journal of acquired immune deficiency syndromes 2:63-69.

7. Levy, J. A. 1993. Pathogenesis of human immunodeficiency virus infection. Microbiological reviews 57:183-289.

8. Daluge, S. M., D. J. Purifoy, P. M. Savina, M. H. St Clair, N. R. Parry, I. K. Dev, P. Novak, K. M. Ayers, J. E. Reardon, G. B. Roberts, and et al. 1994. 5-Chloro-2',3'-dideoxy-3'-fluorouridine (935U83), a selective anti-human immunodeficiency virus agent with an improved metabolic and toxicological profile. Antimicrobial agents and chemotherapy 38:1590-1603.

9. Schinazi, R. F., J. P. Sommadossi, V. Saalmann, D. L. Cannon, M. Y. Xie, G. C. Hart, G. A. Smith, and E. F. Hahn. 1990. Activities of 3'-azido-3'-deoxythymidine nucleotide dimers in primary lymphocytes infected with human

immunodeficiency virus type 1. Antimicrobial agents and chemotherapy 34:1061-1067.

10. Schinazi, R. F., R. M. Lloyd, Jr., M. H. Nguyen, D. L. Cannon, A. McMillan, N. Ilksoy, C. K. Chu, D. C. Liotta, H. Z. Bazmi, and J. W. Mellors. 1993. Characterization of human immunodeficiency viruses resistant to oxathiolane-cytosine nucleosides. Antimicrobial agents and chemotherapy 37:875-881.

11. Chu, C. K., R. F. Schinazi, B. H. Arnold, D. L. Cannon, B. Doboszewski, V. B. Bhadti, and Z. P. Gu. 1988. Comparative activity of 2',3'-saturated and unsaturated pyrimidine and purine nucleosides against human immunodeficiency virus type 1 in peripheral blood mononuclear cells. Biochemical pharmacology 37:3543-3548.

12. Balzarini, J., A. van Aerschot, P. Herdewijn, and E. de Clercq. 1989. 5-Chloro-substituted derivatives of 2', 3'-didehydro-2',3'-dideoxyuridine, 3'-fluoro-2',3'-dideoxyuridine and 3'-azido-2',3'-dideoxyuridine as anti-HIV agents. Biochemical pharmacology 38:869-874.

13. Daluge, S. M., S. S. Good, M. B. Faletto, W. H. Miller, M. H. St Clair, L. R. Boone, M. Tisdale, N. R. Parry, J. E. Reardon, R. E. Dornsife, D. R. Averett, and T. A. Krenitsky. 1997. 1592U89, a novel carbocyclic nucleoside analog with potent, selective anti-human immunodeficiency virus activity. Antimicrobial agents and chemotherapy 41:1082-1093.

14. de Clercq, E. 1996. Non-nucleoside reverse transcriptase inhibitors (NNRTIs) for the treatment of human immunodeficiency virus type 1 (HIV-1) infections: strategies to overcome drug resistance development. Medicinal research reviews 16:125-157.

15. Romero, D. L., R. A. Morge, M. J. Genin, C. Biles, M. Busso, L. Resnick, I. W. Althaus, F. Reusser, R. C. Thomas, and W. G. Tarpley. 1993. Bis(heteroaryl)piperazine (BHAP) reverse transcriptase inhibitors: structure-activity relationships of novel substituted indole analogues and the identification of 1-[(5-methanesulfonamido-1H-indol-2-yl)-carbonyl]-4-[3- [(1-methylethyl)amino]-pyridinyl]piperazine monomethanesulfonate (U-90152S), a second-generation clinical candidate. Journal of medicinal chemistry 36:1505-1508.

16. Young, S. D., S. F. Britcher, L. O. Tran, L. S. Payne, W. C. Lumma, T. A. Lyle, J. R. Huff, P. S. Anderson, D. B. Olsen, S. S. Carroll, and et al. 1995. L-743, 726 (DMP-266): a novel, highly potent nonnucleoside inhibitor of the human immunodeficiency virus type 1 reverse transcriptase. Antimicrobial agents and chemotherapy 39:2602-2605.

17. Kim, E. E. B., C.T.; Dwyer, M.D.; Murcko, M.A.; Rao, B.G.; Tung, R.D.; Navia, M.A. 1995. Crystal structure of HIV-1 protease in complex with VX-478, a potent and orally bioavailable inhibitor of the enzyme. J Amer Chem Soc 117:1181-1182.

18. Vacca, J. P., B. D. Dorsey, W. A. Schleif, R. B. Levin, S. L. McDaniel, P. L. Darke, J. Zugay, J. C. Quintero, O. M. Blahy, E. Roth, and et al. 1994. L-735,524: an orally bioavailable human immunodeficiency virus type 1 protease inhibitor. Proceedings of the National Academy of Sciences of the United States of America 91:4096-4100.

19. 1997. New drugs--reports of new drugs recently approved by the FDA. Saquinavir. Bioorganic & medicinal chemistry 5:465-466.

20. Patick, A. K., H. Mo, M. Markowitz, K. Appelt, B. Wu, L. Musick, V. Kalish, S. Kaldor, S. Reich, D. Ho, and S. Webber. 1996. Antiviral and resistance studies of AG1343, an orally bioavailable inhibitor of human immunodeficiency virus protease. Antimicrobial agents and chemotherapy 40:292-297.

21. Kempf, D. J., K. C. Marsh, J. F. Denissen, E. McDonald, S. Vasavanonda, C. A. Flentge, B. E. Green, L. Fino, C. H. Park, X. P. Kong, and et al. 1995. ABT-538 is a potent inhibitor of human immunodeficiency virus protease and has high oral bioavailability in humans. Proceedings of the National Academy of Sciences of the United States of America 92:2484-2488.

22. Sham, H. L., D. J. Kempf, A. Molla, K. C. Marsh, G. N. Kumar, C. M. Chen, W. Kati, K. Stewart, R. Lal, A. Hsu, D. Betebenner, M. Korneyeva, S. Vasavanonda, E. McDonald, A. Saldivar, N. Wideburg, X. Chen, P. Niu, C. Park, V. Jayanti, B. Grabowski, G. R. Granneman, E. Sun, A. J. Japour, J. M. Leonard, J. J. Plattner, and D. W. Norbeck. 1998. ABT-378, a highly potent inhibitor of the human immunodeficiency virus protease. Antimicrobial agents and chemotherapy 42:3218-3224.

23. Robinson, B. S., K. A. Riccardi, Y. F. Gong, Q. Guo, D. A. Stock, W. S. Blair, B. J. Terry, C. A. Deminie, F. Djang, R. J. Colonno, and P. F. Lin. 2000. BMS-232632, a highly potent human immunodeficiency virus protease inhibitor that can be used in combination with other available antiretroviral agents. Antimicrobial agents and chemotherapy 44:2093-2099.

24. Kilby, J. M., S. Hopkins, T. M. Venetta, B. DiMassimo, G. A. Cloud, J. Y. Lee, L. Alldredge, E. Hunter, D. Lambert, D. Bolognesi, T. Matthews, M. R. Johnson, M. A. Nowak, G. M. Shaw, and M. S. Saag. 1998. Potent suppression of HIV-1 replication in humans by T-20, a peptide inhibitor of gp41-mediated virus entry. Nature medicine 4:1302-1307.

25. Sterne, J. A., M. A. Hernan, B. Ledergerber, K. Tilling, R. Weber, P. Sendi, M. Rickenbach, J. M. Robins, and M. Egger. 2005. Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death: a prospective cohort study. Lancet 366:378-384.

26. Kohl, N. E., E. A. Emini, W. A. Schleif, L. J. Davis, J. C. Heimbach, R. A. Dixon, E. M. Scolnick, and I. S. Sigal. 1988. Active human immunodeficiency virus protease is required for viral infectivity. Proceedings of the National Academy of Sciences of the United States of America 85:4686-4690.

27. McQuade, T. J., A. G. Tomasselli, L. Liu, V. Karacostas, B. Moss, T. K. Sawyer, R. L. Heinrikson, and W. G. Tarpley. 1990. A synthetic HIV-1 protease inhibitor with antiviral activity arrests HIV-like particle maturation. Science 247:454-456.

28. Kaplan, A. H., J. A. Zack, M. Knigge, D. A. Paul, D. J. Kempf, D. W. Norbeck, and R. Swanstrom. 1993. Partial inhibition of the human immunodeficiency virus type 1 protease results in aberrant virus assembly and the formation of noninfectious particles. Journal of virology 67:4050-4055.

29. Seelmeier, S., H. Schmidt, V. Turk, and K. von der Helm. 1988. Human immunodeficiency virus has an aspartic-type protease that can be inhibited by pepstatin A. Proceedings of the National Academy of Sciences of the United States of America 85:6612-6616.

30. Wlodawer, A., M. Miller, M. Jaskolski, B. K. Sathyanarayana, E. Baldwin, I. T. Weber, L. M. Selk, L. Clawson, J. Schneider, and S. B. Kent. 1989. Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. Science 245:616-621.

31. Wlodawer, A., and J. Vondrasek. 1998. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. Annual review of biophysics and biomolecular structure 27:249-284.

32. Wlodawer, A., and J. W. Erickson. 1993. Structure-based inhibitors of HIV-1 protease. Annual review of biochemistry 62:543-585.

33. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. Nucleic acids research 28:235-242.

34. Collins, J. R., S. K. Burt, and J. W. Erickson. 1995. Flap opening in HIV-1 protease simulated by 'activated' molecular dynamics. Nature structural biology 2:334-338.

35. Scott, W. R., and C. A. Schiffer. 2000. Curling of flap tips in HIV-1 protease as a mechanism for substrate entry and tolerance of drug resistance. Structure 8:1259-1265.

36. Piana, S., P. Carloni, and M. Parrinello. 2002. Role of conformational fluctuations in the enzymatic reaction of HIV-1 protease. Journal of molecular biology 319:567-583.

37. Prabu-Jeyabalan, M., E. Nalivaika, and C. A. Schiffer. 2000. How does a symmetric dimer recognize an asymmetric substrate? A substrate complex of HIV-1 protease. Journal of molecular biology 301:1207-1220.

38. Swain, A. L., M. M. Miller, J. Green, D. H. Rich, J. Schneider, S. B. Kent, and A. Wlodawer. 1990. X-ray crystallographic structure of a complex between a synthetic protease of human immunodeficiency virus 1 and a substrate-based hydroxyethylamine inhibitor. Proceedings of the National Academy of Sciences of the United States of America 87:8805-8809.

39. Baca, M., and S. B. Kent. 1993. Catalytic contribution of flap-substrate hydrogen bonds in "HIV-1 protease" explored by chemical synthesis. Proceedings of the National Academy of Sciences of the United States of America 90:11638-11642.

40. Beck, Z. Q., G. M. Morris, and J. H. Elder. 2002. Defining HIV-1 protease substrate selectivity. Current drug targets 2:37-50.

41. Griffiths, J. T., L. H. Phylip, J. Konvalinka, P. Strop, A. Gustchina, A. Wlodawer, R. J. Davenport, R. Briggs, B. M. Dunn, and J. Kay. 1992. Different requirements for productive interaction between the active site of HIV-1 proteinase and substrates containing -hydrophobic*hydrophobic- or -aromatic*pro- cleavage sites. Biochemistry 31:5193-5200.

42. Silva, A. M., R. E. Cachau, H. L. Sham, and J. W. Erickson. 1996. Inhibition and catalytic mechanism of HIV-1 aspartic protease. Journal of molecular biology 255:321-346.

43. Lebon, F., and M. Ledecq. 2000. Approaches to the design of effective HIV-1 protease inhibitors. Current medicinal chemistry 7:455-477.

44. Rodriguez-Barrios, F., and F. Gago. 2004. HIV protease inhibition: limited recent progress and advances in understanding current pitfalls. Current topics in medicinal chemistry 4:991-1007.

45. Velazquez-Campoy, A., S. Muzammil, H. Ohtaka, A. Schon, S. Vega, and E. Freire. 2003. Structural and thermodynamic basis of resistance to HIV-1 protease inhibition: implications for inhibitor design. Current drug targets 3:311-328.

46. Lin, Y., X. Lin, L. Hong, S. Foundling, R. L. Heinrikson, S. Thaisrivongs, W. Leelamanit, D. Raterman, M. Shah, B. M. Dunn, and et al. 1995. Effect of point mutations on the kinetics and the inhibition of human immunodeficiency virus type 1 protease: relationship to drug resistance. Biochemistry 34:1143-1152.

47. Stewart, K. D., and D. J. Kempf. 2004. An 'inside-the-box' approach to drug resistance. Chemistry & biology 11:1327-1328.

48. Shafer, R. W., D. Stevenson, and B. Chan. 1999. Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database. Nucleic acids research 27:348-352.

49. Ohtaka, H., and E. Freire. 2005. Adaptive inhibitors of the HIV-1 protease. Progress in biophysics and molecular biology 88:193-208.

50. Ala, P. J., E. E. Huston, R. M. Klabe, D. D. McCabe, J. L. Duke, C. J. Rizzo, B. D. Korant, R. J. DeLoskey, P. Y. Lam, C. N. Hodge, and C. H. Chang. 1997. Molecular basis of HIV-1 protease drug resistance: structural analysis of mutant proteases complexed with cyclic urea inhibitors. Biochemistry 36:1573-1580.

51. Kuroda, M. J., M. A. el-Farrash, S. Choudhury, and S. Harada. 1995. Impaired infectivity of HIV-1 after a single point mutation in the POL gene to escape the effect of a protease inhibitor in vitro. Virology 210:212-216.

52. Schock, H. B., V. M. Garsky, and L. C. Kuo. 1996. Mutational anatomy of an HIV-1 protease variant conferring cross-resistance to protease inhibitors in clinical trials. Compensatory modulations of binding and activity. The Journal of biological chemistry 271:31957-31963.

53. Muzammil, S., P. Ross, and E. Freire. 2003. A major role for a set of non-active site mutations in the development of HIV-1 protease drug resistance. Biochemistry 42:631-638.

54. Clemente, J. C., R. E. Moose, R. Hemrajani, L. R. Whitford, L. Govindasamy, R. Reutzel, R. McKenna, M. Agbandje-McKenna, M. M. Goodenow, and B. M. Dunn. 2004. Comparing the accumulation of active- and nonactive-site mutations in the HIV-1 protease. Biochemistry 43:12141-12151.

55. Doyon, L., G. Croteau, D. Thibeault, F. Poulin, L. Pilote, and D. Lamarre. 1996. Second locus involved in human immunodeficiency virus type 1 resistance to protease inhibitors. Journal of virology 70:3763-3769.

56. Zhang, Y. M., H. Imamichi, T. Imamichi, H. C. Lane, J. Falloon, M. B. Vasudevachari, and N. P. Salzman. 1997. Drug resistance during indinavir therapy is caused by mutations in the protease gene and in its Gag substrate cleavage sites. Journal of virology 71:6662-6670.

57. Zennou, V., F. Mammano, S. Paulous, D. Mathez, and F. Clavel. 1998. Loss of viral fitness associated with multiple Gag and Gag-Pol processing defects in human immunodeficiency virus type 1 variants selected for resistance to protease inhibitors in vivo. Journal of virology 72:3300-3306.

58. Yusa, K., and S. Harada. 2004. Acquisition of multi-PI (protease inhibitor) resistance in HIV-1 in vivo and in vitro. Current pharmaceutical design 10:4055-4064.

59. Yoshimura, K., R. Kato, M. F. Kavlick, A. Nguyen, V. Maroun, K. Maeda, K. A. Hussain, A. K. Ghosh, S. V. Gulnik, J. W. Erickson, and H. Mitsuya. 2002. A potent human immunodeficiency virus type 1 protease inhibitor, UIC-94003 (TMC-126), and selection of a novel (A28S) mutation in the protease active site. Journal of virology 76:1349-1358.

60. Freire, E. 2002. Designing drugs against heterogeneous targets. Nature biotechnology 20:15-16.

61. Tie, Y., P. I. Boross, Y.-F. Wang, L. Gaddis, F. Liu, X. Chen, J. Tozser, R. W. Harrison, and I. T. Weber. 2005. Molecular basis for substrate recognition and drug resistance from 1.1 to 1.6 A^0 resolution crystal structures of HIV-1 protease mutants with substrate analogs. 5265-5277.

62. Luque, I., M. J. Todd, J. Gomez, N. Semo, and E. Freire. 1998. Molecular basis of resistance to HIV-1 protease inhibition: a plausible hypothesis. Biochemistry 37:5791-5797.

63. Rose, R. B., C. S. Craik, and R. M. Stroud. 1998. Domain flexibility in retroviral proteases: structural implications for drug resistant mutations. Biochemistry 37:2607-2621.

64. Ellman, J., B. Stoddard, and J. Wells. 1997. Combinatorial thinking in chemistry and biology. Proceedings of the National Academy of Sciences of the United States of America 94:2779-2782.

65. Barril, X., R. E. Hubbard, and S. D. Morley. 2004. Virtual screening in structure-based drug discovery. Mini reviews in medicinal chemistry 4:779-791.

66. Oprea, T. I. 2000. Property distribution of drug-related chemical databases*. Journal of computer-aided molecular design 14:251-264.

67. Xue, L., J. W. Godden, and J. Bajorath. 1999. Database Searching for Compounds with Similar Biological Activity Using Short Binary Bit String Representations of Molecules. In J. Chem. Inf. Comput. Sci. 881-886.

68. Willett, P., J. M. Barnard, and G. M. Downs. 1998. Chemical Similarity Searching. In J. Chem. Inf. Comput. Sci. 983-996.

69. Good, A. C., and R. A. Lewis. 1997. New methodology for profiling combinatorial libraries and screening sets: cleaning up the design process with HARPick. Journal of medicinal chemistry 40:3926-3936.

70. Gillet, V. J., W. Khatib, P. Willett, P. J. Fleming, and D. V. S. Green. 2002. Combinatorial Library Design Using a Multiobjective Genetic Algorithm. In J. Chem. Inf. Comput. Sci. 375-385.

71. Schneider, G., M. L. Lee, M. Stahl, and P. Schneider. 2000. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. Journal of computer-aided molecular design 14:487-494.

72. Schneider, G. 2002. Trends in virtual combinatorial library design. Current medicinal chemistry 9:2095-2101.

73. Gehlhaar, D. K., K. E. Moerder, D. Zichi, C. J. Sherman, R. C. Ogden, and S. T. Freer. 1995. De novo design of enzyme inhibitors by Monte Carlo ligand generation. Journal of medicinal chemistry 38:466-472.

74. Grzybowski, B. A., A. V. Ishchenko, C. Y. Kim, G. Topalov, R. Chapman, D. W. Christianson, G. M. Whitesides, and E. I. Shakhnovich. 2002. Combinatorial computational method gives new picomolar ligands for a known enzyme. Proceedings of the National Academy of Sciences of the United States of America 99:1270-1273.

75. Bohm, H. J. 1994. On the use of LUDI to search the Fine Chemicals Directory for ligands of proteins of known three-dimensional structure. Journal of computer-aided molecular design 8:623-632.

76. Pearlman, D. A., and M. A. Murcko. 1996. CONCERTS: dynamic connection of fragments as an approach to de novo ligand design. Journal of medicinal chemistry 39:1651-1663.

77. Sun, Y., T. J. Ewing, A. G. Skillman, and I. D. Kuntz. 1998. CombiDOCK: structure-based combinatorial docking and library design. Journal of computer-aided molecular design 12:597-604.

78. Murray, C. W., D. E. Clark, T. R. Auton, M. A. Firth, J. Li, R. A. Sykes, B. Waszkowycz, D. R. Westhead, and S. C. Young. 1997. PRO_SELECT: combining structure-based drug design and combinatorial chemistry for rapid lead discovery. 1. Technology. Journal of computer-aided molecular design 11:193-207.

79. Elaine C. Meng, B. K. S. I. D. K. 1992. Automated docking with grid-based energy evaluation. In J. Comput. Chem. 505-524.

80. Jones, G., P. Willett, and R. C. Glen. 1995. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. Journal of molecular biology 245:43-53.

81. Goodsell, D. S., and A. J. Olson. 1990. Automated docking of substrates to proteins by simulated annealing. Proteins 8:195-202.

82. Rarey, M., S. Wefing, and T. Lengauer. 1996. Placement of medium-sized molecular fragments into active sites of proteins. Journal of computer-aided molecular design 10:41-54.

83. Baxter, C. A., C. W. Murray, D. E. Clark, D. R. Westhead, and M. D. Eldridge. 1998. Flexible docking using Tabu search and an empirical estimate of binding affinity. Proteins 33:367-382.

84. Schulz-Gasch, T., and M. Stahl. 2003. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. Journal of molecular modeling (Online) 9:47-57.

85. David, L., R. Luo, and M. K. Gilson. 2001. Ligand-receptor docking with the Mining Minima optimizer. Journal of computer-aided molecular design 15:157-171.

86. Kairys, V., and M. K. Gilson. 2002. Enhanced docking with the mining minima optimizer: acceleration and side-chain flexibility. Journal of computational chemistry 23:1656-1670.

87. Garrett M. Morris, D. S. G. R. S. H. R. H. W. E. H. R. K. B. A. J. O. 1998. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. 1639-1662.

88. Given, J. A., and M. K. Gilson. 1998. A hierarchical method for generating low-energy conformers of a protein-ligand complex. Proteins 33:475-495.

89. K. W. Foreman, A. T. P. J. B. R. K. A. D. 1999. Comparing search strategies for finding global optima on energy landscapes. 1527-1532.

90. Sheridan, R. P., Kearsley,S.K. 1995. Using a genetic algorithm to suggest combinatorial libraries. J. Chem. Inf. Comput. Sci 35:310-320.

91. DongXiang Liu, H. J., KaiXian Chen, and RuYun Ji. 1998. A New Approach to Design Virtual Combinatorial Library with Genetic Algorithm Based on 3D Grid Property. J. Chem. Inf. Comput. Sci 38:233-242.

92. Hammer, S. M., K. E. Squires, M. D. Hughes, J. M. Grimes, L. M. Demeter, J. S. Currier, J. J. Eron, Jr., J. E. Feinberg, H. H. Balfour, Jr., L. R. Deyton, J. A. Chodakewitz, and M. A. Fischl. 1997. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. AIDS Clinical Trials Group 320 Study Team. The New England journal of medicine 337:725-733.

93. Molla, A., G. R. Granneman, E. Sun, and D. J. Kempf. 1998. Recent developments in HIV protease inhibitor therapy. Antiviral research 39:1-23.

94. Coffin, J. M. 1995. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science 267:483-489.

95. Prabu-Jeyabalan, M., E. A. Nalivaika, N. M. King, and C. A. Schiffer. 2003. Viability of a drug-resistant human immunodeficiency virus type 1 protease variant: structural insights for better antiviral therapy. Journal of virology 77:1306-1315.

96. King, N. M., M. Prabu-Jeyabalan, E. A. Nalivaika, and C. A. Schiffer. 2004. Combating susceptibility to drug resistance: lessons from HIV-1 protease. Chemistry & biology 11:1333-1338.

97. Surleraux, D. L., A. Tahri, W. G. Verschueren, G. M. Pille, H. A. de Kock, T. H. Jonckers, A. Peeters, S. De Meyer, H. Azijn, R. Pauwels, M. P. de Bethune, N. M. King, M. Prabu-Jeyabalan, C. A. Schiffer, and P. B. Wigerinck. 2005.

Discovery and selection of TMC114, a next generation HIV-1 protease inhibitor. Journal of medicinal chemistry 48:1813-1822.

98. King, N. M., M. Prabu-Jeyabalan, E. A. Nalivaika, P. Wigerinck, M. P. de Bethune, and C. A. Schiffer. 2004. Structural and thermodynamic basis for the binding of TMC114, a next-generation human immunodeficiency virus type 1 protease inhibitor. Journal of virology 78:12012-12021.

99. Morris, R. J. 2004. Statistical pattern recognition for macromolecular crystallographers. Acta crystallographica 60:2133-2143.

100. Chen, Z., Y. Li, E. Chen, D. L. Hall, P. L. Darke, C. Culberson, J. A. Shafer, and L. C. Kuo. 1994. Crystal structure at 1.9-A resolution of human immunodeficiency virus (HIV) II protease complexed with L-735,524, an orally bioavailable inhibitor of the HIV proteases. The Journal of biological chemistry 269:26344-26348.

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

102. E. E. Kim, C. T. B., M. D. Dwyer, M. A. Murcko, B. G. Rao, R. D. Tung, and M. A. Navia. 1995. Crystal structure of HIV-1 protease in complex with VX-478, a potent and orally bioavailable inhibitor of the enzyme. 117:1181 – 1182.

103. Krohn, A., S. Redshaw, J. C. Ritchie, B. J. Graves, and M. H. Hatada. 1991. Novel binding mode of highly potent HIV-proteinase inhibitors incorporating the (R)-hydroxyethylamine isostere. Journal of medicinal chemistry 34:3340-3342.

104. Ohtaka, H., A. Schon, and E. Freire. 2003. Multidrug resistance to HIV-1 protease inhibition requires cooperative coupling between distal mutations. Biochemistry 42:13659-13666.

105. Velazquez-Campoy, A., S. Vega, and E. Freire. 2002. Amplification of the effects of drug resistance mutations by background polymorphisms in HIV-1 protease from African subtypes. Biochemistry 41:8613-8619.

106. http://hivdb.stanford.edu/cgi-bin/PRMut.cgi.

107. Nikitin, S., N. Zaitseva, O. Demina, V. Solovieva, E. Mazin, S. Mikhalev, M. Smolov, A. Rubinov, P. Vlasov, D. Lepikhin, D. Khachko, V. Fokin, C. Queen, and V. Zosimov. 2005. A very large diversity space of synthetically accessible compounds for use with drug design programs. Journal of computer-aided molecular design 19:47-63.

108. Sheridan, R. P., S. G. SanFeliciano, and S. K. Kearsley. 2000. Designing targeted libraries with genetic algorithms. Journal of molecular graphics & modelling 18:320-334, 525.

109. Pegg, S. C., J. J. Haresco, and I. D. Kuntz. 2001. A genetic algorithm for structure-based de novo design. Journal of computer-aided molecular design 15:911-933.

110. Brown, R. D., and Y. C. Martin. 1997. Designing combinatorial library mixtures using a genetic algorithm. Journal of medicinal chemistry 40:2304-2313.

111. Gillet, V. J., W. Khatib, P. Willett, P. J. Fleming, and D. V. Green. 2002. Combinatorial library design using a multiobjective genetic algorithm. Journal of chemical information and computer sciences 42:375-385.

112. Wright, T., V. J. Gillet, D. V. Green, and S. D. Pickett. 2003. Optimizing the size and configuration of combinatorial libraries. Journal of chemical information and computer sciences 43:381-390.

113. Agrafiotis, D. K., and W. Cedeno. 2002. Feature selection for structure-activity correlation using binary particle swarms. Journal of medicinal chemistry 45:1098-1107.

114. van Soest, A. J., and L. J. Casius. 2003. The merits of a parallel genetic algorithm in solving hard optimization problems. Journal of biomechanical engineering 125:141-146.

115. Greg Burns, R. D., James Vaigl. 1994. LAM: An Open Cluster Environment for MPI.

116. Sprous, D. G., D. R. Lowis, J. M. Leonard, T. Heritage, S. N. Burkett, D. S. Baker, and R. D. Clark. 2004. OptiDock: virtual HTS of combinatorial libraries by efficient sampling of binding modes in product space. Journal of combinatorial chemistry 6:530-539.

117. Kim, E. E., Baker, C.T., Dwyer, M.D., Murcko, M.A., Rao, B.G., Tung, R.D., Navia, M.A. . 1995. Crystal Structure of HIV-1 Protease in Complex with Vx-

478, a Potent and Orally Bioavailable Inhibitor of the Enzyme J.Am.Chem.Soc 117.

118. Baldwin, E. T., T. N. Bhat, S. Gulnik, M. V. Hosur, R. C. Sowder, 2nd, R. E. Cachau, J. Collins, A. M. Silva, and J. W. Erickson. 1993. Crystal structures of native and inhibited forms of human cathepsin D: implications for lysosomal targeting and drug design. Proceedings of the National Academy of Sciences of the United States of America 90:6796-6800.

119. Irwin, J. J., and B. K. Shoichet. 2005. ZINC--a free database of commercially available compounds for virtual screening. Journal of chemical information and modeling 45:177-182.

120. Gilson, M. K., H. S. Gilson, and M. J. Potter. 2003. Fast assignment of accurate partial atomic charges: an electronegativity equalization method that accounts for alternate resonance forms. Journal of chemical information and computer sciences 43:1982-1997.

121. Dash, C., A. Kulkarni, B. Dunn, and M. Rao. 2003. Aspartic peptidase inhibitors: implications in drug development. Critical reviews in biochemistry and molecular biology 38:89-119.

122. Fusek, M., and V. Vetvicka. 2005. Dual role of cathepsin D: ligand and protease. Biomedical papers of the Medical Faculty of the University Palacky, Olomouc, Czechoslovakia 149:43-50.

123. Koehl, P. 2006. Electrostatics calculations: latest methodological advances. Current opinion in structural biology 16:142-151.

124. Baker, N. A. 2005. Improving implicit solvent simulations: a Poisson-centric view. Current opinion in structural biology 15:137-143.

125. Prabu-Jeyabalan, M., E. Nalivaika, and C. A. Schiffer. 2002. Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes. Structure 10:369-381.

126. Rajeshwar D. Bindal, J. T. G., and John A. Katzenellenbogen. 1990. J. Am. Chem. SOC. 112:7861-7868.

127. Kim, E. E., Baker, C.T., Dwyer, M.D., Murcko, M.A., Rao, B.G., Tung, R.D., Navia, M.A. 1995. Crystal Structure of HIV-1 Protease in Complex with Vx-478, a Potent and Orally Bioavailable Inhibitor of the Enzyme J.Am.Chem.Soc 117:1181.

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

129. Stephen L. Mayo, B. D. O., William A. Goddard III. 1990. DREIDING: a generic force field for molecular simulations J. Phys. Chem 94:8897-8909.

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

131. Foo, L. C., and M. Mafauzy. 1999. Does the use of mean or median Z-score of the thyroid volume indices provide a more precise description of the iodine deficiency disorder status of a population? European journal of endocrinology / European Federation of Endocrine Societies 141:557-560.

132. Prejdova, J., M. Soucek, and J. Konvalinka. 2004. Determining and overcoming resistance to HIV protease inhibitors. Current drug targets 4:137-152.

133. Weber, J., J. R. Mesters, M. Lepsik, J. Prejdova, M. Svec, J. Sponarova, P. Mlcochova, K. Skalicka, K. Strisovsky, T. Uhlikova, M. Soucek, L. Machala, M. Stankova, J. Vondrasek, T. Klimkait, H. G. Kraeusslich, R. Hilgenfeld, and J. Konvalinka. 2002. Unusual binding mode of an HIV-1 protease inhibitor explains its potency against multi-drug-resistant virus strains. Journal of molecular biology 324:739-754.

134. Fernandes, M. X., V. Kairys, and M. K. Gilson. 2004. Comparing Ligand Interactions with Multiple Receptors via Serial Docking. 1961-1970.

135. Lamb, M. L., K. W. Burdick, S. Toba, M. M. Young, A. G. Skillman, X. Zou, J. R. Arnold, and I. D. Kuntz. 2001. Design, docking, and evaluation of multiple libraries against multiple targets. Proteins 42:296-318.

136. Chema, D., D. Eren, A. Yayon, A. Goldblum, and A. Zaliani. 2004. Identifying the binding mode of a molecular scaffold. Journal of computer-aided molecular design 18:23-40.

137. Lowrie, J. F., R. K. Delisle, D. W. Hobbs, and D. J. Diller. 2004. The different strategies for designing GPCR and kinase targeted libraries. Combinatorial chemistry & high throughput screening 7:495-510.

138. Pierce, A. C., G. Rao, and G. W. Bemis. 2004. BREED: Generating novel inhibitors through hybridization of known ligands. Application to CDK2, p38, and HIV protease. Journal of medicinal chemistry 47:2768-2775.

139. Lipinski, C. A. 2003. Chris Lipinski discusses life and chemistry after the Rule of Five. Drug discovery today 8:12-16.

140. Mohan, V., A. C. Gibbs, M. D. Cummings, E. P. Jaeger, and R. L. DesJarlais. 2005. Docking: successes and challenges. Current pharmaceutical design 11:323-333.

141. Feher, M. 2006. Consensus scoring for protein-ligand interactions. Drug discovery today 11:421-428.