

Computational Ligand Design and Analysis in
Protein Complexes Using Inverse Methods,
Combinatorial Search, and Accurate Solvation
Modeling

by

Michael Darren Altman

BACHELOR OF SCIENCE IN BIOLOGY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY, 1999

Submitted to the Department of Chemistry
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Biological Chemistry

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2006

© Massachusetts Institute of Technology 2006. All rights reserved.

Author

Department of Chemistry

✓ May 8, 2006

Certified by

Bruce Tidor

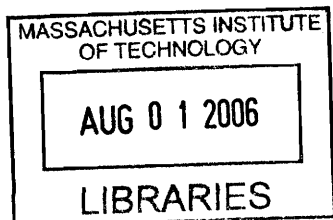
Professor of Biological Engineering and Computer Science

Thesis Supervisor

Accepted by

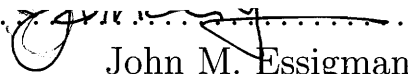
Robert W. Field

Chairman, Department Committee on Graduate Students

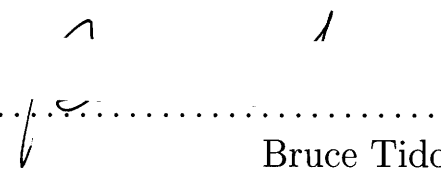


ARCHIVES

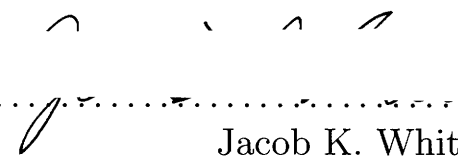
This thesis has been examined by a committee as follows:

Certified By 

John M. Essigmann
Thesis Committee Chair

Certified By 

Bruce Tidor
Thesis Supervisor

Certified By 

Jacob K. White
Reader

Computational Ligand Design and Analysis in Protein Complexes Using Inverse Methods, Combinatorial Search, and Accurate Solvation Modeling

by

Michael Darren Altman

Submitted to the Department of Chemistry
on May 8, 2006, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Biological Chemistry

Abstract

This thesis presents the development and application of several computational techniques to aid in the design and analysis of small molecules and peptides that bind to protein targets. First, an inverse small-molecule design algorithm is presented that can explore the space of ligands compatible with binding to a target protein using fast combinatorial search methods. The inverse design method was applied to design inhibitors of HIV-1 protease that should be less likely to induce resistance mutations because they fit inside a consensus substrate envelope. Fifteen designed inhibitors were chemically synthesized, and four of the tightest binding compounds to the wild-type protease exhibited broad specificity against a panel of drug resistance mutant proteases in experimental tests. Inverse protein design methods and charge optimization were also applied to improve the binding affinity of a substrate peptide for an inactivated mutant of HIV-1 protease, in an effort to learn more about the thermodynamics and mechanisms of peptide binding. A single mutant peptide calculated to have improved binding electrostatics exhibited greater than 10-fold improved affinity experimentally. The second half of this thesis presents an accurate method for evaluating the electrostatic component of solvation and binding in molecular systems, based on curved boundary-element method solutions of the linearized Poisson–Boltzmann equation. Using the presented FFTSVD matrix compression algorithm and other techniques, a full linearized Poisson–Boltzmann equation solver is described that is capable of solving multi-region problems in molecular continuum electrostatics to high precision.

Thesis Supervisor: Bruce Tidor

Title: Professor of Biological Engineering and Computer Science

Acknowledgments

First and foremost, I would like to thank my wife Rhonda for all of her support and patience throughout my graduate career. I would also like to thank my parents and sister for supporting and encouraging me throughout the past seven years.

I am also indebted to Prof. Bruce Tidor, my thesis advisor, for continuous academic advice, for teaching me the most important aspects of being a good scientist, and for encouraging me to work on small-molecule design problems. I would also like to thank Prof. Jacob White, who treated me as his own graduate student, and encouraged me to explore my interest in numerical methods for molecular design.

The work presented in this thesis would not have been possible without the support of current and previous members of the Tidor lab. I am especially indebted to Jay Bardhan, my close collaborator on all of the boundary-element PBE solver work, and to David Green who began the inverse design for small molecules project. I would also like to thank Brian Joughin, Shaun Lippow, Bracken King, Mala Radhakrishnan, Kathryn Armstrong, Aurore Zyto, David Huggins, Bambang Adiwijaya, Joshua Apgar, Caitlin Bever, Katharina Wilkins, Jared Toettcher, Patricio Ramirez, Fa Chunsriviro, Woody Sherman, Justin Caravella, Zachary Hendsch, Alexander Akhiezer, Philip Kim, Roy Kimura, Erik Kangas, Mark Bathe, Karl Hanf, Alessandro Senes, and Lukasz Weber for their contributions over the years.

Throughout my thesis work I had the opportunity to take part in two collaborative research efforts. The first centered around testing the substrate envelope hypothesis to avoid drug resistance using HIV-1 protease as a model system. I would like to thank Profs. Celia Schiffer, Tariq Rana, Mike Gilson, Ron Swanstrom, and Bob Shafer for working with me on this project and dedicating time in their laboratories to test my computational predictions experimentally. Many students from these laboratories were also involved in the project, and I would like to personally thank Moses Prabu-Jeyabalan, Madhavi Nalam, Sripriya Chellappan, Akbar Ali, Kiran Reddy, and Hong Cao for their time and effort. The second collaboration was with the laboratory of Prof. Jacob White, and focused on developing more accurate numerical techniques for performing continuum electrostatic calculations. I would like to thank several present and former members of his laboratory, including Jay Bardhan, Shihhsien Kuo, David Willis, and Zhenhai Zhu for their advice and contributions.

Contents

1	General Introduction	9
2	Inverse Design of Small-molecule Ligands for Drug Discovery	19
2.1	Introduction	20
2.2	Theoretical and computational approach	23
2.3	Results and Discussion	35
2.4	Materials and Methods	51
3	Computational Design of Substrate Envelope Inhibitors to Avoid Resistance Mutations: HIV-1 Protease as a Model System	63
3.1	Introduction	64
3.2	Computational strategy	67
3.3	Results and Discussion	72
3.4	Conclusions and Future Work	83
3.5	Materials and Methods	86
4	Computational design and experimental study of tighter binding peptides to an inactivated mutant of HIV-1 protease	93
4.1	Introduction	94
4.2	Results and Discussion	97
4.3	Conclusions	117

4.4	Materials and Methods	120
5	FFTSVD: A Fast Multiscale Boundary Element Method Solver Suitable for BioMEMS and Biomolecule Simulation	127
5.1	Introduction	128
5.2	Background examples	131
5.3	The FFTSVD algorithm	135
5.4	Computational results	146
5.5	Discussion	154
6	Accurate Solution of Multi-region Continuum Electrostatic Problems Using the Linearized Poisson–Boltzmann Equation and Curved Boundary Elements	157
6.1	Introduction	158
6.2	Theory	163
6.3	Computational details	182
6.4	Results and Discussion	186
6.5	Conclusions	198
7	General Conclusions	201

Chapter 1

General Introduction

Automated design of small molecules and peptides that bind tightly and specifically to protein targets remains a formidable challenge in computational biochemistry. Two primary challenges must be overcome to develop such an algorithm. First, the space of molecular diversity, conformation, and orientation that must be searched to identify tight binders is extremely prohibitive. For example, consider the space of short dodecameric peptides containing natural amino acids, for which there are 20^{10} sequences. All reasonable conformations of these peptides would need to be tested against all reasonable conformations of the protein target in all possible orientations within a potential binding site in order to identify those with desired properties. When designing small molecules, the space of molecular diversity is almost limitless, bounded only by organic chemistry. Consequently, almost all molecular design algorithms employ techniques to search this space highly approximately, in an attempt to focus in on regions of the space likely to be enriched in geometries compatible with tight binding.

Even if searching such a space was possible, there would still be a challenge in correctly determining which members of the space, if any, suggest that the designed small molecule or peptide is indeed going to bind the target tightly. Current molecular design algorithms handle this through the use of a scoring function, which takes as input the geometry of the ligand-protein complex and outputs a numerical score indicating the likelihood that this structure predicts tight binding. Due to the massive search space that must be evaluated, scoring functions need to be very fast, and are consequently highly approximate. Often, these scoring functions are based on phenomenology instead of physical models, and even the energetic models themselves

are computed inexactly.

This thesis presents the development and application of several computational techniques that attempt to address some of these challenges in molecular design. The first method, known as inverse design, can reformulate the molecular design problem in a way that reduces the size of the search space while still retaining the ability to design tight-binding molecules. Inverse design methods originated in the field of airplane wing design in the 1950s, and were viewed as an alternative to traditional engineering design approaches. In traditional forward design, a prototype is initially built using guidelines and tested to determine whether or not it meets a set of design specifications. If not, the object being designed is modified in a way more likely to yield better results, and it is tested again against the specifications. This process is iterated until the design is successful.

In the realm of protein design, the forward approach would most closely correspond to starting with a polypeptide sequence, solving the protein-folding problem, and testing to see if the folded protein has the desired structure or function. If not, mutations would be made to the sequence and the process repeated. Obviously, this approach is intractable considering the difficulty of the protein-folding problem [1,2]. For small molecule design, forward design methods correspond to taking a molecule and using docking calculations [3–6] to determine its best position inside the binding site and using scoring functions to determine the likelihood of tight binding. If the score is low, then a different small molecule would be selected, docked, and evaluated for fitness. Although this “virtual screening” method seems computationally inefficient, it is one of the most commonly used techniques for computational drug discovery [5, 7, 8].

In contrast, inverse design methods begin with the set of specifications, and try to directly solve for designs that meet these goals. Inverse methods were first proposed for the design of proteins by Drexler [9] and Pabo [10] in the early 1980’s. If the goal

of a protein design effort is to design a protein with a particular structure, inverse methods begin with a fixed model for the protein backbone. Given this backbone, the only design problem remaining is to search over the space of amino acid side-chain identities and conformations to determine those that stabilize the fold. It turns out that the search over side chains, independent of backbone flexibility, is an approachable problem, unlike the protein-folding problem that allows all degrees of freedom to vary simultaneously. Given certain assumptions about the scoring function used to determine side-chain fitness, and given a discretization of side-chain geometries into rotamers, fast combinatorial search algorithms have been developed to solve the side-chain placement problem [11–14]. The hallmark of inverse design methods is that they take certain specifications as constraints or restraints, leaving behind a tractable search problem in a dramatically reduced space.

Although inverse design methods have been extensively applied to design proteins with higher stability [12, 15], novel structures [16, 17], and alternative binding interfaces [18–21], they have not been extensively applied to the small-molecule design problem in drug discovery. Chapter 2 of this thesis presents one possible implementation of an inverse strategy for small-molecule ligand design. Using the structure of a protein target, and knowledge of a binding site for ligands, we solve an inverse shape problem and an inverse electrostatics problem to determine the “optimal” theoretical ligand that would have the best binding properties given our computational models. These optimal shape and electrostatic features serve the same role as the fixed protein backbone in inverse protein design. The remaining search problem is to identify real molecules that reproduce the properties of the “optimal” theoretical ligand.

The inverse shape problem, given the computational models presented in this thesis, is relatively straightforward. Two components of the energetic models used here are only functions of the ligand geometry. These include the van der Waals contributions to the binding energy, and the non-polar contribution to the solvation free

energy, which is often referred to as the hydrophobic effect. In the models used here, the van der Waals binding energy gets more favorable as ligand atoms are packed into the binding site such that they make close contacts with the receptor. Consequently, the more of the binding site that the ligand occupies, the more opportunities it has to make favorable van der Waals packing interactions. Similarly, the non-polar solvation contribution is modeled as being directly proportional to the surface area buried upon binding [22], meaning that the more of the binding site the ligand fills, the more favorable the binding energy is predicted to be. Therefore, given these energetic models, the optimal ligand shape would be a negative image of the binding site, as this shape maximally fills the site.

The inverse electrostatic problem, however, is significantly more challenging. The goal of the inverse electrostatic problem is to determine the charge distribution that a ligand should have to minimize the electrostatic component of its binding free energy. If biological systems occurred in vacuum, the solution to the inverse electrostatics problem would be trivial, as the binding energy of the ligand could be increased without bound by assigning it larger and larger charge with an opposite sign to the receptor's. However, the presence of an aqueous environment disfavors large ligand charges, as ligands must pay a desolvation penalty before binding. This cost has been shown to eventually exceed the benefit obtained from increased direct electrostatic interaction with the target protein [23, 24], leading to the presence of a charge optimum. Previous work has solved the inverse electrostatics problem in the context of a mixed discrete-continuum electrostatics model, where solute molecules and solvent are represented by homogeneous regions of low and high dielectric, respectively. Point charges within the low dielectric region represent the partial atomic charges of atoms, and a Debye–Hückel treatment is used to represent ionic strength in the solvent region. The classical electrostatic properties of this system are governed by the Poisson–Boltzmann equation [25–29], and the use of its linearized form permits the

ligand charges that optimize the binding free energy to be obtained using a matrix equation derived from its solution in the bound and unbound state [23, 24, 30, 31]. The optimal charge distribution within the optimal shape can be obtained on a cubic lattice, which is sufficient to approximate optimal electrostatic properties.

Given the optimal shape and electrostatic properties of an ideal ligand, how can the essentially unlimited space of real small molecules be effectively searched to identify those most similar to these properties? Two techniques were considered to solve this problem. In the first method, the optimal shape and charge distribution were considered as hard constraints, and computer vision algorithms, such as object-based model recognition [32] or geometric hashing [33, 34], were used to identify molecules that looked like portions of the ideal ligand from a set of generated compounds. These compounds could then be combinatorially extended with various functional groups in an attempt to sequentially match more of the ideal target. However, obtaining consistent results with this method was difficult, most likely due to the fact that the optimal shape and charge distribution had little similarity to real molecules.

Chapter 2 of this thesis presents an alternative method that met with more success. Rather than treat the shape and charge distribution as hard constraints, molecules being designed were directly tested for near-optimality by scoring the compounds in the objective functions used to determine the optimal shape and charge distributions. This scoring function was designed to have a similar functional form to those used in the side-chain placement problem in inverse protein design, allowing all of the fast combinatorial search algorithms developed in that field to be applied here. Molecular diversity was approximated by using a scaffold and functional group scheme, where the space of discrete functional group attachments was combinatorially searched for many discrete positions of the scaffold inside the binding site. Because it was rare that the shape of designed compounds exactly matched the optimal ligand shape, methods were developed to re-score compounds using the real molecular shape such that

guarantees provided by the combinatorial search techniques were nearly maintained. Overall, these features allowed the development of a full small-molecule design technique, applicable to solving problems in *de novo* ligand design as well as the ability to evaluate combinatorial libraries in their fitness for a binding site.

Chapter 3 of this thesis presents an application of the inverse design method to solving a real biological problem, which is the design of novel inhibitors that are less likely to induce drug resistance mutations. Drug resistance is a growing problem in the treatment of rapidly evolving pathogens [35–42], and drug development strategies need to be developed to minimize its occurrence. One such strategy, which is most applicable to cases where the drug target is highly mutable and an essential enzyme for the pathogen, is to create inhibitors that mimic the interactions that substrates make with the binding site. This idea, termed the “substrate envelope hypothesis” states that if an inhibitor stays within the consensus volume of substrates, resistance mutations that would disrupt inhibitor binding would also disrupt substrate processing, rendering the pathogen non-viable.

In order to begin testing the substrate envelope hypothesis as an inhibitor design principle, HIV-1 protease was employed as a model system. The structures of substrate complexes with inactivated HIV-1 protease have been determined [43,44], and the consensus substrate envelope was used as a computational small-molecule inverse design target rather than an optimal shape that fills the entire binding site. In order to target the envelope, inverse design was performed using a molecular scaffold derived from known HIV-1 protease inhibitors, and a naive functional group library from chemical catalogs was used to diversify the scaffold at three positions. Fifteen computationally designed compounds were synthesized by collaborators and tested for binding against the wild-type protease. Of these compounds, four had inhibition constants (K_i) within 30–50 nM. To begin testing whether or not these compounds would induce resistance mutations, the inhibition constant for these compounds against a

panel of three drug-resistance HIV-1 protease mutants was experimentally measured. Overall, the compounds lost no more than 6–13 fold inhibition against the mutant proteases, relative to wild type. This is in sharp contrast to first generation HIV-1 protease inhibitors, which often lose more than 1000-fold inhibition relative to wild type in drug resistant mutants. This small molecule design project was also very useful in validating and identifying limitations in the computational inverse design technique. A comparison of predicted to experimental binding energies highlighted weaknesses associated with rigid treatment of the protease in initial design calculations. Switching to design in a protease structure more compatible with the scaffold used for design results in improved rankings. Crystal structures of the four tightest binding design inhibitors were also determined by collaborators, and the structures agreed well with prediction when design was performed without the substrate envelope constraint.

In Chapter 4, two inverse design methods were directly applied to improving the binding affinity of peptides for an inactivated mutant of HIV-1 protease. As demonstrated in Chapter 3, knowledge of how substrates bind to the protease is of value in learning about drug resistance. One important inhibitor property that has been shown to be useful in predicting drug resistance profiles is the balance between enthalpic and entropic contributions to the binding free energy [45, 46], as determined through calorimetry. Consequently, it was of interest to examine this balance when the peptide substrates themselves bind to an inactivated protease. However, in calorimetry experiments, the substrates bind so weakly ($K_d \geq 5 \mu M$) to the inactivated protease that accurate measurement proved difficult. Two computational techniques were applied to suggest peptide sequences predicted to improve binding, starting with the tightest binding substrate sequence as a template. The first technique, charge optimization [23, 24], replaces the partial atomic charges on the peptide side chains with those that optimize the electrostatic component of the free energy of

binding. This procedure identified two peptide residues with suboptimal electrostatics, and in one case suggested a threonine-to-valine mutation that was predicted to improve binding. In addition, protein design techniques were applied to evaluate all single, double, and triple mutant peptide sequences for improved binding properties. This procedure recapitulated the mutation identified by charge optimization and suggested several other mutations. Three designed peptides were tested for binding using calorimetry, and the single threonine to valine mutation resulted in a greater than 10-fold improvement in binding. A triple mutant sequence predicted from protein design methods yielded a more modest 2–3 fold improvement.

Chapters 5 and 6 of this thesis detail methods to address another challenge in molecular design, which is the accurate scoring of molecular geometries to determine those that are predictive of favorable energetics or tight binding. Specifically, they propose an alternative way to solve the linearized Poisson–Boltzmann equation (LPBE) for the mixed discrete-continuum solvation model used to compute the electrostatic component of free energies throughout this thesis. The issue with current techniques is that, given the amount of computational resources currently available, their solution never seems to be fully converged. Therefore, when making predictions based on this popular solvation model, it is difficult to know whether errors are due to the model itself, or the inability to solve it accurately.

In order to address this problem, we have applied boundary-element methods (BEM) to solve the LPBE, which are capable of achieving higher accuracy than traditional finite-difference methods (FDM). The boundary-element method a popular technique for solving the partial differential equations that arise in electrostatics problems, especially in the electrical engineering community [47]. Finite-difference methods have fundamental inaccuracies when solving the molecular electrostatics problem because they represent molecular geometries and point charges using a rectilinear grid. In contrast, boundary-element methods using specialized curved panels

can exactly represent molecular geometries as well as the point charges that represent partial atomic charges, allowing for enhanced accuracy. However, boundary-element methods also have their limitations. One example is that the BEM generates a dense matrix equation that must be solved. The size of the dense system of equations grows quadratically in the discretization of molecular surfaces, which increases rapidly as larger systems are considered or the desired accuracy is increased. Consequently, many matrix compression techniques have been developed to allow this dense system to be stored approximately, reducing the time needed to construct and multiply by this operator [48–50]. However, one limitation of existing methods is that increasing the accuracy of the compression procedure results in large time and memory usage penalties. Chapter 5 presents the development of the FFTSVD matrix compression algorithm, which takes the best features of several existing boundary-element solvers and combines them into a single algorithm that is capable of achieving high accuracy with lesser computational expense.

Given the FFTSVD algorithm, and other essential work performed by collaborators, it was possible to overcome all of the practical challenges associated with the BEM when solving problems in continuum molecular electrostatics with the LPBE. The culmination of this work is presented in Chapter 6, where a complete boundary-element method is presented capable of solving biomolecular electrostatics problems to high precision. The presented curved BEM was tested against standard FDM techniques for solving the LPBE on several classes of continuum electrostatic calculations commonly performed in the field. When calculating the electrostatic component of solvation energies of molecules, absolute rigid binding free energies within a protein–protein complex, or non-rigid binding free energies, the curved BEM solution appeared to have significantly better convergence properties with increasing compute time expended, unlike FDM where the solution appeared to continually change. The only class of calculations for which FDM performed adequately was differential electro-

static binding energies between mutant and wild-type protein complexes that could exploit error cancellation due a large portion of the complex remaining unchanged. Overall, the development of this highly accurate LPBE solver confirmed the suspicion that solutions from the FDM may not be well converged, and suggests that more accurate solution methods may be required to make accurate predictions with the LPBE continuum model.

Chapter 2

Inverse Design of Small-molecule Ligands for Drug Discovery

Abstract

A new strategy for the design of small-molecule ligands for binding macromolecular targets is described. The new formulation is comparable to inverse protein design, where instead of predicting a protein's fold from its sequence, all sequences compatible with a given backbone fold are considered. Analogously, this method differs from docking and other traditional approaches that take molecules or molecular fragments and try to determine their bound structure within the target site. The inverse design method described here begins with a desired ligand shape along with computed binding potentials within, and constructs molecules that are energetically compatible with these features. The inverse design procedure described has a number of characteristics. It is combinatorial and discrete in that ligands are constructed by the joining of molecular fragments to scaffolds in individual rotameric and conformational states. It incorporates the guaranteed discrete search algorithms dead-end elimination and A*, which together provide the global optimum and an ordered list of all solutions higher in energy than the global optimum up to a chosen threshold. It is hierarchical in that the priorities of design candidates are initially ranked with a very rapid, physically based energy function, and the candidate list ordering is subsequently refined with increasingly accurate energy functions. Grid-based pre-calculations are used to speed energy computations, and the initial selection of the ligand shape allows solutions to the linearized Poisson-Boltzmann equation to be used to describe the electrostatic contributions to solvation and interaction from the earliest stages of the search. Because of its combinatorial and complete nature, this approach has the potential to evaluate the expected feasibility of combinatorial libraries for targeting a given binding site, as well as being useful in more traditional *de novo* design efforts and lead refinement.

2.1 Introduction

Automated exploration of the space of small molecules compatible with tight binding to a macromolecular target remains a formidable challenge in computational chemistry [51, 52]. The difficulties in developing such an algorithm are mainly two-fold. The first problem is the combinatorial complexity of molecular diversity, conformation, and orientation that must be effectively searched to identify molecules predicted to bind tightly as well as their predicted binding modes. Traditional *de novo* small molecule design algorithms fall into several categories depending on the way the search problem is handled. Many methods build ligands into a target site using sequential atom- or fragment-based growth, using strategies such as Monte Carlo, evolutionary algorithms, and tree search to select favorable attachments and orientations [53–61]. Other algorithms determine optimal placements for individual functional groups followed by techniques to link them together into full molecules [62, 63]. Several additional techniques allow atoms distributed throughout the site to merge and form real molecules under the influence of potential functions [64, 65]. Molecular docking algorithms have also been used to explore a binding site using pre-existing libraries of compounds [5, 7, 8, 66].

Alternatively, the field of protein design, with its recent successes [12, 15–21] has taken a different approach to solving the problem of combinatorial complexity in molecular design. This success can be attributed, in part, to the phrasing of protein design as an inverse problem [9, 10]. In inverse protein design, the backbone is constrained in a desired conformation, leaving behind a tractable search over discrete side chain identities and conformations to find sequences predicted to stabilize the fold [67–69]. Given a pairwise-additive scoring function, this smaller search space can be deterministically searched by algorithms such as dead-end elimination (DEE) [11, 14, 70] and A* [13] to identify global optimal and progressively sub-optimal energy structures. These deterministic search methods have been used in the small-

molecule docking problem to account for ligand or receptor flexibility [71–73], but they have not been previously applied, to our knowledge, in *de novo* ligand generation.

The second challenge in developing an automated small-molecule design procedure is a scoring function that can enrich the search space in molecules that are likely to bind tightly. Due to the vast search space involved in small-molecule design, traditional approaches often use computationally efficient, but approximate, energy models [3, 4, 74–81]. One aspect that is often neglected or highly approximated is solvation, due to the high computational cost of computing its effects accurately. Solvation has been shown to be important to correctly rank and predict small-molecule binders [82, 83], suggesting that it may be important to account for in ligand generation strategies. The strong geometry dependence of solvation makes it especially difficult to model during ligand design, where the shape of the molecules being considered changes as fragments are sequentially added.

In this report, we propose one possible implementation of an inverse algorithm to design tight binding small molecules that includes solvation as part of the scoring. This implementation has several important characteristics. Firstly, the method reduces the combinatorial complexity of small-molecule design by applying early constraints, in the same fashion as inverse protein design. A fixed shape is selected within the target binding site, which serves as a limit on the size and location of designed compounds, and also as the dielectric boundary for continuum solvation calculations. In addition, the search space of molecular diversity is described using a discrete set of molecular scaffolds and functional groups that can be attached. These scaffolds are placed discretely throughout the fixed shape, leaving behind a tractable combinatorial search over functional group space for each.

Secondly, the search over small-molecule diversity is approximated with a scaffold and functional group approach, where the search over functional group space is complete and deterministic. The conformations of functional groups grown from

each scaffold placement are sampled discretely, and the contribution to the binding score for any set of functional groups is pairwise additive. This allows existing deterministic algorithms such as dead-end elimination (DEE) [11, 14, 70] and A* [13] to generate a guaranteed rank-ordered list of high scoring molecules, topped by the global optimum. Thirdly, fast grid-based energy functions [74, 84–86] are employed to accelerate the pairwise binding score calculation. These include a pairwise solvation model, where screened interaction and desolvation potentials derived from Poisson–Boltzmann (PB) theory are recorded on a grid inside the fixed shape. Evaluation of a molecule against these grid potentials provides an approximate electrostatic binding free energy that assumes a fixed dielectric boundary. Lastly, to correct for grid-based energies and the approximate pairwise solvation model, the top scoring compounds across all scaffold placements are hierarchically re-ranked in more detailed energetic models that include explicit-atom calculations and full non-pairwise PB solvation models that take into account the correct molecular shape.

Given the complete and deterministic nature of this inverse ligand design strategy, it is well suited as a tool for exploring the space of molecules compatible with binding a target site. To demonstrate this, we applied the method to identify ligands compatible with small engineered model binding sites in the hydrophobic core of T4 lysozyme. The presence of a complete search space allowed for an energetic analysis of designed compounds, revealing that those most similar to known binders tend to have better electrostatic complementarity. Using the T4 lysozyme system as an example, we also demonstrate the importance of the approximate pairwise solvation model in designing molecules with the correct polarity for a binding site. We also applied the inverse method in a more traditional *de novo* design protocol to identify small molecules well suited for the *E. coli* chorismate mutase active site. Top scoring ligands reproduced the hydrogen bonding patterns and functional group usage of known inhibitors, even though they employed different scaffold geometries. Finally, the completeness and

determinism of the search protocol allows for a reasonably fair comparison between different combinatorial libraries targeting the same binding site. To this end, we used the method to evaluate the fitness of an example combinatorial library for binding the HIV-1 protease active site by comparing the compounds created from this library with those from a peptide-based reference design.

Overall, the inverse ligand design method presented here is applicable to a variety of molecular design problems including exploring the space of molecules compatible with a target site, *de novo* ligand generation, and combinatorial library evaluation. It has several advantages over existing methods due to its complete nature and accurate electrostatic modeling.

2.2 Theoretical and computational approach

2.2.1 Constraints applied for search feasibility

Our implementation of an inverse ligand design procedure, outlined in Figure 2-1, relies on several constraints and approximations in order to make the resulting search over molecular diversity feasible. The first constraint, which is also the first stage in the design procedure, is to select a target shape that limits the scope of the designed molecules, in that the atom centers of generated compounds must stay within its volume. A second important constraint in the inverse ligand design methodology is the chemical space searched. Molecular diversity is described using a scaffold–functional group approach; the space of ligands searched consists of a set of specified molecular cores (scaffolds) decorated with a given library of functional groups. This dovetails nicely with combinatorial library and parallel synthesis approaches to laboratory ligand construction, and can ensure that the search focuses on synthesizable molecules. Scaffolds are placed within the binding site in specified discrete conformations and orientations, and the functional groups grown from them are derived from generated

discrete conformational ensembles.

A further methodological constraint is the need for a pairwise additive function to describe the binding free energy. A hierarchical series of energy functions can be used to score and re-order the space of feasible binders, in which case only the base energy function needs to be pairwise additive. Discrete representation and pairwise energetics are essential for the search algorithms used to identify low-energy molecules in the combinatorial space.

The overall goal of these constraints is to phrase small-molecule design in a manner similar to inverse protein design, with the specific aim of adopting the combinatorial search algorithms than have been developed to solve protein design problems. With these constraints applied up front, it is possible to implement an inverse small-molecule design strategy as explained below.

2.2.2 Target shape selection

In addition to serving as a constraint on ligand size, the target shape also serves as an approximation to the final shape of the ligand throughout the combinatorial search over molecular space. In continuum models, the electrostatic component of binding is highly dependent on the shape of the ligand and the receptor, such that shape changes in one part of the ligand can affect the electrostatic contributions of other regions. These effects are captured directly in continuum solvation models but require significant sampling of solvent degrees of freedom to capture in explicit solvent models. As a result, small-molecule design approaches that place ligand groups progressively can lead to sub-optimal or even poor solutions, as addition of functional groups can change the electrostatic contributions of those already placed. By selecting a target ligand shape and assuming that the molecule occupies all of this volume throughout the search procedure, this complication is avoided as well as allowing for fast approximations of solvation energies as described later.

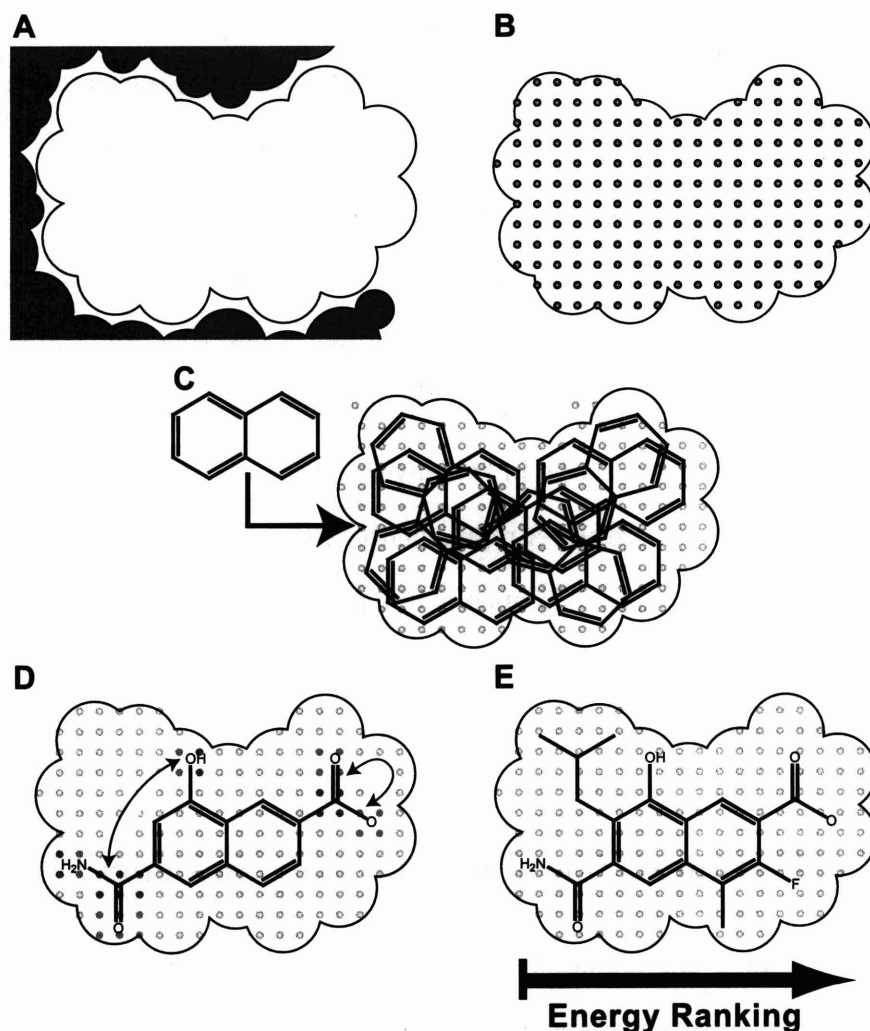


Figure 2-1: Overview of the inverse ligand design algorithm. First, a ligand envelope is created in the active site (A), which will serve as a hard constraint on the size and shape of designed ligands as well as an approximation to the true molecular shape in solvation calculations. Next, grid-based energy functions are computed on cubic lattices within the shape (B) for van der Waals packing, electrostatic desolvation upon binding, and screened electrostatic interaction. Then, a discrete conformational ensemble of molecular scaffolds is discretely sampled throughout the site in a variety of orientations (C). For each scaffold placement, self and pairwise contributions to the binding energy are computed for discrete functional group attachments using the grid-based energies (D). Finally, fast and guaranteed combinatorial search algorithms determine the global minimal binding energy configuration of functional groups for each scaffold placement (E), as well as a ranked list of configurations with increasing energy. The best scoring compounds can then be pooled across all scaffold placements and re-ranked in increasingly sophisticated binding energy models.

The target shape also represents an idealized version of the shape for a completed ligand, and the objective function for design attempts to fill as much of this volume as possible. Given this role, it makes sense to select a target shape that would predict ligands of maximal affinity or encourage interaction with certain parts of the site of interest. In most physical models of molecular interaction, the hydrophobic (shape) contribution to binding has been shown to correlate with burial of surface area [87], implying that the optimal target shape for binding is one that fills the binding site completely. Therefore, a negative image of the binding site, approximated by a sphere packing, is used as the target shape in this inverse design procedure. In buried active sites, it suffices to pack spheres on a regular cubic lattice and accept any sphere that does not clash with the receptor. In open active sites, spheres can be placed and their geometry minimized or optimized such that they pack desired regions of the binding interface without extending too far into solvent.

2.2.3 Calculation of binding potentials and energies

Once the target shape is selected, the goal of the combinatorial search procedure is to design molecules into this shape that interact favorably with the receptor. Rather than directly scoring the interaction between a molecule in the shape and the receptor, we employ grid-based energy functions [74, 84–86] and pre-compute binding energies or potentials on a regular cubic lattice inside the target shape. Grids can then be interpolated onto molecules or molecules projected onto the grids in order to compute binding energies.

The scoring function employed in the combinatorial search procedure contains three primary components, a van der Waals packing term, a screened electrostatic interaction term, and desolvation penalties for both the designed ligand and the receptor. Grids for van der Waals energies are computed by placing a particular parameterized atom type at each grid point and computing its van der Waals interaction

energy with the rest of the receptor. This is repeated for every atom type. To derive the van der Waals binding energy for a given molecule, the energetic contribution of each atom is calculated by trilinearly interpolating energies from the surrounding eight points of the appropriate grid.

Given that the target shape is fixed throughout the combinatorial search portion of the algorithm, calculation of grid-based potentials useful in the evaluation of electrostatic interaction and desolvation is straightforward using the linearized Poisson–Boltzmann equation (LPBE), and follows directly from charge optimization methods [23, 24, 30, 31]. As shown previously for the LPBE [24, 31], the electrostatic component of the binding free energy given a fixed shape for the bound and unbound states, fixed charges on the receptor, and a set of basis points within the ligand can be written as

$$\Delta G_{\text{elec}} = \vec{q}^T L \vec{q} + \vec{q}^T \vec{C} + R_{\text{des}} \quad (2.1)$$

where \vec{q} is a vector whose elements give the charge values at ligand basis points, L is the ligand desolvation matrix, \vec{C} is the screened interaction vector, and R_{des} is the receptor desolvation penalty (a scalar). The ligand desolvation matrix is computed as the difference between two matrices, $S_{\text{bound}} - S_{\text{unbound}}$, where each is a solvation energy matrix such that $\vec{q}^T S \vec{q}$ equals the solvation energy of the ligand in either the bound or unbound state. The symmetric S matrices are computed by charging each ligand basis point to a value of $1.0e$ with all other ligand points and receptor atoms set to $0.0e$ and solving the LPBE in the specified geometry. The resulting potentials at all of the basis points become one column of the S matrix. The interaction vector \vec{C} is derived similarly, except only the bound state is considered and receptor atoms are charged to their parameterized values. The screened electrostatic potential projected by the receptor charges at each ligand grid point becomes the elements of C . The receptor desolvation penalty, R_{des} , is computed by taking the difference between two states, where the receptor is charged and the ligand, uncharged, is either bound or unbound.

Because the target ligand shape is fixed, this number is a constant throughout the combinatorial search.

If the basis points within the ligand are set to a regular cubic lattice, the electrostatic binding energy of any molecule can be approximated by trilinearly projecting each partial atomic charge to the grid points and applying Equation 2.1. This approximation estimates the electrostatic binding free energy of the molecule within the target shape, rather than the correct molecular surface of the ligand derived from radius parameters. Keeping the target shape constant allows for precomputation of L , \vec{C} , and R_{des} and is the basis for the fast grid-based electrostatics and solvation approximation.

One important identity of Equation 2.1 is that two superimposed grid charge distributions, \vec{q}_1 and \vec{q}_2 can be distributed as shown in Equation 2.2 for the case of a symmetric L matrix.

$$\Delta G_{\text{elec}, \vec{q}_1 + \vec{q}_2} = \vec{q}_1^T L \vec{q}_1 + \vec{q}_2^T L \vec{q}_2 + 2\vec{q}_1^T L \vec{q}_2 + \vec{q}_1^T C + \vec{q}_2^T C + R_{\text{des}} \quad (2.2)$$

This distribution of terms can be extended to any number of superimposed charge distributions and provides the basis for a pairwise decomposition of electrostatic and solvation energies as subsequently described.

In addition to van der Waals and electrostatic solvation terms, additional components of the score during the combinatorial search include a bump check against the target shape, ensuring that any molecule outside it has an infinite energy, as well as functional group–scaffold and functional group–functional group bump checks to ensure that designed molecules are not self-intersecting. Bump checks were chosen over traditional molecular mechanics internal energies because it is unclear when designing a small molecule how much internal strain is paid upon synthesis rather than binding.

2.2.4 Pairwise decomposition of the scoring function

In order to use existing implementations of combinatorial search algorithms such as DEE and A*, the scoring function must be pairwise decomposable in functional group conformation and identity. This means that the total energy of a given scaffold with added functional groups must satisfy

$$\Delta G_{\text{binding}} = E_{\text{const}} + \sum_{i=1}^n E_{\text{self}_i} + \frac{1}{2} \sum_{i=1}^n \sum_{j=1, j \neq i}^n E_{\text{pair}_{i,j}} \quad (2.3)$$

where n is the number of attached functional groups in discrete geometries. This expression consists of a constant term (E_{const}), a sum over the self energy contributions of each functional group alone (E_{self_i}), and a sum over contributions for each pair of functional groups ($E_{\text{pair}_{i,j}}$). When functional groups are attached to the scaffold, a hydrogen on both the scaffold and functional group are removed and a bond is formed between the two antecedent atoms. Therefore, the blunt scaffold, with growable hydrogens removed, is invariant in a particular placement of the scaffold and search over functional groups, contributing only to the constant term in Equation 2.3. This constant term includes the grid-based van der Waals and electrostatic binding energy of the blunt scaffold, as well as the constant receptor desolvation penalty and can be written as

$$E_{\text{const}} = \text{vdW}_{\text{scaffold,blunt}} + \vec{q}_{\text{scaffold,blunt}} \text{ }^T \vec{C} + \vec{q}_{\text{scaffold,blunt}} \text{ }^T L \vec{q}_{\text{scaffold,blunt}} + R_{\text{des}} \quad (2.4)$$

where $\vec{q}_{\text{scaffold,blunt}}$ are the grid projected partial atomic charges of the blunt scaffold with L , \vec{C} , and R_{des} derived from the electrostatic approximation.

The self energy for each attached functional group in a discrete conformation is the grid-based van der Waals contribution for all atoms except the hydrogen removed when attached to the scaffold, and the electrostatic energy is computed by placing

the charge from the removed scaffold hydrogen on the functional group antecedent atom, and the charge on the removed functional group hydrogen on a dummy atom at the position of the scaffold antecedent atom to ensure charge conservation after attachment. The self electrostatic binding energy of the functional group can be computed through grid projection and application of Equation 2.1. In addition to electrostatic interaction and self desolvation, the indirect desolvation between the functional group and the scaffold must also be added to E_{self_i} , which can be computed from the cross terms in the symmetric desolvation matrix, as shown in Equation 2.2. Therefore, a final expression for E_{self_i} is

$$E_{\text{self}_i} = \text{vdW}_{\text{fg,blunt}} + \vec{q}_{\text{fg,blunt}}^\top \vec{C} + \vec{q}_{\text{fg,blunt}}^\top L \vec{q}_{\text{fg,blunt}} + 2\vec{q}_{\text{fg,blunt}}^\top L \vec{q}_{\text{scaffold,blunt}} \quad (2.5)$$

where $\vec{q}_{\text{fg,blunt}}$ are the projected grid charges for the charge-swapped functional group, $\vec{q}_{\text{scaffold,blunt}}$ are the projected grid charges for the scaffold with growable hydrogens removed, and L , C , and R_{des} are derived from the electrostatic approximation. Any functional group geometry that fails a bump check with the scaffold or with the shape is removed from further consideration in the combinatorial search. Optionally, a discrete functional group conformation can also be removed from the space if it fails certain chemical criteria, such as making specified hydrogen bonding interactions.

The contribution of a pair of functional groups to the binding free energy only contains the indirect solvation effects between them, as shown in Equation 2.6. If the two functional groups fail a bump check and clash, their pair energy contribution is infinite.

$$E_{\text{pair}_{i,j}} = \begin{cases} \infty, & \text{if functional groups bump} \\ 2\vec{q}_{\text{fg}_i,\text{blunt}}^\top L \vec{q}_{\text{fg}_j,\text{blunt}} & \text{otherwise} \end{cases} \quad (2.6)$$

The pairwise energy decomposition presented above is crafted such that Equa-

tion 2.3 sums to same energy as if the entire molecule, complete with scaffold and functional groups, is evaluated with the grid-based van der Waals and electrostatics/solvation functions.

2.2.5 Scaffold placement

Once the target shape has been selected and precomputed grids generated within for van der Waals and electrostatics/solvation scoring, the next step in the inverse design procedure is to place the scaffold throughout the target shape in discrete conformations and orientations. These scaffold placements serve as the launching point for functional group attachment, and serve a similar role to a set of fixed backbone geometries used in inverse protein design calculations. Finding optimal scaffold placements is similar to the problem of molecular docking, except that the ligand molecule is incomplete. As a result, finding the lowest energy docked structures for the scaffold may not be optimal, because these placements may limit functional group attachment. To solve this problem, we exhaustively sample scaffold orientations for a pre-generated conformational ensemble through uniform sampling of translational and rotational space. Any scaffold orientation that is within the target shape and passes a grid-based van der Waals energy cutoff is accepted. Another possibility for generating scaffold positions might involve docking the scaffold with minimal or representative functional groups attached and keeping only the scaffold placement.

2.2.6 Combinatorial search over functional groups

For each scaffold placed within the target shape, a complete search is performed over the space of possible functional group identities and conformations at each growable hydrogen position. Pairwise energies are computed as described above, and the dead-end elimination (DEE) [11, 14, 70] and A* [13] algorithms are used to identify the global minimum functional group structure as well as a ranked list upwards in energy.

Operationally, the ranked list is extended until a certain number of unique molecules has been discovered and up to a given number of conformations are retained for each. The results of the combinatorial search across all scaffold placements are then pooled and sorted in energy score in order to determine the best structures as predicted by this approximate scoring function.

2.2.7 Hierarchical rescoring of top structures

The pairwise energy function presented above has several deficiencies that can be corrected by successively passing the highest scoring molecules to better and better energy functions in a hierarchical fashion. The most obvious of these deficiencies is the approximate pairwise solvation model because it assumes that all ligands have the same molecular shape. To initially correct for this approximation, the highest scoring molecules from the combinatorial search can be re-evaluated with fast but non-pairwise approximations to converged linearized Poisson–Boltzmann calculations that incorporate more knowledge of the actual molecular shape. To this end, we have implemented three approximate methods that meet these criteria and are applicable to certain classes of design problems.

The first of these “medium resolution” methods is suited for design problems where the target site is deeply buried from solvent, and the designed molecules fill the target shape reasonably well. In this case, the solvation and interaction potentials in the bound state computed from the grid-based fixed-shape assumption should agree very well with those computed with the LPBE because the shape of the bound state is mostly invariant to the ligand geometry. However, in the unbound state, the target shape may be a poorer representation of the ligand geometry and a correction is required. The advantage here is that the unbound state with the ligand alone is much smaller than the bound state and requires only a relatively fast LPBE calculation to obtain a converged answer. This procedure is summarized in Equation 2.7, which

is analogous to Equation 2.1 except that the ligand desolvation matrix is broken up into its two solvation matrix components and the unbound state solvation energy is computed directly with LPBE calculations rather than the grid-based approximation.

$$\Delta G_{\text{elec}} = (\vec{q}^T S_{\text{bound}} \vec{q} - \Delta G_{\text{solv,unbound,LPBE}}) + \vec{q}^T \vec{C} + R_{\text{des}} \quad (2.7)$$

A second approximation to the full LPBE electrostatic binding calculation, which is more generally applicable, is to apply a fast method that approximates the change in solvation energy when the shape of the ligand shrinks from the target shape to the true molecular surface derived from parameterized radii. One such method, first proposed by Arora and Bashford [88] and later extended [89], involves integrating the energy density in the volume that changes from being inside the target shape to inside solvent assuming that the electric field is the same as computed from the target shape geometry. This is shown in Equation 2.8

$$\Delta \Delta G_{\text{solv}} \approx \frac{\epsilon_{\text{mol}}}{\epsilon_{\text{solv}}} \int \frac{\epsilon_{\text{mol}} - \epsilon_{\text{solv}}}{8\pi} \|\vec{E}\|^2 dV \quad (2.8)$$

where ϵ_{mol} and ϵ_{solv} are the molecular and solvent dielectric constants, \vec{E} is the electric field computed when the target shape is present, and the volume integral is taken over regions that were in the target shape but became solvent when the actual ligand radii were used to generate the dielectric boundary.

With this formalism, we can correct each desolvation component of the electrostatic binding energy. The ligand desolvation upon binding is corrected by taking the difference between evaluation of the integral in both the bound and unbound states. Electric fields are calculated through finite-difference of the grid-based solvation potentials, derived from multiplying the projected ligand charges by the appropriate solvation matrix S . Integration is carried out by assuming the electric field is constant within a grid cube and summing the product of the square of the electric field

magnitude with the volume of each cube in the differential region. Receptor desolvation is similarly corrected through integration using the electric field derived from the interaction potential \vec{C} .

One final method, which is slower but should provide a reasonable approximation in most cases, is simply to run a full electrostatic binding calculation by solving the LPBE at very low discretization, such that the solve time is fast but the answer still correlates with a fully converged calculation.

The top ranking structures after re-evaluation with fast corrections to the fixed shape solvation approximation, which should be few in number, are then subjected to further hierarchical improvements to the energy function that take longer amounts of time to compute. These include a converged LPBE electrostatic binding calculation using the correct molecular surface for both the bound and unbound states and the addition of a surface area term to model the hydrophobic contribution to solvation. Additional calculations that could be performed at this stage include computing ligand and internal strain through modeling of the unbound state, assessing the error in the atomic charge swapping procedure used for functional group attachment by repeating partial atomic charge determination on the entire compound and recomputing binding energetics, or consideration of ligand conformational entropy losses upon binding.

At each stage of hierarchical re-evaluation, enough structures are considered such that correlation between the more and less sophisticated energy model is observed. With enough correlation, one eventually reaches a point where additional top ranking structures from the lesser model are unlikely to score better than a certain threshold in the more accurate energy function. In this manner, we attempt to statistically guarantee that the top molecules in our best energy function have been identified even though we are unable to perform direct guaranteed searches on the best energy function due to non-pairwise effects.

2.3 Results and Discussion

2.3.1 Validation with engineered model binding sites in T4 lysozyme

To understand the operation of the inverse small molecule design method, we applied the procedure to two model binding sites previously engineered into the hydrophobic core of T4 lysozyme by Matthews and co-workers [90–92]. The first site contains a single L99A mutation [90,91], which opens a hydrophobic cavity that can be bound by hydrophobic small molecules [91,93]. The second site adds the additional M102Q mutation [92], which introduces a single hydrogen-bond acceptor into the cavity and causes it to bind polar molecules as well as hydrophobics. These sites are well suited for validation and testing because they are feasible to sample thoroughly due to their small size, have lists of experimentally known binders and non-binders, and although artificial, seem to present many of the same computational challenges found in natural binding sites [92,94].

The space of potential ligands for this study was constructed combinatorially from previous database screening efforts by Murcko and co-workers [95,96] that identified the scaffolds and functional groups most commonly used in the Comprehensive Medicinal Chemistry (CMC) database [97]. Conformational ensembles were generated for these scaffolds, followed by uniform sampling in discrete positions and orientations throughout the target shape. The space of functional group attachments was combinatorially searched for each scaffold placement, and the best structures were identified. The total number of structures searched was 3.2×10^{10} for the L99A hydrophobic site and 1.1×10^9 for the L99A/M102Q polar site. For each of the two model binding sites, structures in the top 10 kcal/mol of the combinatorial search were re-evaluated by correcting the unbound state contribution to the solvation energy with LPBE calculations (because the sites are deeply buried), and by switching from grid-based to

Table 2.1: Combinatorial search space sizes for inverse designs

Target Site	Scaffold Library	Functional Library	Search Space
L99A T4 lysozyme	CMC derived ^a	CMC derived ^b	3.2×10^{10}
L99A/M102Q T4 lysozyme	CMC derived ^a	CMC derived ^b	1.1×10^9
<i>E. coli</i> chorismate mutase	CMC derived ^a	CMC derived ^b	1.5×10^{19}
<i>E. coli</i> chorismate mutase	Figure 2-7	CMC derived ^b	2.5×10^{19}
HIV-1 protease	Bignelli-based (Figure 2-10A)	Figure 2-10B,2-10C	3.7×10^{12}
HIV-1 protease	Tetrapeptide	Natural L-Amino Acids	1.2×10^{16}

^a Reference [95]

^b Reference [96]

^c Reference [98]

^d Reference [99]

explicit-atom van der Waals binding energies.

The correlation between the grid-based energy function used in the combinatorial search and the first set of “medium resolution” corrections is shown in Figure 2-2A for the L99A hydrophobic site. Although there is some degree of spread, mainly due to the target-shape based electrostatics approximation, there is significant correlation. We chose to draw a horizontal bounding line 3 kcal/mol above the best corrected score; it is likely that nearly all structures with a corrected score better than this threshold have been identified due to the correlation between uncorrected and corrected energies. In this manner, we can be reasonably confident that the best molecules in this non-pairwise correction have been identified. While it is not possible to provide absolute guarantees that no low-energy structures with one energy function are missed by sorting with a different function, statistical confidence limits can be developed but are beyond the scope of the current work.

These top structures were then further re-ranked using full LPBE binding calculations and a surface-area dependent hydrophobic solvation term. The relationship between the first set of corrections and the full energy model is shown in Figure 2-2B, which indicates a strong correlation and high confidence that few, if any, low-energy structures are missing.

One of the advantages of complete search is that the large space of generated ligand structures can be analyzed in terms of the trade-offs between portions of the

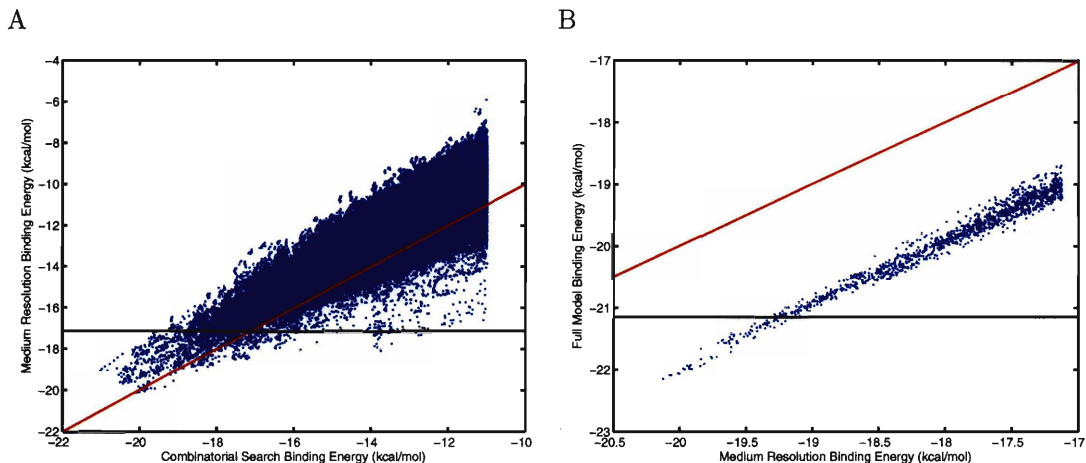


Figure 2-2: Correlation between energy models in different stages of the inverse design procedure. The approximate energy model used in combinatorial search shows good correlation with the first round of energy function correction (A), allowing a bounding line (black) to be drawn. It is unlikely that many, if any, additional structures re-ranked by the first round of corrections will score better than this threshold. The same is true when the first round of corrections is compared to the full energy model (B). The red lines indicate $y = x$. The correlation in (B) is offset from $y = x$ due to the addition of a favorable hydrophobic solvation term in the full energy model. Lower energies indicate a more favorable binding score.

scoring function. Figure 2-3A contains the top 10 kcal/mol of combinatorially generated structures from the L99A hydrophobic site plotted as a function of their van der Waals and electrostatic contribution to the binding free energy after the first round of “medium resolution” energy function correction. The results in the Figure reveal trends in the design space. Firstly, there is an observed lower bound to the electrostatic binding free energy at approximately +1.5 kcal/mol. Even the most electrostatically favorable ligands in the space are unable to fully recover the electrostatic desolvation of the ligand and receptor through intramolecular interactions in the complex.

A second feature is that in the regime of the best total binding energies, there is a direct trade-off between van der Waals packing and electrostatics/solvation, in that no ligands exist in the search space that simultaneously make the best packing and best electrostatic interactions. Possible explanations for this finding include

an inherent incompatibility with simultaneously optimizing these terms, or possibly that the functional groups selected for the library are more optimal for packing than electrostatic interactions. To test the second idea, the search procedure was repeated using a functional group library that contained aliphatic or vinylic versions of all polar groups, which could be more electrostatically optimal for the hydrophobic site but still capable of making the same packing interactions. However, the same trend was observed, indicating that polar groups may be selected for their packing interactions even though they pay additional desolvation penalties.

A final feature is that when ranking compounds on the sum of the two energy terms (indicated by the black line with slope -1 and perpendicular arrows) the first compounds selected are more optimized for packing rather than electrostatic interactions. These compounds have excellent shape complementarity to the site (Figure 2-3B) but pay desolvation penalties due to their use of polar functional groups. These molecules differ strongly from those known to bind this nonpolar site, and may indicate deficiencies in the energy function components or their weighting.

A comparison of computed binding energy contributions with known experimental binding results can reveal their relative reliability. Figure 2-3A shows generated molecules known to bind L99A T4 lysozyme in green, while generated molecules known not to bind are in red. The results show a rather strong relationship between experimental binding and computed electrostatic binding affinity, and a weak or non-existent correlation with computed van der Waal's binding affinity. This could indicate a fundamental inaccuracy in the treatment of van der Waals and packing interactions, as the use of minimization or reduced van der Waal's radii [100] do not improve the correlation (data not shown). Combined with previous findings that natural ligands are electrostatically optimized [31, 101, 102], these results could suggest that tight binding ligands may be those that are nearly electrostatically optimized while maintaining a reasonable packing score. While a full analysis of these effects

is beyond the scope of the current work, we note that the current energy model, which is in common use, tends to reward packing interactions of buried unsatisfied polar groups beyond that of buried hydrophobic groups. In Figure 2-3A, there are several green points (binders) that have similar electrostatic scores as red points (non-binders). These molecules are modestly polar ligands that happen to bind the nonpolar site, such as hexafluorobenzene and anisole [91], and are computed to have suboptimal electrostatic complementarity because they cannot make direct polar interactions upon binding.

Results for exploring ligand space in the polar L99A/M102Q site were similar to that of the L99A nonpolar site (Figure 2-4). An electrostatic lower bound was observed, as well as the same trade-offs between electrostatics and packing in the structures with the best total score. Once again, compounds known to bind the polar site were heavily biased towards better electrostatic than packing scores. Due to the presence of a single hydrogen bond acceptor, the ligands with near optimal electrostatics and good packing were not limited to hydrophobics, but also included molecules with single hydrogen-bond donors (Figure 2-5). Many of these designed molecules are similar to those known to bind and recapitulate the single hydrogen bond observed in their crystal structures [92].

Overall, validation in these model binding sites proved useful in identifying the strengths and weaknesses in the search procedure and scoring model. Even though the pairwise energy model used in the combinatorial search is very approximate, it has enough correlation with non-pairwise corrections and the full-energy model such that identifying the highest ranked compounds can be statistically ensured. In addition, the combinatorial search is complete and deterministic, which allows for identification of the regions in the scoring function enriched in molecules likely to bind the target site. Even though the total energy may be biased to favor the use of polar groups to make packing interactions, structures closer to electrostatic optimality bear more

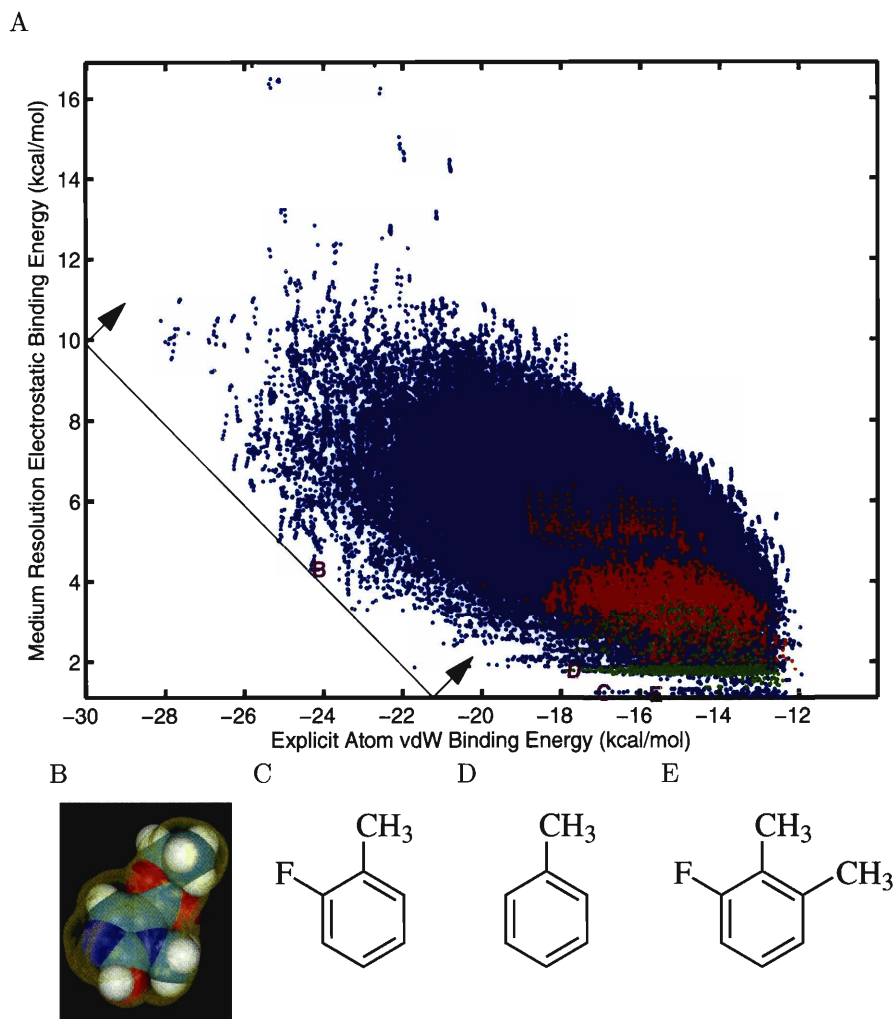


Figure 2-3: Energetic decomposition of ligand binding energies within 10 kcal/mol of the highest ranking structure from combinatorial search in the L99A T4 lysozyme site. Comparison of the explicit-atom van der Waals score and initial shape-corrected electrostatics score for each structure shows tradeoffs in the high scoring region (A). Green points are compounds known to bind the site, red points are designed compounds known not to bind, and blue points are other designed compounds. Green and red points were plotted in an alternating fashion to avoid hiding data. The black line with arrows is a line with slope -1 passing through the structure with the best total energy. As the line moves in the direction of the arrows, the top compounds in total energy sum are recovered. The structures of selected designed compounds are shown in (B–E). Their location in the energy decomposition (A) is denoted by letters. In (B), the designed compound (atom colors, vdW representation) is shown inside the target shape (transparent yellow). Toluene (D) is a known binder of the L99A T4 lysozyme site [91].

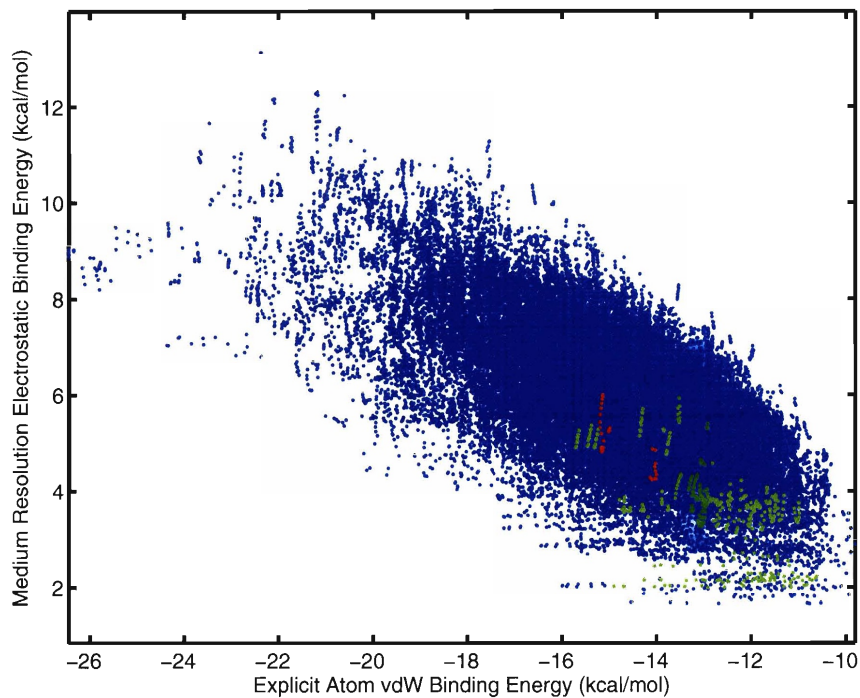


Figure 2-4: Energetic decomposition of ligand binding energies within 10 kcal/mol of the highest ranking structure from combinatorial search in the L99A/M102Q T4 lysozyme site. Green points are compounds known to bind the site, red points are designed compounds known not to bind, and blue points are other designed compounds. Green and red points were plotted in an alternating fashion to avoid hiding data.

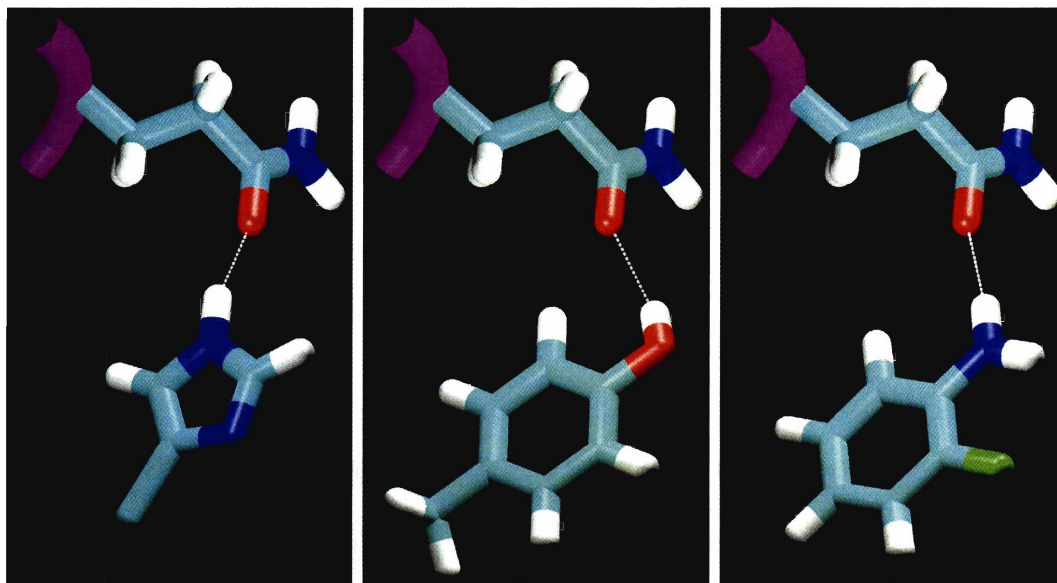


Figure 2-5: Examples of designed molecules with near-optimal electrostatics in the L99A/M102Q T4 lysozyme site. Each molecule makes a single hydrogen bond to the carbonyl of the glutamine 102 side chain in a similar fashion to structures of known binders. The extra light blue atom in the first panel is chlorine and the green atom in the last panel is fluorine.

similarity to known binders, providing a guideline for identifying such molecules in subsequent analysis.

2.3.2 Importance of including desolvation effects in the combinatorial search

In order to measure the importance of including the approximate solvation model in the search procedure, design in the polar L99A/M102Q site was repeated except that all entries of the ligand desolvation matrix L were set to zero, leaving behind only the screened interaction energy to model electrostatics, as is common in many scoring functions. After combinatorial search, structures in the top 10 kcal/mol, designed with and without desolvation in the energy function, were re-evaluated for rigid binding using a full LPBE electrostatics model with correct molecular shapes. For reference, the correlation between the approximate grid-based electrostatics model including desolvation and the full electrostatics model is shown in Figure 2-6A. In contrast, the correlation between the grid-based electrostatic score lacking desolvation and the full LPBE model is greatly reduced, as can be seen in Figure 2-6B. Without a desolvation penalty, the combinatorial search strongly favors positively charged amino groups to interact with the side-chain carbonyl of E102. In fact, the generation of ligands with $+2e$ and $+3e$ charges was common, corresponding to the horizontal stripes in Figure 2-6B, where desolvation penalty is dominated by the net charge. These results show that even an approximate solvation model, such as the pairwise grid-based fixed-shape assumption, can be useful in the generation of compounds with appropriate polarity for the binding site.

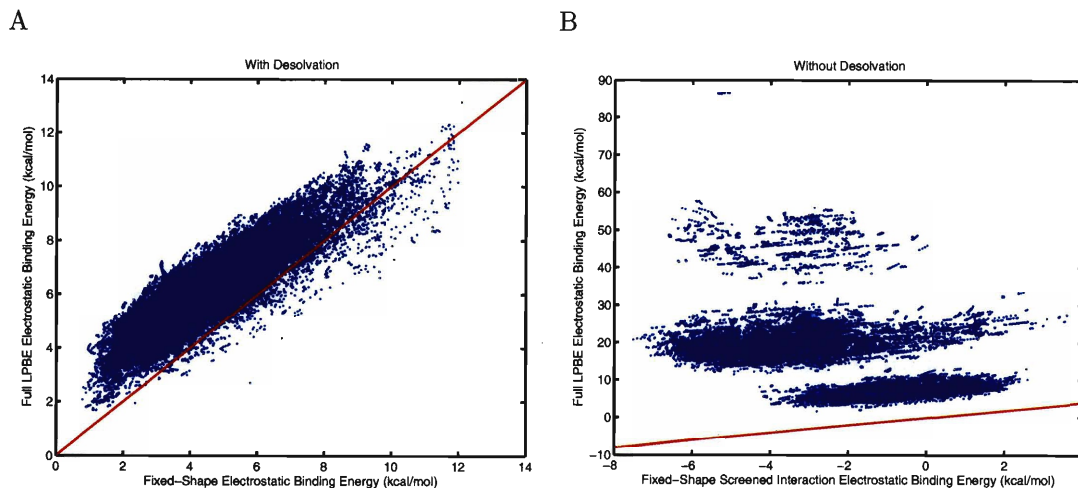


Figure 2-6: The effect of ignoring desolvation penalties in the combinatorial search of the L99A/M102Q T4 lysozyme site. The correlation between the pairwise electrostatic model including a desolvation approximation and the full electrostatics model (A) is superior to the correlation obtained when only using screened interactions to model electrostatics during the combinatorial search (B).

2.3.3 Exploring the space of molecules complementary to the *E. coli* chorismate mutase binding site

In order to further validate and test the inverse ligand design procedure, we applied the algorithm to explore the space of molecules complementary to the binding site of *E. coli* chorismate mutase. This target site is interesting for several reasons. Firstly, the active site is highly positively charged and binds highly negatively charged substrates and ligands [98, 103], which implies that modeling solvation accurately is important for this site. Secondly, this target is pharmaceutically relevant because chorismate mutases are essential enzymes for plants, bacteria, and fungi, but not found in higher organisms [104–106]. Lastly, there are relatively few classes of molecules known to bind this site, and they tend to only have moderate affinity [98, 107–111], indicating that exploration of new ways to generate site complementarity may be useful.

To initially target *E. coli* chorismate mutase, the most common scaffolds in the CMC database [95] were discretely sampled throughout the sphere-packed site, result-

ing in approximately 90,000 placements. Functional groups from the CMC database analysis [96] were grown in discrete conformations from any hydrogen on each of these scaffold placements, leading to a total search space of 1.5×10^{19} structures (Table 2.1) after eliminating conformations that clash with the shape or the scaffold. Once again, the top 10 kcal/mol of structures from the combinatorial search were passed to a first round of “medium resolution” energy function correction including explicit-atom van der Waals interactions and correcting for solvation in the unbound state. As before, the corrected binding energies were split into their van der Waals packing and electrostatic/solvation components. Tradeoffs were seen between packing and electrostatic contributions to the binding energy, and a set of 1000 structures that ranked the best in electrostatics while still maintaining good packing scores was examined further.

Within this set, about 75% used one of two scaffolds from the library, namely cyclohexane and tetrahydropyran. The prevalence of these scaffolds is interesting given that known inhibitors, mostly bicyclic transition state analogs presenting two carboxylate groups, contain such ring structures [98,107–109]. These generated molecules had either a $-1e$ or $-2e$ net charge, and interacted with the site primarily through carboxylate or nitro groups making either one or two electrostatic contacts with active-site arginines (Figure 2-8). This interaction pattern is similar to the known binding mode of an oxabicyclic diacid transition state analog [103] (Figure 2-8A).

With the top compounds utilizing only a small number of scaffolds, it might be possible to improve scores and attain higher diversity by expanding the scaffold library to include additional structures. A previous study had identified several ring structures likely to be well suited for binding chorismate mutases [98], and the combinatorial search procedure was repeated using these scaffolds as well as alternative versions in which some methylene groups were replaced by ether oxygens (Figure 2-7). With this new scaffold library, the combinatorial search complexity remained about the same (Table 2.1) while first-round corrected scores of the highest ranked

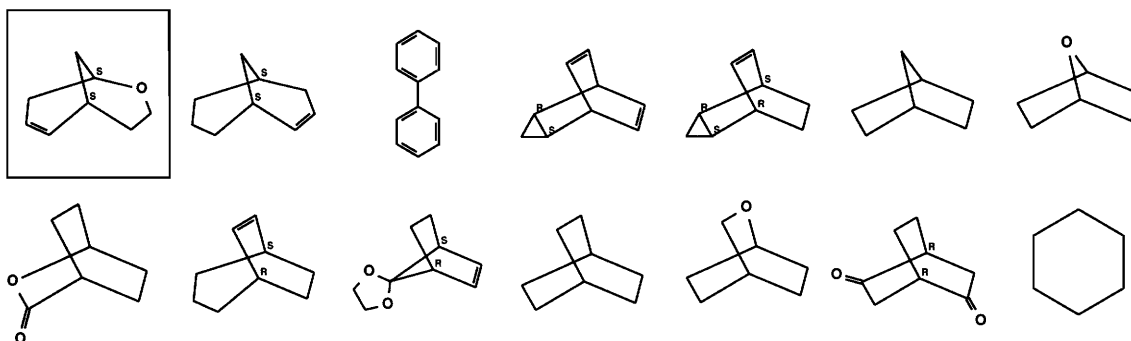


Figure 2-7: Extended scaffold library used for design in the *E. coli* chorismate mutase active site. Scaffolds shown here are derived from Husain *et al.* [98], scheme 1 and 2. The boxed scaffold is taken from a known transition state analog inhibitor [107,108].

compounds improved by almost 2 kcal/mol due to a combination of improved packing and electrostatics. Although these scaffolds are different from those derived from the CMC, the mode of interaction for generated compounds was extremely similar, primarily using carboxylates or nitro groups to make one or two bidentate electrostatic contacts with active site arginines (Figure 2-8).

In order to quantify the prevalence of electrostatic contacts in designed compounds, the hydrogen bonding geometries for carboxylate and nitro groups used to interact with active-site arginines were compared to the geometry observed in the crystal structure of a transition state analog inhibitor bound to *E. coli* chorismate mutase [103]. The comparison was made by computing the root-mean-square deviation (RMSD) between the designed and experimentally observed groups. For the approximately 32,000 structures within the top 10 kcal/mol of the combinatorial search, 99% made at least one carboxylate/nitro – active-site arginine interaction that was within 1.5 Å RMSD of an observed geometry. In addition, 21% of the designed compounds made two such interactions. As shown in Figure 2-9, the distribution of designed carboxylate/nitro group RMSD as compared to the crystal structure is heavily biased towards lower values, demonstrating how the inverse design methodology identified this structural determinant of binding.

One of the scaffolds in the expanded library (Figure 2-7, box) is derived from

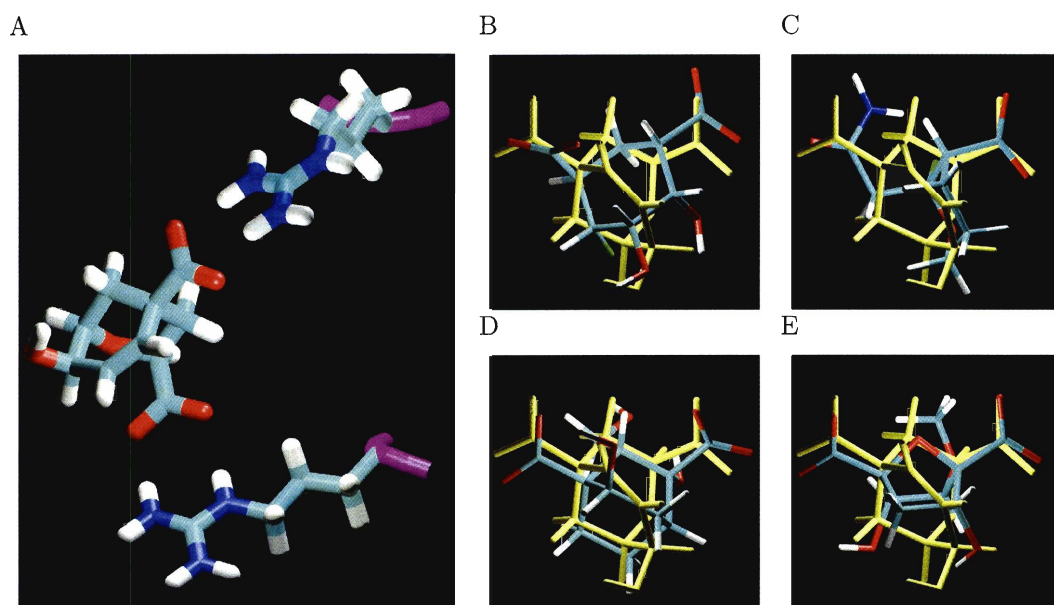


Figure 2-8: Results of exploring the space of small molecules compatible with binding the *E. coli* chorismate mutase active site. The crystal structure of a known transition state analog (TSA) inhibitor (A, left center) interacts with the active site primarily through bidentate electrostatic contacts between its carboxylates and the side chains of Arg11 and Arg28. Molecules designed using scaffolds derived from the CMC database by Bemis and Murcko [95] (B–C, atom colors) and an extended scaffold library derived from Husain *et al.* [98] (Figure 2-7) (D–E, atom colors) replicate these interactions, as compared to the TSA in yellow.

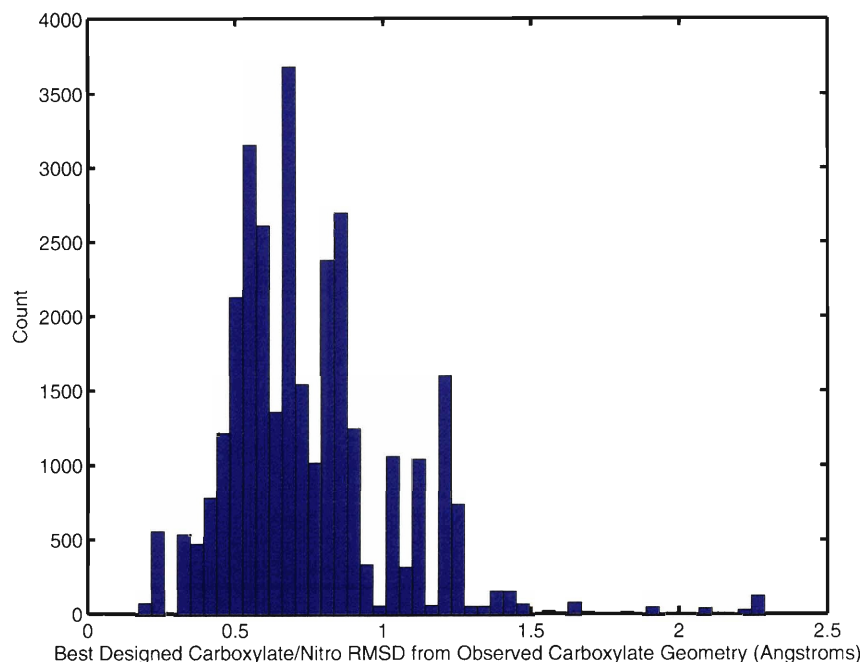


Figure 2-9: Histogram of the RMSD between designed carboxylate or nitro groups and carboxylates found in a crystal structure of a known transition state analog. If the designed compound contained more than one such group, the lowest RMSD value was used.

a known transition state analog inhibitor [107, 108] (Figure 2-8A). Although the design algorithm did not reproduce the exact functional group arrangement on this inhibitor given the settings used for design, it did reproduce the inhibitor with one additional hydroxyl group. This structure was ranked 18,837 in electrostatics (after first-round corrections) in the design with the extended library, which is relatively good considering the size of the search space (Table 2.1). In addition, the relative ranking of compounds based on the known inhibitor scaffold improved as scaffold placement sampling increased, highlighting the importance of adequate sampling in the ability to recover known actives.

Application of inverse combinatorial search to the *E. coli* chorismate mutase system highlights several of its advantages. These include rapidly identifying the major determinants of binding in this system and generating molecules with functional group patterns and binding modes similar to those of known inhibitors. It also demonstrates

the ability of the combinatorial search procedure to identify different scaffolds that can make similar interactions within the target site, a useful tool in developing multiple strategies for inhibitor design.

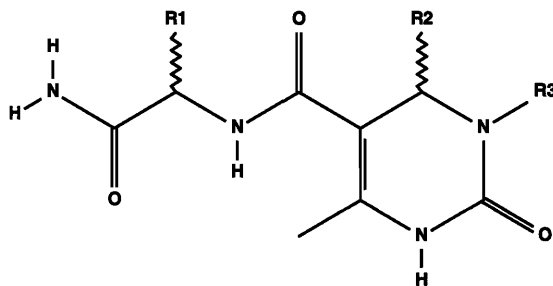
2.3.4 Evaluating scaffolds and combinatorial libraries in HIV-1 protease

When combinatorial chemistry techniques are employed to identify molecules that bind a given target, it would be useful to bias libraries towards scaffolds and functional groups enriched in molecules compatible with the site. The completeness and determinism of the inverse ligand design algorithm make it an ideal candidate for performing such a task. To demonstrate and assess its effectiveness in this mode, we used the technique to evaluate the feasibility of a combinatorial synthetic scheme for the development of HIV-1 protease inhibitors.

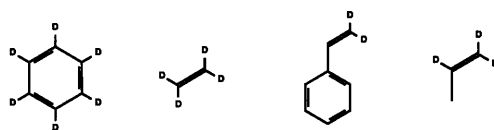
The example combinatorial scheme shown in Figure 2-10 is based on a solid-phase Biginelli reaction [99, 112], where a β -ketoester, aldehyde, and urea are condensed to form a dihydropyrimidone ring scaffold [113] (Figure 2-10A). In this particular case, the β -ketoester is attached to an α -amino acid linker, allowing for efficient solid-phase synthesis [99] and diversification with natural amino acids (except proline) at the R1 position. The R2 position can be varied by changing the aldehyde used in the condensation, and a set of four such molecules were considered (Figure 2-10B). A ring nitrogen can also be reacted with the 24 chloroformates shown in Figure 2-10C to generate a carbamate linkage at R3. The stereochemistry of attachment at the R1 and R2 positions are also unspecified, allowing for further diversity.

To judge the fitness of this example combinatorial library for the HIV-1 protease active site, we began by applying the inverse inhibitor design procedure using the scaffold and functional groups shown in Figure 2-10. The structure for HIV-1 protease was taken from a bound complex with the inhibitor TMC-114, and the fixed

A (Scaffold)



B (R2)



C (R3)

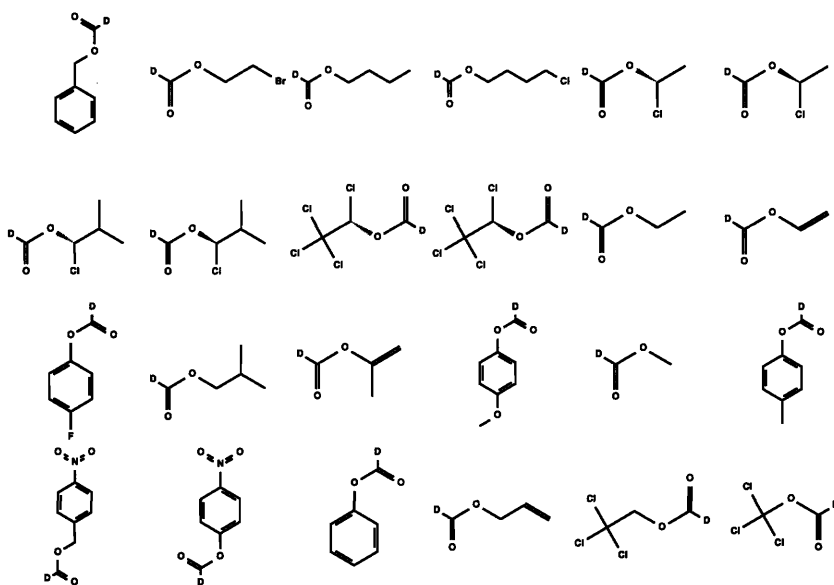


Figure 2-10: Example combinatorial library used to target the HIV-1 protease active site. The scaffold (A) is derived from a synthetic scheme based on solid-phase Biginelli chemistry, and contains three sites for diversification, denoted by R1, R2, and R3. Waved bonds indicate that both possible stereochemistries are permitted. The R1 position was joined to all natural amino acid side chains (except proline) in either stereochemistry. The R2 position was diversified with the groups shown in (B), which are derived from their respective aldehydes. Chloroformate-derived moieties shown in (C) were joined to the R3 position on the scaffold. The atoms labeled D in (B) and (C) indicate hydrogen atoms that are also possible points of attachment to the scaffold.

target shape was generated by minimizing the energy of a set of spheres placed inside the known location of the TMC-114 inhibitor until the central four pockets of the active site were well packed. In a similar fashion to the designs presented above, a pre-generated conformational ensemble for the scaffold was sampled throughout the fixed target shape for approximately 88,000 discrete placements, and the functional groups were grown from the scaffold in accordance with the synthetic scheme to identify global minimal energy configurations. In the case of unspecified stereochemistry, functional groups were grown in one or the other configuration but not both simultaneously. The total size of the search space was 3.3×10^{12} ligand structures (Table 2.1).

The designed molecules obtained after searching the Biginelli-based combinatorial scheme for HIV-1 protease compatibility need to be compared against a reference in order to measure fitness. HIV-1 protease, being an enzyme that cleaves proteins, should have an active site that is compatible with binding peptides. Consequently, we repeated the inverse design procedure using a tetrameric peptide backbone (with blocked termini) as a scaffold, and all of the natural L-amino acid side chains (except proline) as functional groups, to serve as a control experiment. If the combinatorial compounds from the Biginelli scheme score significantly better than designed peptides, it suggests that this combinatorial scaffold/library combination is feasible. It is important to note that the level of sampling utilized in the combinatorial library and peptide reference designs was kept as similar as possible to ensure a fair comparison.

Figure 2-11A shows an energy histogram of the top 1,000 scoring compounds from both the combinatorial Biginelli and reference peptide designs in the full energy model. These compounds were identified through hierarchical energy functions by taking all designed structures within the top 15 kcal/mol of the combinatorial search, re-ranking them in an initial set of energetic corrections, and sending the top 5 kcal/mol to the full binding energy model to identify the highest ranking 1,000

structures. Top compounds derived from the Biginelli scheme scored approximately 8 kcal/mol worse on average than tetrapeptides derived from the reference design. This energetic difference primarily came from reduced van der Waals interactions, as members of the combinatorial library were only able to partially fill the HIV-1 binding site (Figure 2-12). To quantify this lack of shape complementarity, the volume of the target shape filled by the best scoring molecules from each design was recorded. As shown in Figure 2-11B, the highest scoring molecules from the Biginelli scheme are consistently filling around 150 Å³ less of the target volume than designed tetrapeptides.

In sum, this study suggests that the particular Biginelli chemistry-based combinatorial library, shown in Figure 2-10, is poorly suited for the HIV-1 protease active site due to its inability to fill the site adequately and score better than model peptides. This example also serves to highlight the ease in which combinatorial libraries can be evaluated with this inverse design method. Comparing combinatorially designed molecules against a reference library can quickly gauge feasibility through both energetic scoring and shape complementarity.

2.4 Materials and Methods

2.4.1 Protein structure preparation

Structures of mutant T4 lysozyme used in this study were obtained from the Protein Data Bank (PDB), with accession codes 181L [91] and 1LGU [92]. The structure 181L contains L99A T4 lysozyme complexed with benzene, which was removed from the binding site. Structure 1LGU contains the L99A/M102Q double mutant with a beta-mercapto ethanol molecule and a water molecule in the binding site, which were also removed. For the *E. coli* chorismate mutase study, the structure 1ECM [103] was obtained from the PDB, containing the enzyme bound to a transition state analog.

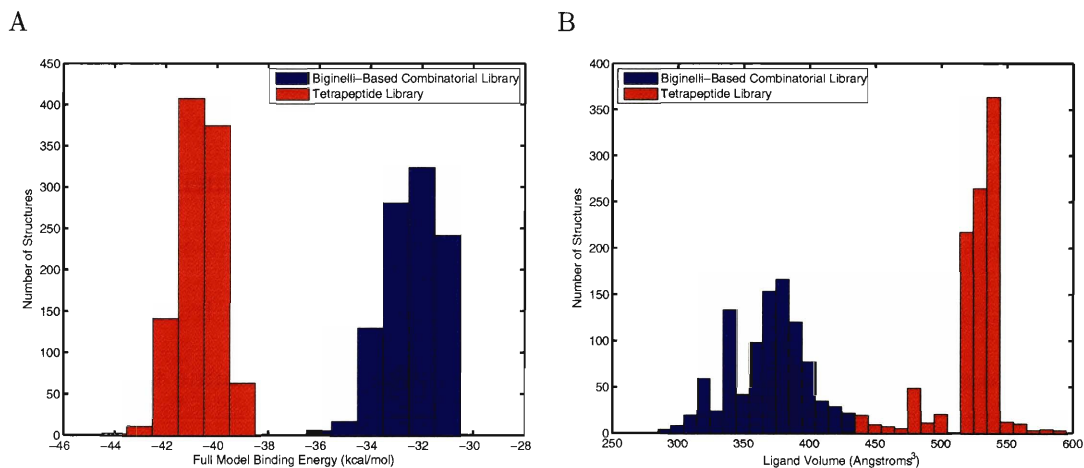


Figure 2-11: Comparison of the top 1,000 scoring molecules in the full binding energy model for two libraries targeting the HIV-1 protease active site. A comparison in binding energy score (A), shows that molecules designed from a Biginelli chemistry-based library (Figure 2-10) score worse on average than compounds designed from a reference tetrapeptide library by approximately 8 kcal/mol. Comparing the volume of the target shape filled by these compounds (B) offers an explanation for this energetic discrepancy, as Biginelli-based compounds fill substantially less of the site.

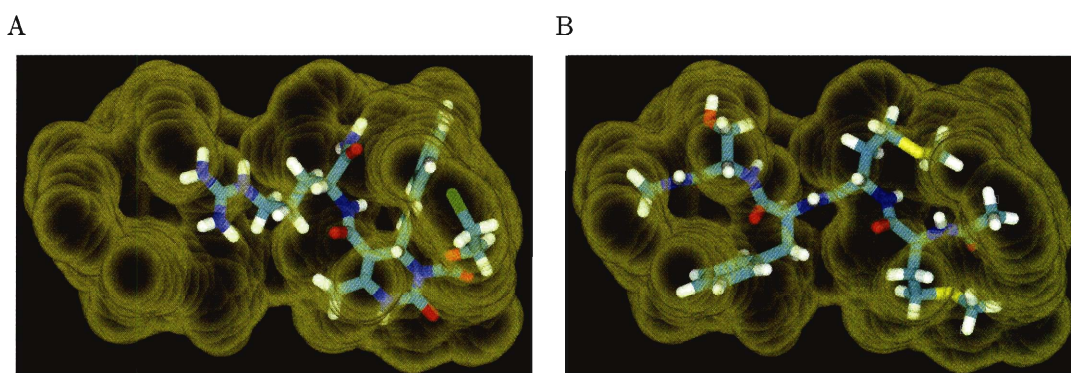


Figure 2-12: Visual comparison of a top ranking compound from the Biginelli-based combinatorial scheme (A), and a top ranking tetrapeptide design (B), in terms of their ability to fill the target shape (transparent yellow spheres). In general, compounds based on the Biginelli-derived scaffold are only able to partially fill the site.

The analog from the second active site (chain M) was removed from the structure as it was the site targeted for design. The single active site water molecule was left in this structure because it mediates hydrogen bonds between the analog and the enzyme. HIV-1 protease structure 1T3R [114] was obtained from the PDB, containing the protease bound to TMC-114, was used for the scaffold evaluation study after removal of the inhibitor. The flap water molecule, which mediates ligand binding in this structure, was left in place during design. All protein structures were further prepared by removing additional water molecules not completely buried by the protein. In addition, all glutamine, asparagine and histidine side chains were visually inspected for 180-degree flips about their final rotatable bond that would optimize the hydrogen bonding network. Ionizable residues were left in their standard states at pH 7. All proteins were parameterized using the general CHARMM22 force field [115]. Side chains with missing density were built back into the structure using default geometry, and missing terminal density was handled by adding acetyl and methylamine blocking groups to the missing N and C termini, respectively. Hydrogen atom positions were built on all protein structures using the HBUILD functionality [116] in the CHARMM computer program [117], using a distance dependent dielectric constant of 4.

2.4.2 Scaffold and functional group library preparation

Scaffolds and functional groups used for design in T4 lysozyme and *E. coli* chorismate mutase were derived from previous database screening studies of the Comprehensive Medicinal Chemistry (CMC) database [97] by Bemis and Murcko (reference [95] Chart 2 and reference [96] Chart 1). Scaffolds that contain double-bond attachment sites (indicated by double dots in Reference [95] Chart 2) were combinatorially attached to every double bonded functional group and added to the scaffold library. In addition, only the site of attachment indicated for CMC-derived functional groups was allowed during design. Additional scaffolds for chorismate mutase were derived from Husain

et al., Scheme 1 and 2. For the HIV-1 protease scaffold evaluation study, the Biginelli chemistry-based combinatorial scheme was derived from Zhang and Rana [99] and functional groups for these scaffolds were the natural amino acid side chains (except proline) at R1, and groups derived from 4 commercially available aldehydes and 24 commercially available chloroformates at the R2 and R3 positions, respectively (Figure 2-10). For the reference peptide design, tetraglycine with N-terminal acetyl and C-terminal methylamide blocking groups was used as the scaffold and the functional groups were the natural amino acids side chains (except proline) in the L configuration grown from the alpha carbons.

All scaffolds and functional groups were built using GAUSSVIEW 3 [118], and subjected to quantum mechanical geometry optimization using GAUSSIAN 98 [119] at the RHF/6-31G* level of theory. Partial atomic charges were fit to the electrostatic potential using the RESP method [120], and molecular mechanics parameters were assigned from the general all-atom CHARMM22 force field [115] using a rule-based method. Conformations were generated for each scaffold and functional group by allowing torsional sampling according to simple rules based on the hybridization of the bonded atoms (sp^3 - sp^3 every 120° starting at 180° , sp^3 - sp^2 every 30° starting at 0° , sp^2 - sp^2 every 90° starting at 0°). In addition, torsional angles were allowed to vary by an additional $\pm 10^\circ$ about these cardinal angles, and self-clashing conformations were removed using a 75% van der Waals radii bump check.

2.4.3 Defining the ligand envelope

For the completely buried target sites of T4 lysozyme and *E. coli* chorismate mutase, packing was achieved by placing spheres with a 1.5 Å van der Waals radius on a regular cubic lattice of 0.25 Å throughout the site. Spheres were assigned a van der Waals well depth of -0.1 kcal/mol. If a sphere had an unfavorable van der Waals interaction with the protein, it was rejected. For the open HIV-1 protease active site,

spheres with a 1.5 Å van der Waals radius were placed on a regular cubic lattice of 0.4 Å inside the volume that the TMC-114 inhibitor would occupy. Sphere positions were then minimized to convergence using the van der Waals potential in CHARMM, keeping the protease fixed. A special van der Waals function was employed where the optimal distance between two spheres was 0.5 Å, while the optimal distances for sphere–protein interactions used the standard mixing rules. This procedure used the known ligand as seed positions for the spheres, while allowing them to expand further to fill the site upon minimization.

2.4.4 Calculation of grid-based potentials

Grid-based energies for van der Waals scoring were computed on a regular cubic lattice placed within each target shape. For T4 lysozyme and *E. coli* chorismate mutase, a grid spacing of 0.125 Å was used, while 0.2 Å was used for the work in HIV-1 protease. The parameters used for van der Waals energies were derived from the general CHARMM22 [115] parameter set, which contains specified parameters for particular pairs of atom types in addition to using traditional mixing rules. As a result, separate grids were computed for each van der Waals atom type rather than using one grid and factoring out the mixing rules. To generate the grid-based energies, an atom of specified type was placed at each grid point within the target shape and its van der Waals interaction energy was computed with the receptor. After the grids were generated, the van der Waals interaction energy of any given molecule was approximated by summing the contribution from each atom, derived from trilinear interpolation from the surrounding eight points on the appropriate atom-type grid.

Grid-based electrostatic solvation and interaction potentials were also computed on regular cubic lattices within the target sites, with a spacing of 0.5 Å for T4 lysozyme and *E. coli* chorismate mutase, and 0.75 Å for HIV-1 protease. Elements in the L matrix and \vec{C} vector as well as R_{des} were computed with a continuum

electrostatic model by solving the linearized Poisson-Boltzmann equation (LPBE) using a locally modified version of DELPHI [25–27, 121]. In all cases, the LPBE was solved using a molecular dielectric constant of 4, a solvent dielectric constant of 80, a molecular surface probe radius of 1.4 Å, an ionic-strength of 145 mM, and an ion-exclusion radius of 2.0 Å. When computing the elements of L or \vec{C} , a $129 \times 129 \times 129$ finite-difference grid was employed, and a $257 \times 257 \times 257$ grid was used to compute R_{des} . A focusing scheme was used where the system being solved first occupied 23% and then 92% of the grid, transferring boundary potentials from the lower to higher resolution run. When computing elements of the L matrix and \vec{C} vector, an additional over-focusing stage at 184% fill, centered at the grid point being computed, was used to improve accuracy. In all cases, the final grid resolution was at least 0.25 Å at the maximum percent fill. Radii and charges for the receptor proteins were derived from the PARSE parameter set [22], and a radius of 1.5 Å was used for all target shape spheres.

Elements in the L matrix were calculated by taking the difference between solving the LPBE in the bound state, where the receptor and target shape defined the geometry, and the unbound state where only the target ligand shape was present. For the bound and unbound states, each grid point was charged to a value of one while all other grid points and receptor atoms had a charge of zero. After solving the LPBE, the resulting potentials at the charged grid point as well as all other grid points become one column of the solvation matrix S for that state. After computing the full S matrices for each state, the relation $L = S_{\text{bound}} - S_{\text{unbound}}$ was used to compute L . Elements in the \vec{C} vector were derived from bound state calculations, where the receptor atoms are charged to their parameter values and the grid points had zero charge. The screened Coulombic potential generated at the grid points defined the \vec{C} vector. R_{des} was computed by taking the difference between the bound- and unbound-state solvation energy when the receptor was charged and the fixed target

shape uncharged.

The approximate electrostatic binding energy for any given molecule inside the target shape was computed by first spreading each partial atomic charge on to the surrounding eight grid points with trilinear projection. If several partial atomic charges were projected to the same grid point, their contribution was summed. These grid charges were then used in Equation 2.1. In general, only a small number of grid points receive charge, and consequently Equation 2.1 was implemented using sparse vectors in order to decrease computation time.

2.4.5 Scaffold placement

Every generated conformation of each scaffold was discretely placed throughout the target shape using rigid uniform sampling. In all systems studied, scaffolds were sampled in translation on a regular 0.5 \AA cubic lattice. At each lattice point, many rigid rotations were considered by equal sampling in X, Y, and Z rotation using Euler angles. The angular step size for rotational sampling was set such that the arc length swept out by the atom furthest from the molecule center was 1.0 \AA for T4 lysozyme and 2.0 \AA for *E. coli* chorismate mutase and HIV-1 protease. Each placed scaffold was then checked against the fixed target shape to ensure that all atom centers were within the shape and at least 0.5 \AA away from its boundary. The extra distance padding was necessary to ensure that partial atomic charges were not spread to grid points outside the fixed shape. If the scaffold placement failed this check, it was rejected. In addition, the grid-based van der Waals energy was computed for each scaffold placement, and if this energy was unfavorable the placement was also rejected.

2.4.6 Functional group attachment and pairwise energy evaluation

Pairwise energies for the functional groups attached to each scaffold placement were computed in three stages. Firstly, all growable hydrogen atoms were removed from the scaffold, as the remaining atoms were constant throughout the search. The van der Waals and electrostatic grid-based energies were computed for this blunt scaffold and stored along with the precomputed receptor desolvation R_{des} as the constant term in the pairwise energy decomposition shown in Equation 2.4. Secondly, functional groups in discrete conformations were attached to each growable site by removing a hydrogen and creating a single bond between the two blunt atoms along the same vector used by the scaffold hydrogen. The bond length was set to the equilibrium value from the CHARMM22 [115] parameter set. Rotations were generated about the newly formed bond with the same hybridization-based rules used for conformational sampling. Re-attachment of the hydrogen was also permitted at each growable position, with its original bond length preserved. Newly attached discrete functional groups were examined to ensure that all atom centers were no closer than 0.5 Å to the target shape surface, and that the functional group did not intersect the blunt scaffold when 75% van der Waals radii were used. The 75% cutoff was selected because it allowed clashing yet synthesizable geometries, such as two methyl groups grown from neighboring hydrogens on a benzene ring, to be designed without penalty. If a discrete functional group failed one of these checks, it was removed from consideration in the rest of the search procedure.

When the scaffold and functional group hydrogens were removed for attachment and the two molecules fused, their partial atomic charges needed to be adjusted to maintain integral charge and account for inductive effects. In this work, a simplistic method was used where the partial atomic charge on the removed scaffold hydrogen was placed on the atom bonded to the removed functional group hydrogen, and the

charge on the removed functional group hydrogen was placed on an associated point charge located at the position of the atom attached to the removed scaffold hydrogen. The role of the point charge is to ensure that the transferred charge is part of the functional group rather than the scaffold in energy calculations. Although this method only accounts for charge transfer across the new bond, any pairwise decomposable charge spreading procedure would be allowable. This means that charge could have been spread anywhere in the functional group or the blunt scaffold, but not into a different functional group. In the case of reattaching a hydrogen, the partial charge was kept the same as it was in the original scaffold molecule. The grid-based van der Waals and electrostatic energy of the blunt functional group with point charge (or hydrogen), as well as the indirect desolvation between the functional group and the blunt scaffold was then computed using Equation 2.5 to generate the self-energy terms.

The final stage of pairwise energy calculation was to compute the pair contributions to binding, which are either the indirect desolvation terms between each pair of discrete functional groups, or an infinite energy if the two functional groups clash when 75% van der Waals radii were used (Equation 2.6). If one of the two functional groups being considered was a re-attached hydrogen, the bump check was skipped because the hydrogen was forbidden from being eliminated. To increase the computational efficiency of the quadratic matrix product in Equation 2.6, which was the slowest step in the energy calculation, the projected grid charges for all acceptable discrete functional groups were precomputed, and the inner $L\vec{q}$ product was stored and reused.

2.4.7 Combinatorial search

For each scaffold placement, the global minimum (binding) energy configuration (GMEC) of functional groups, as well as a ranked list of structures with progressively

increasing binding energy, was determined with a locally developed implementation of the dead-end elimination (DEE) [11,14,70] and A* [13] algorithms. After determining the GMEC, a ranked list of configurations with increasing energy was computed in 2.5 kcal/mol increments until the list contained 10 conformations for 100 “unique” ligands, or a limit of 25 kcal/mol above the global minimum was reached. Ligands were considered “unique” if the receptor–ligand complex differed by more than rotations about dihedral angles in the functional group attachments. In other words, “unique” ligands differ in their functional group attachment pattern, and not solely in conformation. These best 1,000 (10×100) structures were pooled across all scaffold placements, sorted on total energy, and the top one million structures were stored.

2.4.8 Hierarchical rescoring of highly ranked molecules

The molecules with the most favorable computed binding affinity from the combinatorial search procedure were first re-evaluated with fast but non-pairwise corrections to the energy function, termed “medium resolution”. These include approximations to correct for the fixed-shape assumption in the electrostatic calculation and an improvement from grid-based to explicit-atom van der Waals packing energies. The best compounds found in the T4 lysozyme and *E. coli* chorismate mutase buried sites were electrostatically corrected by computing the unbound state solvation energy via the LPBE and true shape rather than grid-based energies (Equation 2.7). The solvation free energy in the unbound state was computed with a locally modified version of DELPHI [25–27,121] using the same parameters as for grid generation except that a $65 \times 65 \times 65$ grid was used. PARSE radii [22] were assigned to the ligand.

The compounds with the most favorable binding affinity from combinatorial search in the solvent exposed HIV-1 protease site were corrected using the presented variational method [88,89]. The electric field in the bound or unbound states, present in the integrand of Equation 2.8, was computed through finite difference of the grid-

based potentials, $\phi = 2S\vec{q}$, where \vec{q} are the projected grid charges for the designed molecule and S is the solvation matrix for the appropriate state. In order to perform numerical volume integration, the computed electric field was assumed to be constant in the surrounding cube for each grid point, making its contribution equal to an evaluation of the integrand at the point times the grid spacing cubed. A grid cube contributed to the integral only if its associated point was inside the target shape, but became inside solvent when the dielectric boundary shrank from the target shape to a molecular surface derived from PARSE radii [22]. In order to determine solvent accessibility for a grid point, triangulated surfaces were generated in DELPHI using its boundary-element features [122], and point-in-polyhedron tests were performed using the GNU triangulated surface library (GTS) [123].

The best molecules after re-scoring with these fast but approximate corrections were subjected to further re-evaluation with a more sophisticated method of computing binding energies, utilizing a full Poisson–Boltzmann/surface area (PBSA) model [22]. In this approach, the differential solvation energies between the bound and unbound states were added to the vacuum ($\epsilon = 4$) interaction energy between ligand and receptor. Solvation energies were computed by solving the linearized Poisson–Boltzmann equation with a locally modified version of DELPHI [25–27, 121] using a $129 \times 129 \times 129$ grid and the same parameters as for computing grid potentials. PARSE radii [22] were used for both the receptor and designed molecules, while PARSE charges were used only for the receptor. Charges on the designed molecules were the same as those used in the combinatorial search, which consisted of quantum mechanically derived charges for scaffolds and functional groups merged as described above. The hydrophobic contribution to solvation was computed as directly proportional to the solvent accessible surface area buried upon binding, using a probe radius of 1.4 Å and a scaling factor of 5 cal/Å² [22]. Vacuum electrostatic interactions used a constant dielectric constant of 4, and explicit-atom van der Waals interactions were carried

over directly from the “medium resolution” energy function correction.

Chapter 3

Computational Design of Substrate Envelope Inhibitors to Avoid Resistance Mutations: HIV-1 Protease as a Model System¹

Abstract

Drug resistance is a growing concern in the treatment of rapidly evolving pathogens. One of the most common mechanisms of drug resistance, especially in viral infections, is the accumulation of escape mutations in the drug target that reduce drug binding while maintaining normal protein function. Consequently, it is necessary to develop new therapeutics that are less likely to induce escape mutations, especially in cases where the exact mechanisms of resistance are unknown. One such strategy, which is especially applicable to drug targets that are essential enzymes for the pathogen, is to design inhibitors that mimic the structural features and binding modes of substrates. This idea, termed the substrate envelope hypothesis, suggests that target enzymes would not be able to mutate against inhibitors that remain inside a consensus substrate shape without also disrupting substrate binding. In order to begin testing the substrate envelope hypothesis as a drug design principle, we have employed HIV-1 protease as a model system. Computational inverse small-molecule design techniques were applied to design inhibitors that would stay within a consensus substrate envelope and were predicted to have high affinity. Fifteen envelope inhibitors were chemically synthesized and tested for binding against wild-type HIV-1 protease, with four compounds having inhibition constants (K_i) ranging from 30–50 nM. These four compounds exhibited broad specificity against a panel of three drug resistance HIV-1 protease sequences, losing no more than 6–13 fold affinity relative to wild-type. A comparison of predicted to experimental binding energies as well as predicted to crystallographic structures for inhibitors highlights the limitations and strengths of our computational models. Overall, these results suggest that designing inhibitors that mimic substrates may be a viable strategy in the development of therapeutics with better resistance properties.

¹All experimental work presented in this chapter was performed by collaborators at the University of Massachusetts Medical School in the laboratories of T. M. Rana and C. A. Schiffer. Specifically, chemical synthesis was performed by A. Ali and K. Reddy, HIV-1 protease inhibition assays were performed by H. Cao, and crystallography was performed by M. Nalam.

3.1 Introduction

One of the most important limiting factors in the current treatment of rapidly evolving pathogens is drug resistance [35–42]. Although drug resistance can be caused by a variety of mechanisms [40, 124–127], one of the most common, especially in viruses, is when the drug target mutates such that drug binding is reduced while the normal function of the target is maintained [40, 126, 128, 129]. Due to the increasing prevalence of this form of resistance, drugs need to be developed that do not induce these viable escape mutations. Most current efforts to design therapeutics with improved resistance profiles have focused on analyzing the failure modes of existing drugs and designing new compounds that have high efficacy against known resistant mutants [114, 130–132]. Although this strategy has met with success [132, 133], it has also resulted in the identification of new, unanticipated modes of resistance [134, 135]. It is clear that a strategy is needed to design drugs that will not encourage escape mutations to form even when the possible modes of resistance are unknown.

One such strategy, which is especially applicable to targets that are essential enzymes, is to design drugs that mimic the structural features of substrates [44, 136–139]. If a drug molecule makes the same interactions and contacts with the target as substrate, meaning that its determinants of affinity are the same, it could be difficult for mutations to evolve that block inhibitor binding but maintain substrate recognition. Consequently, escape mutations should never form because they would render the pathogen non-viable. This idea can be summarized as the “substrate envelope hypothesis,” meaning that inhibitors that stay within a consensus substrate shape should be less likely to induce resistance mutations than those that exceed the envelope and provide handles for escape mutations to lower inhibitor affinity selectively (Figure 3-1).

The goal of this work is to begin testing the substrate envelope hypothesis as an inhibitor design principle using HIV-1 protease as a model target and computational

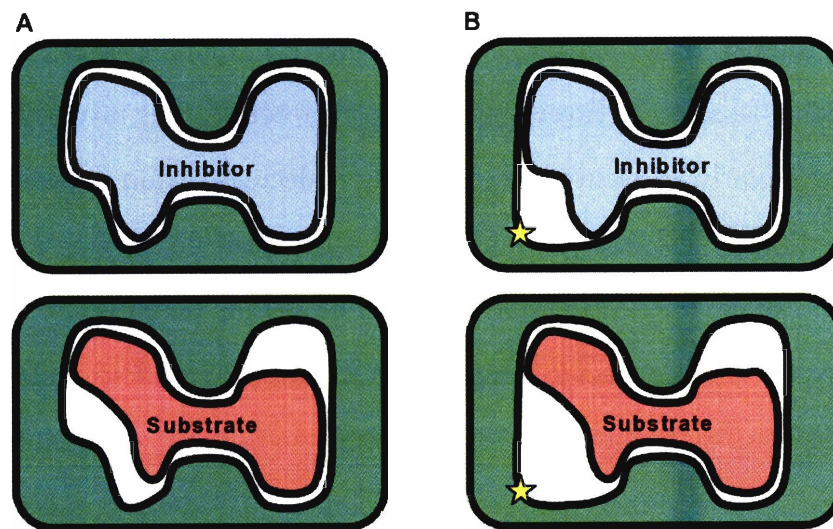


Figure 3-1: Example of the substrate envelope hypothesis. In the wild-type drug target (A), the traditional inhibitor (top) occupies more of the binding site and makes more contacts than a substrate (bottom). In (B), the drug target has mutated to expand the active site in a region that only contacts the inhibitor (star). The inhibitor (top) loses contacts and consequently binding affinity, while the substrate (bottom) loses no affinity as it never contacted the mutable residue. If the inhibitor had been designed to look more like the substrate, this resistance mutation would have little effect on its binding affinity.

ligand design techniques to generate inhibitors that mimic the substrate shape. HIV-1 protease was selected as a model system due to the vast amount of structural, inhibitor, and resistance data available. Mechanisms of drug resistance in HIV-1 protease have been well studied [36, 40, 140], and binding modes of several HIV-1 protease peptide substrates have been determined [43, 44]. The substrate complexes suggest a consensus substrate envelope [44] that can serve as a drug design target. In order to design inhibitor molecules that stay within the envelope and are predicted to have high affinity to the protease, we employed a combinatorial small-molecule design technique based on an inverse approach (Chapter 2). The inverse ligand design strategy is particularly well suited to generating molecules within a specified envelope because a fixed target shape is a requirement in the algorithm, serving as both a limit on the size and shape of the ligand as well as a molecular boundary for electrostatic modeling. Inverse small-molecule design takes a library of scaffold molecules, places them discretely within the envelope, and performs guaranteed combinatorial searches over a discrete space of functional group attachments to identify molecules that fit in the shape and are predicted to have high affinity. For this work, the scaffold employed was a variant of the core of the clinical inhibitor amprenavir [130] (Figure 3-2), which is known to fit well inside the substrate envelope [136], is relatively easy to chemically synthesize, and has three sites for functional group attachment. The functional group library, however, was selected naively, and consisted of reagents from chemical catalogs.

In order to measure how well the substrate envelope inhibitors perform, a set of fifteen compounds suggested by computation were chemically synthesized. These compounds were assayed for binding against the wild-type protease and all exhibited inhibitory activity, with four compounds having K_i values in the range of 30–50 nM. In addition, the affinity of these top four compounds was measured for three drug resistant proteases derived from co-evolving mutations in clinical isolates [141, 142].

Compared to two clinical inhibitors, these compounds exhibited broad specificity, losing similar fold affinity to the clinical inhibitor amprenavir, known for its effectiveness *in vivo* against multi-drug resistant HIV-1 strains [143]. Ultimately, these compounds need to be tested in accelerated evolution studies [144, 145], where HIV-1 protease is grown *in vitro* under strong selection pressure from protease inhibitors to determine if it is possible to evolve escape mutants against them.

Overall, the successful design of compounds that stay within the HIV-1 protease substrate envelope and have broad specificity against resistant mutants supports the substrate envelope hypothesis, setting the stage for more complete validation studies. Currently, efforts are underway to design a second round of inhibitors that stay within the envelope and have higher affinity, as well as to specifically design inhibitors that exit the substrate envelope. These compounds should have poor resistance profiles, providing negative controls for the substrate envelope hypothesis in a design format.

3.2 Computational strategy

In order to computationally design compounds predicted to stay within the substrate envelope and have high affinity to the wild-type HIV-1 protease, we employed a previously described inverse small-molecule design algorithm (Chapter 2). In this section, we provide a brief overview of the method and how it was directly applied to the design of compounds capable of binding within the envelope, highlighting choices that were specific to this system as well as any deviations from the established inverse design protocol.

3.2.1 Substrate envelope selection

The inverse small-molecule design strategy requires several inputs. The first of these is a fixed envelope that serves as a hard outer constraint on the size and shape of

designed ligands. Two fixed shapes, based on the superimposed structures of five peptide substrate-inactivated HIV-1 protease complexes, were generated by placing spheres on a regular grid surrounding the superimposed substrates and keeping those that were within 1.0 Å of a substrate atom center in either 1 of the 5, or 3 of the 5 substrates. The shape generated when requiring that spheres are near atoms in 3 of the 5 substrates provided a tighter envelope.

3.2.2 HIV-1 protease structure selection

Another requirement of the inverse design algorithm is that the fixed envelope must be placed inside a protein structure such that binding energetics for molecules built inside can be evaluated. When designing compounds that are attempting to mimic substrate geometries, it makes sense that the protein receptor should be well suited for binding substrates. Consequently, both envelopes were placed back into one of the inactivated HIV-1 protease structures, specifically the reverse transcriptase-RNaseH (RT-RH) substrate complex (PDB code 1KJG), using the same alignment that generated the substrate superposition used for envelope generation. This structure was selected because the RT-RH substrate is the largest peptide in volume, and its associated protease accommodated the substrate envelope well. However, there were a few substrate envelope spheres that clashed with the protease structure when the envelope was placed in the site, and these spheres were removed. In addition, extra spheres were added to the envelope within 3.5 Å around the catalytic dyad. This was necessary because the substrates do not closely interact with the catalytic residues, but most inhibitors make direct hydrogen bonding interactions. Adding these spheres should not increase the chance of escape mutations because the catalytic aspartates are invariant. In some sense, the role of the extra spheres is to more closely mimic the transition state envelope at the site of catalysis.

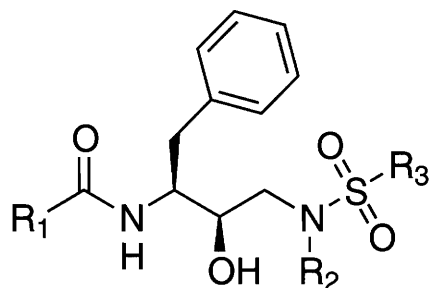
One final modification that needed to be made to the protease structure was

the “re-activation” of the catalytic residues. The protease–complex structures were solved in the context of a deactivating D25N mutation in order to prevent substrate processing. Consequently, the catalytic residues were mutated back to aspartate, using the same side-chain dihedral angles as the crystallographic asparagines. The protonation state of the aspartyl dyad in HIV-1 protease has been the subject of much debate [146–148]. Therefore, two protonation states were considered separately in the inverse design procedure. One model has the dyad doubly deprotonated, and the other is singly protonated such that a hydrogen bond is formed that bridges the two aspartates.

3.2.3 Scaffold and functional group search space

The inverse small-molecule design method expresses the combinatorial space of inhibitor molecules as molecular scaffolds with functional group attachments. For this work, it was important to select a molecular scaffold that was easy to chemically synthesize and that fits well inside the substrate envelope. The scaffold selected to form the core of all designed compounds is shown in Figure 3-2. It has three sites for attachment, denoted by the R groups. Given the synthetic scheme for molecules based on this scaffold, functional groups at the R1 position are derived from carboxylic acids, groups at the R2 position are derived from primary amines, and attachments at R3 come from sulfonyl chlorides. This scaffold is very similar to the core of the clinically approved inhibitor amprenavir (APV) [130], except that an amide linkage is used to attach the R1 group rather than a carbamate linkage.

Functional groups were selected naively from chemical databases. A set of 2,327 carboxylic acids and 379 primary amines for the R1 and R2 positions respectively were obtained from the ZINC database [149] (as of December 2004). 274 sulfonyl chlorides were obtained for the R3 position directly from the Sigma–Aldrich and Maybridge catalogs.



R₁: From Carboxylic Acids (R₁-COOH)
 R₂: From Primary Amines (R₂-NH₂)
 R₃: From Sulfonyl Chlorides (R₃-SO₂Cl)

Figure 3-2: Hydroxyethyl sulfonamide core used as the scaffold for substrate envelope inhibitor design. The scaffold can be diversified at three positions, R₁, R₂, and R₃, using functional groups derived from carboxylic acids, primary amines, and sulfonyl chlorides, respectively. The scaffold is similar to the core of the clinical inhibitor amprenavir [130] and the pre-clinical inhibitor TMC-114 [114].

3.2.4 Grid-based energies and scaffold placement

The next stage in the inverse design inhibitor procedure was to compute pairwise grid-based potentials and energies within the fixed envelopes that will be used for fast evaluation of binding energetics [74, 84–86]. Overall, four sets of grids were computed, for both the tight and loose substrate envelopes and both protonation states of the protease. Once grid generation was complete, the next stage involved finding docked conformations for the scaffold that had a good disposition for growing functional groups that could interact favorably with the grid potentials. Two methods were used to achieve this goal. The first method used uniform sampling to build the scaffold into the envelope by rigidly docking a scaffold fragment and extending it to a complete molecule while sampling torsional degrees of freedom. Scaffold placements that remained in the envelope, scored well against the grid-based potentials, and could successfully grow prototypical functional groups from all three attachment points were accepted.

The second scaffold sampling method used knowledge of the known binding mode of this scaffold observed in crystal structures. The structure of the amprenavir com-

plex (PDB accession code 1HPV) [130] was aligned to the substrate structure and the coordinates of amprenavir were retained. Scaffold placements were then derived by threading onto the known binding mode to generate an ensemble of scaffold positions that had an RMSD less than 1.0 Å. Again, placements were accepted if they scored well and could grow three functional groups.

3.2.5 Combinatorial search and hierarchical re-scoring

A total of eight independent inverse design runs were carried out, to cover all combinations of both the tight and loose substrate envelopes, the singly and doubly deprotonated protease structures, and both the uniform and knowledge-based scaffold sampling techniques. For each scaffold placement in each independent design, discrete conformations of the functional group library were grown from each attachment site, computing the self energy of each functional group as well as pair energies between groups. Because the energy function used in this stage of the inverse design is pairwise additive, global minimal and rank-ordered energy arrangements of functional groups could be identified using the deterministic and guaranteed search algorithms dead-end elimination (DEE) [11, 14, 70] and A* [13]. The top ranking compounds across all scaffold placements in each design run were pooled and hierarchically re-evaluated using more sophisticated energy models to correct for approximations such as the use of grid-based energy functions and a fixed envelope in electrostatic calculations. In this manner, we could ensure that the top ranking compounds had been identified in a full binding energy model, which included explicit-atom van der Waals interactions, a full continuum solvation treatment of electrostatic effects, and a non-polar solvation term directly proportional to the surface area buried upon binding.

3.3 Results and Discussion

3.3.1 Development of a substrate envelope inhibitor library

After inverse design calculations were performed to generate ligands within the substrate envelope and predicted to have high affinity, the space of top-ranking compounds was hand pruned to select a set of twenty compounds for chemical synthesis (Figure 3-3). Several criteria were used to generate the final library when examining the top compounds from all eight design efforts. First, nine compounds were selected (Figure 3-3, Compounds 1–9) where a functional group attachment made direct hydrogen bonding interactions with backbone atoms of the protease in at least one of the design runs. The next set of five compounds (Figure 3-3, Compounds 10–14) were selected because they scored highly in all eight inverse design conditions, with the idea that these compounds were robust to changes in the design protocol. Two additional compounds (Figure 3-3, Compounds 15–16), were selected because they scored very highly in one particular design, which used the loose envelope, the doubly deprotonated protease structure, and uniform scaffold sampling. The final four compounds (Figure 3-3, Compounds 17–20) were selected because they cross-validated well with other computational small-molecule affinity estimation techniques. The first two of these compounds scored well in a molecular docking algorithm [150,151], and the last two scored well in the commercial docking program Glide [6,152] from Schrödinger, Inc. For the docking studies, the doubly deprotonated model of the protease was used as a target.

3.3.2 Experimental binding affinities to wild-type HIV-1 protease

Of the twenty designed compounds, the inhibitory activity (K_i) of fifteen was experimentally measured against the wild-type HIV-1 protease using an enzymatic inhi-

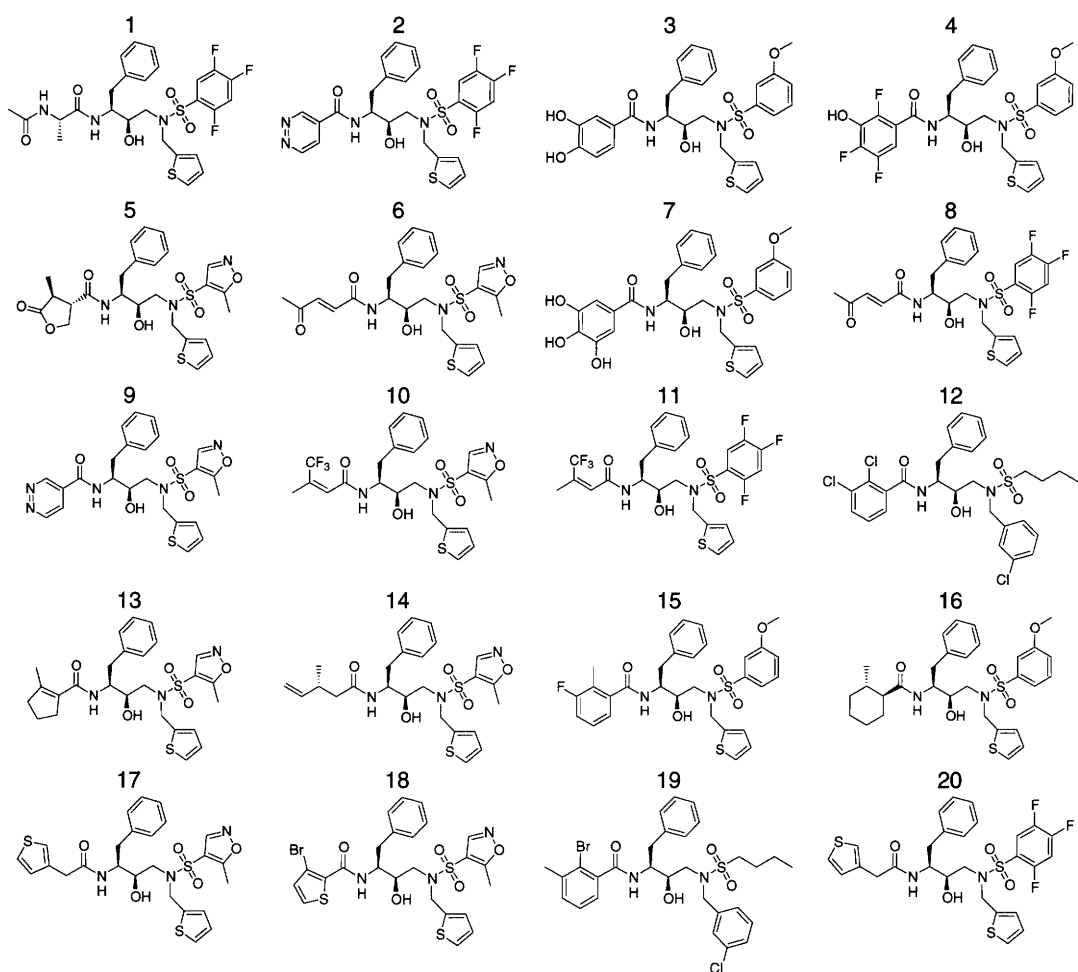


Figure 3-3: Twenty compounds proposed by computational inverse ligand design to bind HIV-1 protease within the substrate envelope and with high affinity.

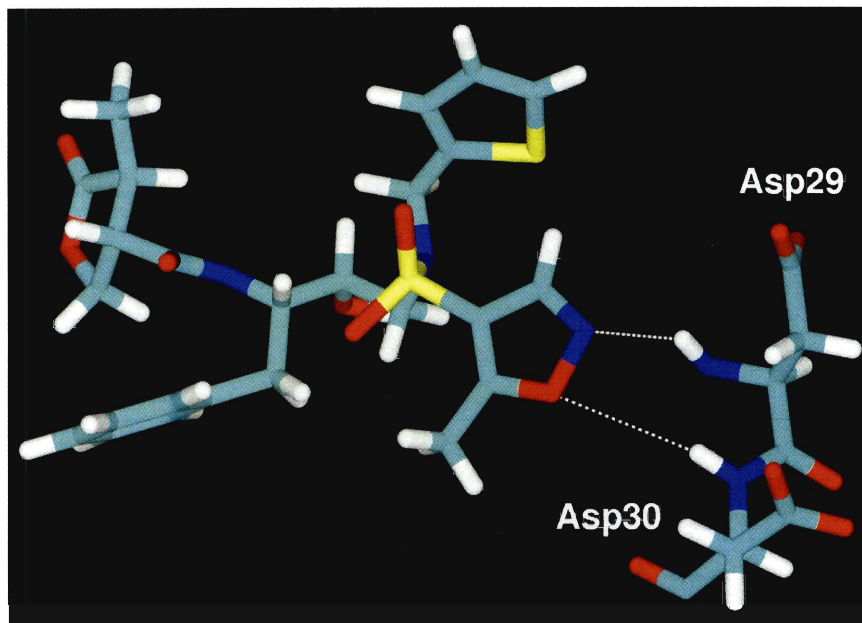


Figure 3-4: Predicted interactions for isoxazole rings designed at the R3 position. The ring oxygen and nitrogen were predicted to make favorable hydrogen bonding interactions with the backbone amide hydrogens of protease residues Asp29 and Asp30. However, the isoxazole group was detrimental to binding, as compounds containing the ring had inhibition constants (K_i) greater than $10 \mu\text{M}$.

bition assay [153]. The results of these assays are shown in Table 3.1. All fifteen compounds tested had measurable inhibitory activity against the wild-type protease, with K_i values ranging from approximately 30–26,000 nM. For reference, the K_i values for two clinically approved inhibitors, lopinavir (LPV) and amprenavir (APV), were also measured.

When comparing the structures of the compounds and their relative binding affinities, several trends emerged. Firstly, the use of a methylated isoxazole ring at the R3 position (Compounds 6, 10, and 14) was detrimental to binding, resulting in inhibition constants greater than $10 \mu\text{M}$. This was surprising considering that these compounds were predicted to make strong hydrogen bonding interactions with the backbone amide hydrogens of protease residues Asp29 and Asp30 (Figure 3-4). The five untested compounds were not measured because they all contained this isoxazole ring and would likely have poor affinity.

Several other groups that were predicted to make hydrogen bonding interactions were well tolerated in this set of compounds, including an alanine substitution at R1 (Compound 1), a catechol ring at R1 (Compound 3), and a meta anisole at R3 in several compounds. In addition, several non-polar groups were well tolerated, including fluorinated groups at R1 and R3 in several compounds, and a thiophene ring at R2.

Overall, even the best binders in the set of designed compounds were still 2–4 orders of magnitude weaker than the clinical inhibitors lopinavir and amprenavir. One possible explanation is the commonly designed thiophene ring at the R2 position, which interacts with the P1/P1' pocket of the protease. The far majority of known tight-binding inhibitors of HIV-1 protease use highly aliphatic or aromatic groups to interact with this hydrophobic pocket [154]. Therefore, it is possible that the thiophene ring may still be too polar for the hydrophobic P1 subsite. However, the thiophene moiety was one of the least polar functional groups available in the naive primary amine library used to diversify the R2 position in the inverse design calculations. Unsubstituted aliphatic and aromatic groups were not present in the set of compounds searched at R2. Therefore, the low affinity observed in this first-round library may be due to limited diversity in the initial selection of functional groups, rather than deficiencies in the design algorithm itself.

3.3.3 Experimental binding affinities to drug resistant HIV-1 proteases

In order to gain initial insight into whether compounds designed to stay within the substrate envelope would be less likely to elicit escape mutations, the inhibition constants for several designed compounds were experimentally measured in drug resistant proteases. Three drug resistant HIV-1 protease mutants were selected from previously identified sets of co-evolving mutations in drug resistant clinical isolates by Shaffer

Table 3.1: Experimentally measured inhibition constants against wild-type HIV-1 protease for two clinically approved inhibitors and fifteen inhibitors computationally designed to stay inside the substrate envelope.

Compound	K_i (nM)	Compound	K_i (nM)
LPV	0.002	10	13232
APV	0.13	11	56
1	33	12	377
2	1064	14	11596
3	50	15	42
4	515	16	582
6	26318	19	650
7	1148	20	2360
8	609		

and coworkers [141] and Swanstrom and coworkers [142]. The three drug resistant proteases contained either “Group 2” mutations (L10I, G48V, I54V, L63P, V82A) which are tightly linked to lopinavir (LPV) resistance, “Group 3” mutations (D30N, L63P, N88D), which co-evolve in response to treatment with nelfinavir (NFV), or “Group 4” mutations (L10I, L63P, A71V, G73S, I84V, L90M) which confer moderate resistance to a variety of inhibitors. The four designed compounds with the highest affinity to the wild-type protease were tested against the three drug resistant proteases, in addition to the two clinical inhibitors lopinavir (LPV) and amprenavir (APV). The results of the binding assays are shown in Table 3.2.

The clinical inhibitor lopinavir (LPV) exhibited the worst resistance profile, losing 1,220-fold inhibition relative to wild type in the Group 2 resistant protease, which was unsurprising given that the Group 2 mutations were known to co-evolve in response to lopinavir treatment [142]. On the other hand, amprenavir (APV) exhibited a relatively flat resistance profile for these protease mutants, losing at most 11-fold inhibition in Group 4. Group 4 mutations, especially I84V, are known to correlate with amprenavir treatment [142]. Although the designed inhibitors bound weaker than the clinical compounds, they had resistance profiles similar to or better than amprenavir, losing 6–13 fold inhibition in the three mutants tested. One interesting

Table 3.2: Experimentally measured inhibition constants against wild-type and three drug resistant HIV-1 proteases for two clinically approved inhibitors and four inhibitors computationally designed to stay inside the substrate envelope.

Compound	Wild-type K_i (nM)	Group 2 K_i (nM)	Group 3 K_i (nM)	Group 4 K_i (nM)	Worst Fold Loss
LPV	0.005 ± 0.02	6.1 ± 2.0	0.04 ± 0.23	0.9 ± 0.7	1220
APV	0.13 ± 0.11	0.15 ± 0.37	0.21 ± 0.44	1.4 ± 1.0	11
1	33 ± 10	267 ± 21	29 ± 3	95 ± 13	8
3	50 ± 7	382 ± 40	66 ± 11	139 ± 11	8
11	53 ± 22	142 ± 40	127 ± 31	674 ± 119	13
15	42 ± 10	257 ± 62	85 ± 11	79 ± 17	6

feature in the resistance patterns of the designed inhibitors was that small changes in structure were sufficient to shift the resistance profile. For example, when an alanine group at R1 in Compound 1 was changed to a fluorinated hydrocarbon in Compound 11, the resistance profile shifted from being Group 2 sensitive to strongly Group 4 sensitive.

Although the designed compounds show broad specificity against this particular set of drug resistant proteases, these results do not guarantee that the designed compounds will not be susceptible to alternative resistance pathways. For example, the clinically approved inhibitor amprenavir (APV), exhibits broad specificity in these assays but induces an alternative resistance pathway in clinical isolates involving the I50V mutation [134]. Therefore, to fully test the substrate envelope hypothesis, it will be necessary to ensure that alternative or previously unknown resistance pathways do not develop in response to treatment with the envelope inhibitors. One possible way to accomplish this would be to use the designed compounds as selection pressure in accelerated evolution studies of HIV-1 in cell culture [144, 145].

3.3.4 Comparison of predicted to experimental binding affinities

In order for computational inhibitor design algorithms to enrich libraries with compounds that bind tightly experimentally, it is important that the scoring functions used for design correlate well with experimental affinities. The energy function used to perform the final ranking of compounds in the inverse design procedure was based on a rigid-binding Poisson–Boltzmann surface area (PBSA) approach [22]. It included contributions to the binding free energy from van der Waals forces, electrostatic solvation as computed with the linearized Poisson–Boltzmann equation, and a non-polar solvation term directly proportional to the surface area buried upon binding. A comparison between computed and experimental binding energies is shown in Figure 3-5A for the 15 designed compounds measured as well as the clinical inhibitor amprenavir (APV).

Overall, little or no correlation was observed between the computed and experimental binding energies for the designed compounds. One possible explanation for this finding is that previous reports have shown that it is difficult for computation to discriminate between compounds that are separated by less than 3–4 orders of magnitude in binding affinity with energy functions similar to those used in this work [155]. However, it is surprising that amprenavir (APV) is not ranked significantly better than the designed compounds, considering that its experimental inhibition constant was five orders of magnitude tighter than the weakest binding designed inhibitors.

In an attempt to improve the correlation, several modifications to the scoring function were tested. These included using several dielectric constants for the molecular interior, several minimization protocols for the designed structures before scoring, and recomputing partial atomic charges for the entire designed molecules, rather than the fragment-based approach used in the inverse design technique. In addition, several methods for weighting the various energy terms in determining a final score

were tested. However, none of these efforts resulted in significantly better agreement with experiment (data not shown). The inability to improve correlation led to a close examination of the predicted structure for amprenavir from inverse design in the minimal shape. Due to the position of the I84 side chain in the substrate-bound protease structure, the isobutyl group of amprenavir was unable to adopt the conformation observed in its own crystal structure due to the presence of a van der Waals clash. Therefore, another possible explanation for the poor energy ranking was that the substrate-bound protease structure lacked the induced fit required to bind amprenavir.

In order to significantly improve the correlation, it was ultimately necessary to change the the model used for inhibitor envelope as well as the structure used for the protease. In the inverse design protocols described thus far, the target shape for design was set to the substrate envelope. Experimentally, there is no requirement that the lowest-energy structure for the inhibitor–protease complex stay within this envelope, and the addition of this artificial constraint on ligand geometry may have been partially responsible for decreased correlation. The fact that the designed inhibitors scored well within the substrate envelope only signifies that there exists a low energy envelope structure, which is not necessarily the overall structural minimum. However, the use of a “maximal” ligand envelope in inverse design, which completely fills the central four pockets of the protease active site, was still insufficient to observe correlation with experimental affinities when used in conjunction with a protease structure derived from a substrate complex. It was also necessary to switch from using the protease from the RT–RH substrate complex to using a protease structure that was bound to an inhibitor containing a scaffold similar to the one used in the designed compounds.

The inverse design procedure was repeated using the protease structure from a complex with the tight-binding small-molecule inhibitor TMC-114 [114]. TMC-114,

a variant of amprenavir, contains a similar scaffold to the one used for design and its known binding mode was used to generate scaffold placements for the inverse design procedure. Instead of using the substrate envelope as the target shape for design, a “maximal” shape was employed that completely packs the central four pockets of the protease with spheres. The functional group library used for design contained the fragments necessary to regenerate the fifteen previously designed compounds that had experimental affinities as well as those for regenerating amprenavir and TMC-114. The highest ranking combinatorially generated structures for each of these compounds were evaluated using the full PBSA scoring function, and the correlation between these energies and those from experiment is shown in Figure 3-5B.

After changing the envelope definition and protease structure, there was some improvement in the correlation, as the scores for amprenavir (APV) and TMC-114 were now predicted to be better than those for the designed compounds. However, there was not much improvement in the relative ranking of the designed compounds, most likely due to their small binding energy differences. Overall, the improved correlation highlights the effect of the receptor model in correctly ranking inhibitor affinities. Although results were only presented for the TMC-114-bound protease structure, re-designing the compounds in a protease structure derived from a complex with amprenavir (PDB accession code 1HPV) [130], which contains the same scaffold as TMC-114, led to similarly improved correlation (data not shown). These results reinforce the importance of induced fit in correctly scoring inhibitor affinity [156].

3.3.5 Comparison of predicted to experimental structures

In addition to observing correlation between predicted and experimentally determined binding energies, it is important that the structural models used for scoring in computational design are predictive of the true experimental structures. To this end, four crystal structures were determined of complexes between wild-type HIV-1 protease

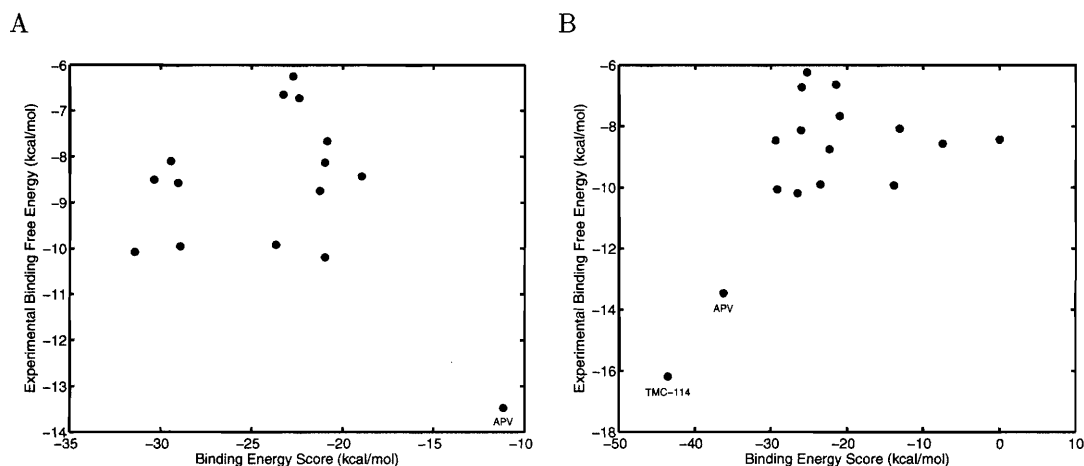


Figure 3-5: Comparisons between predicted and experimental binding affinities. In (A), the fifteen designed compounds with measured affinities, as well as the clinical inhibitor amprenavir (APV), were generated and scored using a substrate envelope inside a protease structure derived from a substrate complex. In (B), these compounds, in addition to the tight-binding inhibitor TMC-114, were designed and scored inside a “maximal” envelope that completely fills the active site of a HIV-1 protease structure derived from a complex with TMC-114. Generating and scoring compounds without the substrate envelope constraint and inside a structure bound to a similar scaffold improves the ability to differentiate between tighter and weaker binders. Experimental K_i values were converted to binding energies by assuming that $K_i = K_d$ and that $\Delta G = -RT \ln K_d$.

and the tightest binding designed inhibitors, Compounds 1, 3, 11, and 15. When comparing predicted to experimental structures, two structural models were available for each designed compound. The first was the initial structure designed against the substrate envelope in a substrate-bound protease used to propose compounds for synthesis, and the second was the structure used for improved energy function correlation designed without the substrate envelope constraint in the TMC-114-bound protease. Because a crystal structure is a representative low energy structure, it made sense to initially compare geometries against the structural prediction designed without envelope constraints.

In general, the structural agreement between the predicted (without envelope) and experimental structures was reasonably high (Figure 3-6). After aligning the protease structures, the root-mean-square deviation (RMSD) in coordinate positions for non-hydrogen atoms of the inhibitors were 0.8, 1.0, 0.9, and 1.2 Å for compounds 1, 3, 11, and 15 respectively. The major structural differences included a flip of the thiophene ring orientation in all four compounds, and misalignments of the positions for the R1 and R3 groups, especially in Compound 15. The prediction of the scaffold geometry and placement agreed well with the crystal structures for all compounds.

When the substrate envelope was superimposed on the crystal structures of the inhibitors, it was clear that the thiophene ring contained in all of the crystallized designed compounds protruded (Figure 3-7A). This finding is significant because it implies that a lower energy structure can be obtained if the inhibitors are allowed to leave the envelope. Consequently, the predicted structures for the inhibitors designed using the envelope constraint in a substrate-bound protease structure had far less similarity to the crystal (Figure 3-7B). In order to stay within the confines of the substrate envelope, the sulfonamide nitrogen inverts and the thiophene ring at R1 rotates almost 180° with respect to its linkage to the scaffold. This predicted geometry, where the substituents interacting with the P1 and P1' protease pockets

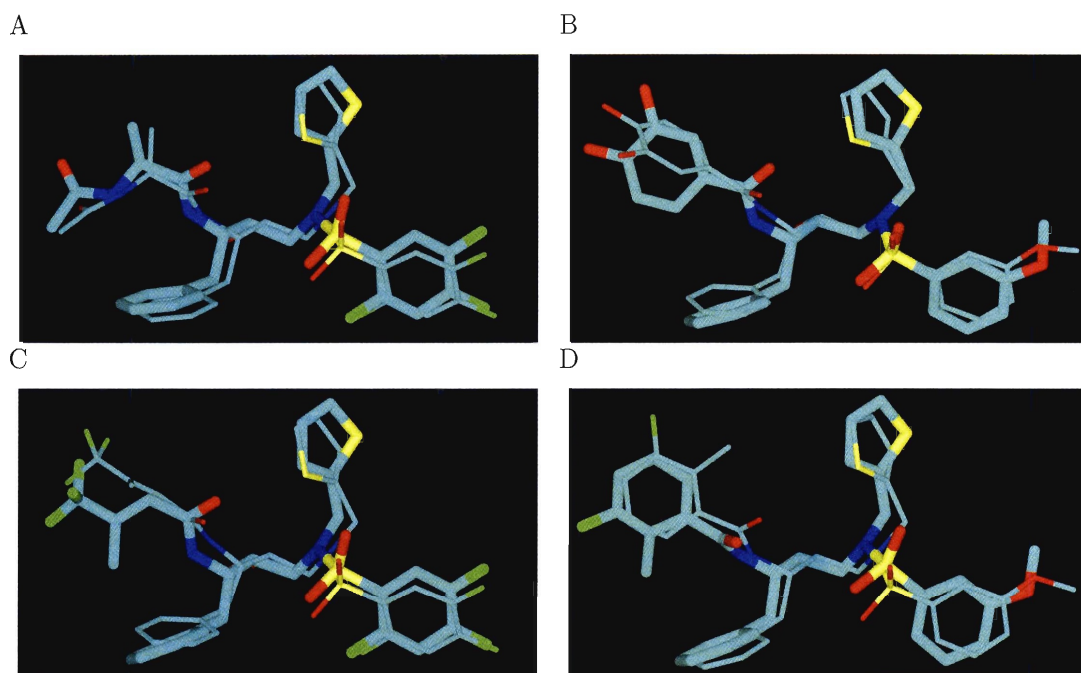


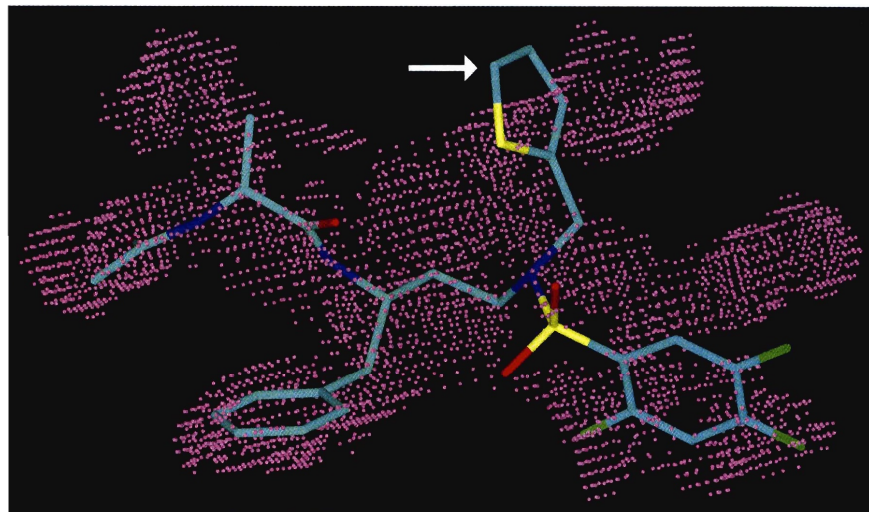
Figure 3-6: Comparison between predicted and experimentally determined structures for four of the tightest binding designed compounds, Compounds 1 (A), 3 (B), 11 (C), and 15 (D). The structure drawn in thicker bonds is the prediction, and the structure with thinner bonds is the crystallographic geometry. Green atoms are fluorine.

are pointing away from each other, is very reminiscent of the binding mode that the substrates utilize [44]. These findings suggest that the designed compounds may be able to retreat inside the envelope and adopt this alternate conformation with a small energy penalty, which may explain the broad specificity that the envelope compounds exhibit towards drug resistant proteases.

3.4 Conclusions and Future Work

In conclusion, this work establishes a framework for and begins to test the substrate envelope hypothesis as a design principle for the development of inhibitors less likely to induce resistance mutations in a drug target. Using HIV-1 protease as a model system, an inverse small-molecule design strategy was employed to design protease inhibitors predicted to have favorable binding energetics while remaining inside the substrate envelope. These compounds were combinatorially generated using a molecular scaffold

A



B

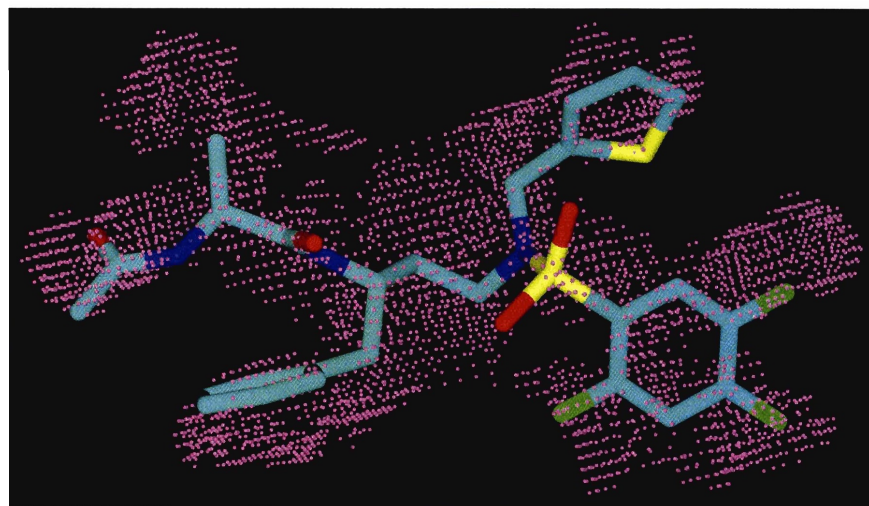


Figure 3-7: Experimental and predicted structures compared with the substrate envelope. In (A), the crystal structure of Compound 1 is superimposed on the substrate envelope (purple dots represent envelope sphere centers). The white arrow shows a portion of the molecule that protrudes from the envelope. In (B), the predicted structure for Compound 1, when designed with the substrate envelope constraint, uses an alternate conformation that stays within. This generates the hypothesis that the broad specificity of these compounds may be due to their ability to easily retreat inside the substrate envelope.

known to bind and stay within the HIV-1 protease substrate envelope, as well as a naive functional group library derived from chemical catalogs. Fifteen designed compounds were chemically synthesized and experimentally tested for inhibition of the wild-type HIV-1 protease. Four of these compounds had inhibition constants (K_i) between 30–50 nM, and these compounds were experimentally tested for binding against a panel of three drug resistant protease mutants. The designed compounds exhibited broad specificity against the drug resistant mutants, losing only 6–13-fold inhibition relative to wild-type in the worst case. This is in sharp contrast to the clinical inhibitor lopinavir (LPV), which loses over 1,200-fold in relative inhibition for a particular drug resistant protease.

Comparison of computed with experimental binding energies initially showed little correlation, even after attempting several modifications to the energy evaluation protocol. However, re-designing compounds in a “maximal” shape that filled the HIV-1 protease active site and utilizing a protease structure bound to an inhibitor containing a similar scaffold to designed compounds were sufficient to correctly rank the inhibitors amprenavir and TMC-114 ahead of the designed compounds. These results point to active site rearrangement and induced fit being important in the correct ranking of inhibitor binding affinity. In addition, a comparison of the predicted and experimental structures for the four tightest-binding designed compounds shows good agreement when the structural prediction is derived from design without the substrate envelope constraint. This was not surprising given that the crystal structure geometries for the designed compounds protrude from the envelope. The predicted structures for the compounds designed within the envelope in a substrate-bound protease structure were significantly different from the crystal geometries and more substrate-like, offering the hypothesis that these designed ligands may be able to retreat inside the envelope in response to resistance mutations.

Although the fact that these inhibitors show broad specificity to drug resistant

mutations is encouraging, it does not guarantee that these inhibitors will not be susceptible to alternative resistance pathways. One way to answer this question would be to use the designed envelope inhibitors as selection pressure in an accelerated evolution study of HIV-1 in cell culture [144,145], to see if it is possible to evolve resistance against the compounds. However, a practical factor limiting these experiments is that the binding affinity for the designed compounds is still significantly worse than the weakest binding clinical inhibitors. In general, the IC_{50} for HIV-1 protease inhibitors in cell-based assays is weaker than their K_i values in solution [145], making it challenging to work with weaker inhibitors in viral culture. Therefore, we are currently designing a second round of envelope inhibitors using a more protease-like set of functional groups and a substrate envelope built into the TMC-114 protease structure for improved scoring. Hopefully, this library will be enriched in tighter-binding inhibitors that still remain inside the envelope.

Another important feature of the work presented here is that the substrate envelope hypothesis has only been tested as a design principle in the positive direction, meaning that compounds designed to stay inside the envelope have been shown to have broad specificity against mutant proteases. It is also important to test the hypothesis in the negative direction, meaning that compound libraries specifically designed to protrude from the envelope yet are predicted to have high affinity should exhibit poor resistance profiles. Consequently, computational design efforts are currently being carried out to design such inhibitors.

3.5 Materials and Methods

3.5.1 HIV-1 protease structure preparation

Crystal structures of five inactivated HIV-1 protease-substrate peptide complexes were obtained from the Protein Data Bank (PDB) (Accession codes 1F7A, 1KJ7,

1KJF, 1KJG, and 1KJH) [43,44] in addition to a structure of HIV-1 protease bound to the inhibitor TMC-114 (Accession code 1T3R) [114]. Although a sixth substrate complex was available, it was not used in this study because of its poorer resolution. For structures with multiple occupancy, the first conformation was used in all calculations. In all structures, the terminal side-chain dihedral angle for asparagine, glutamine, and histidine residues was considered for a 180 degree rotation if it would visually improve the hydrogen bonding network. All water molecules were removed from the protease-substrate complexes except for five that were highly conserved across all structures [44]. In the TMC-114-bound structure, only the flap water was retained. Hydrogen atoms were added to all structures using the HBUILD module [116] in the CHARMM computer program [117] using the general CHARMM22 parameter set [115] and a distance-dependent dielectric constant of 4. In addition, any missing side-chain atoms were built back into the structures using CHARMM and the default geometry in CHARMM22. All ionizable residues were left in their standard states at pH 7.

For subsequent inverse design calculations, the protease from the RT-RH substrate complex (PDB accession code 1KJG) and the TMC-114 bound protease structure were used as design targets. In the case of the RT-RH-bound protease, the inactivating D25N mutations were reversed by building aspartate residues directly on top of the crystallographic asparagines. To create a singly protonated protease structure, the re-activated aspartate in the A chain was protonated on the O δ 2 atom such that it formed a hydrogen bonding interaction with Asp25B across the dimer interface. For both of these protease structures, PARSE radii and charges [22] were assigned for solvation and electrostatic calculations, while CHARMM22 parameters [115] were assigned for computing van der Waals interactions with potential ligands.

3.5.2 Preparation of scaffold and functional group libraries

A 3-D structure for the scaffold used for computational inhibitor design (Figure 3-2) was created using GAUSSVIEW 3 [118]. 3-D structures for 2,327 carboxylic acids and 379 primary amines were obtained from the ZINC database [149] in December 2004 for subsequent attachment to the R1 and R2 sites respectively. Compounds obtained from the ZINC database were limited to the vendors Sigma–Aldrich, Maybridge, and Ryan Scientific. 2-D structures for 274 sulfonyl chlorides were obtained for the R3 position directly from the Sigma–Aldrich and Maybridge catalogs in December 2004, as they were not present in the ZINC database at the time of retrieval. 3-D structures for the R groups needed to reconstruct amprenavir and TMC-114 were built by hand using GAUSSVIEW 3.

Structures were converted from 2-D to 3-D (if necessary), and 3-D structures were sampled in ring conformations, tautomeric states, and protonation states using the program LIGPREP from Schrödinger, Inc. At this stage, carboxylate groups were removed from members of the R1 library, amines were removed from members of the R2 library, and the sulfonyl chloride moiety was removed from members of the R3 library to facilitate attachment to the scaffold. In all cases, the removed groups were replaced by a hydrogen atom designated for attachment. As previously described (Chapter 2), geometries for all scaffolds and functional groups were optimized and partial atomic charges determined using quantum mechanical calculations, with conformational ensembles pre-generated for all functional groups.

3.5.3 Computational inverse inhibitor design to target the substrate envelope

Computational inverse inhibitor design was carried out as previously described (Chapter 2), with the following changes to the protocol. Rather than use a target ligand

shape that filled the entire binding site, inverse design was performed using an envelope based on the HIV-1 protease substrate structures. To generate the envelope, the five HIV-1 protease–substrate peptide complexes were simultaneously aligned on all C α atoms using the program PROFIT [157]. Spheres of radius 1.5 Å were placed on a cubic lattice of dimension 0.5 Å that surrounded the superimposed substrate peptides. Spheres were accepted if they were either within 1.0 Å of any substrate non-hydrogen atom, to create a loose envelope definition, or if they were simultaneously within 1.0 Å of non-hydrogen atoms in 3 substrates, to create a tighter envelope. In both cases, spheres were also accepted if they were within 3.5 Å of the side chain of Asn25 (inactivated catalytic residue) in either monomer to ensure that designed inhibitors would be permitted to interact with these non-mutable residues. These envelopes were placed back into a protease structure derived from a complex with the largest substrate peptide, RT–RH. The orientation of the envelope inside the protease was the same as the one used for substrate superposition.

When placing the scaffold inside the substrate envelope, two techniques were employed that differed from the previously described methods for rigid sampling of pre-generated scaffold conformations (Chapter 2). Because the scaffold used here had numerous rotatable bonds, it was not possible to generate a complete conformational ensemble ahead of time. Instead, the scaffold was broken into rigid pieces across rotatable bonds, and the largest (in this case the benzene ring) was sampled discretely throughout the envelope in translational steps of 0.25 Å and rotational steps in X, Y, and Z such that the arc length swept out by the atom furthest from the geometric center was 1.0 Å. Next, the placed fragments were extended into complete scaffold molecules by attaching the remaining rigid pieces sequentially while sampling about torsional degrees of freedom in 30-degree increments. Scaffolds were accepted if they remained inside the envelope by at least 0.5 Å, had a favorable grid-based van der Waals score, and were able to grow at least one functional group from the library at

each attachment site.

In an alternate procedure, scaffold placements were obtained by threading the scaffold on an experimentally determined binding mode. This mode was generated by aligning a protease structure bound to amprenavir (PDB accession code 1HPV) [130] to the substrate-bound protease, and only retaining atoms that corresponded to the scaffold. To perform threading, the scaffold was again broken into rigid pieces, and the largest was aligned to the known binding mode by minimizing the RMSD. Then, the remaining scaffold pieces were sequentially reattached, sampling torsional angles in 30-degree increments. After each attachment, each conformation was aligned to the known binding mode, minimizing the RMSD between the corresponding atoms. If the best RMSD fit, assuming all of the unplaced atoms were perfectly superimposed, was less than 1.0 Å, the fragment was accepted for another round of attachment. This procedure was repeated until a set of completed scaffolds was generated. Overall, this procedure generated all conformations for the scaffold such that the best-fit RMSD to the known binding mode was less than 1.0 Å. Each best-fit conformation was then finely sampled in translation and rotation, using three steps of 0.125 Å in X, Y, and Z translation and three steps of 0.25 Å arc length in X, Y, and Z rotation. Again, scaffold placements were accepted if all atoms were within the envelope by at least 0.5 Å, the molecules scored less than zero against the van der Waals grids, and at least one functional group library member could be grown from each attachment point.

During the process of computing self energies for functional group attachments, an additional filter was employed beyond those used in the standard inverse design method. Conformations for functional group attachments that contained polar atoms were eliminated from the search space if their polar atoms were buried and not involved in hydrogen bonding interactions with the protease. Polar atoms were considered buried if they were within 4.5 Å of any protease atom, and an acceptor atom and donor hydrogen were considered hydrogen bonded if their distance was less than

3.5 Å. If two polar atoms were bonded, such as in a hydroxyl group, only one needed to make a hydrogen bond if the group was buried. The addition of this filter was useful for two reasons. First, it significantly reduced the size of configurational space sent to the combinatorial search procedure, by pre-eliminating functional groups that would bury unsatisfied polar groups. Second, it attempts to minimize a previously identified shortcoming of this energy function (Chapter 2), where polar groups may be unfairly rewarded in van der Waals interactions for burial in a non-polar region.

3.5.4 Inverse inhibitor design for affinity prediction

To improve the correlation between predicted energies and experimental binding affinities, the inverse design methodology was repeated using a protease structure derived from a complex with the inhibitor TMC-114 (PDB accession code 1T3R) and a ligand envelope that filled the central four pockets of the HIV-1 protease active site (P2–P2'). This “maximal” ligand envelope was generated by initially packing the volume occupied by the known binding mode of TMC-114 with spheres, and then allowing the spheres to energy minimize with a van der Waals potential function as previously described for HIV-1 protease (Chapter 2). For the affinity evaluation, the functional groups allowed in the combinatorial search were limited to those needed to cover the designed compounds (Figure 3-3) and the inhibitors amprenavir and TMC-114.

All affinity predictions presented were calculated using a Poisson–Boltzmann / Surface Area (PBSA) binding energy model, which is the standard energy model used in the inverse design procedure for the final ranking of compounds. This energy function was computed as described previously (Chapter 2), and consists of explicit-atom van der Waals interaction, desolvation penalties and electrostatic interactions computed through solution of the linearized Poisson–Boltzmann equation, and a hydrophobic solvation term directly proportional to the surface area buried

upon binding. In the case that the same molecule was designed multiple times in the combinatorial search procedure, the best PBSA score was chosen to represent its energy.

To compute experimental binding free energies from inhibition constants (K_i), we assumed that the inhibition constant was equal to the dissociation constant ($K_d = K_i$), and used the relationship $\Delta G = -RT \ln K_d$.

3.5.5 Comparison of predicted to experimental structures

In order to compare predicted and experimentally determined structures, predicted complexes were superimposed on crystal structures by minimizing the RMSD between the coordinates of C α atoms in the protease using the program PROFIT [157]. After superposition, the RMSD between non-hydrogen atoms in the inhibitors was measured and the structures analyzed.

Chapter 4

Computational design and experimental study of tighter binding peptides to an inactivated mutant of HIV-1 protease ¹

Abstract

Drug resistance in HIV-1 protease, a barrier to effective treatment, is generally caused by mutations in the enzyme that disrupt inhibitor binding but still allow for substrate processing. Structural studies with mutant, inactive enzyme, have provided detailed information regarding how the substrates bind to the protease yet avoid resistance mutations; insights obtained inform the development of next generation therapeutics. Although structures have been obtained of complexes between substrate peptide and inactivated (D25N) protease, thermodynamic studies of peptide binding have been challenging due to low affinity. Peptides that bind tighter to the inactivated protease than the natural substrates would be valuable for thermodynamic studies as well as to explore whether the binding mode observed for substrate peptides is a function of weak binding. Here, two computational methods, charge optimization and protein design, were applied to identify peptide sequences predicted to have higher binding affinity to the inactivated protease. Of the candidate designed peptides, three were tested for binding with isothermal titration calorimetry, with one measured to have more than a ten-fold improvement over the tightest binding wild-type substrate. Crystal structures were also obtained for the three designed peptide complexes showing good agreement with computational prediction. Thermodynamic studies of the designed peptides show that binding is entropically driven, and structural studies show strong similarities between natural and tighter-binding designed peptide complexes, which may have implications in understanding the molecular mechanisms of drug resistance in HIV-1 protease and other rapidly evolving drug targets including those involved with other infectious agents or cancer treatment.

¹All experimental work presented in this chapter was performed by collaborators at the University of Massachusetts Medical School in the laboratory of C. A. Schiffer. Specifically, the calorimetry experiments were performed by E. A. Nalivaika, and the crystallography was performed by M. Prabu-Jeyabalan.

4.1 Introduction

Current treatment for HIV-1 infection has been increasingly limited by the growing trend of drug resistance [36, 40, 140]. One of the most common forms of resistance is mutations in the protease, an enzyme essential for the viral life cycle and a common target for drug treatment [40, 158–160]. One consistent feature of multi-drug resistant HIV-1 protease is that binding to inhibitors is significantly reduced while substrate binding and processing remains at viable levels [161, 162]. Therefore, in order to understand mechanisms of drug resistance and gain insight into how to design new drugs, it becomes important to understand how the substrate peptides themselves interact with the protease and can be catalytically cleaved by variants containing resistance mutations. Several recent studies have been conducted to address this issue, mostly through structural analysis of complexes between the protease and substrate analogs [163–165], or between inactive protease and actual peptides derived from cleavage sites [43, 44]. These studies have revealed structural motifs that may be used for substrate recognition [44], as well as compensatory structural changes in response to resistance mutations [164–167]. Besides structural analysis of substrate complexes, another useful tool in understanding drug resistance in HIV-1 protease has been to examine the thermodynamics of inhibitor binding to both wild-type and drug resistant protease using calorimetry [136, 168]. These studies support a connection between the enthalpy and entropy of binding with the ability for inhibitors to evade resistance mutations [45, 46]. A combination of these two experimental approaches, to determine the thermodynamics of binding for peptides in the context of the inactivated protease, may be very useful in learning more about how peptides interact with the protease.

Although the structures of six decameric substrate peptides have been solved in complex with an inactivated (D25N) protease [44], obtaining calorimetric measurements of peptide binding has been challenging, most likely due to poor affinity. The protease forms a homodimer that places a pair of aspartic acid residues (Asp25

from each monomer) at the active site. Although the charge state of this aspartyl dyad during substrate binding is uncertain [146–148], it is likely that the D25N inactivating mutation causes a net charge change of at least one charge unit, possibly accounting for the low affinity. Only one substrate peptide, derived from the reverse transcriptase–RNase H (RT–RH) cleavage site (wild-type P5–P5' sequence: GAETF*YVDGA), shows marginally detectable binding and low micromolar affinity for the inactivated D25N protease (Figure 4-3A). In order to obtain clean thermodynamic data, and to determine if there exist fundamental difference in the way weaker and tighter binding peptides interact with the protease, it became necessary to design tighter binding substrate-like peptides to the inactivated protease, using the RT–RH peptide–inactivated protease complex as an initial model.

Computational techniques are well suited for generating suggestions to improve affinity of protein–protein and peptide–protein interfaces; thermodynamic and structural studies of the binding of enhanced affinity peptides can provide insight into substrate binding, as well as into computational methodology. To this end, we applied two computational techniques to design tighter binding peptides to the inactivated protease. The first technique, charge optimization [23, 24, 30, 31, 101, 102, 169–173], has been shown to be a useful tool in identifying chemical groups that have poor electrostatic complementarity for their binding partner. Charge optimization, based on a linear response continuum electrostatic model, varies the partial atomic charges on a protein or peptide side chain so as to balance the cost of desolvation with favorable interactions made upon binding. The resulting charge distributions are optimal in that the electrostatic component of the computed binding free energy is maximally favorable. Mutations to improve binding can be suggested by identifying amino acid substitutions that have charge distributions similar to that of the global optimum, or by repeating the optimization under several charge constraints and selecting mutations that match those that are favorable.

One of the current limitations of charge optimization theory is that the concept makes sense only when the geometry of the bound and unbound states for the ligand being optimized are pre-determined, which is required to express the electrostatic binding free energy as a pairwise additive function of the partial atomic charges [24, 174]. Consequently, amino-acid substitutions suggested by charge optimization must also have a similar shape to the wild-type residue, in addition to a more optimal charge distribution. Because all amino-acid substitutions involve some degree of shape change, one reasonable way to evaluate a suggestion from charge optimization is to build the mutation using molecular mechanics and evaluate its computed change in binding free energy using the same continuum electrostatics model [173]. Although the charge optimization and building procedure does allow for flexibility in the designed side chain, it is still ultimately limited because the initial set of suggested mutations are derived from fixed geometry calculations.

In order to allow for side-chain flexibility in both the designed peptide and the protease active site, relax the requirement that suggested mutations are shape conservative, and to see how much of a difference these features make, we have applied a second computational technique to improve peptide affinity — namely, protein design. Computational protein design methods [67–69] have been successfully applied to stabilize protein folds [12, 15], create new protein folds [16, 17], and design protein interfaces [18–21]. Protein design techniques search a combinatorial space of discrete amino-acid identities and geometries in order to identify either global minimum or low energy sequences and structures with desired properties. In the case of designing tighter binding peptides to the inactivated protease, we allowed the RT–RH peptide side chains to change their amino-acid identity and conformation, protease active site residues to change their conformation only, and searched this combinatorial space using discrete search algorithms [11, 13, 14, 70] to identify global minimum energy complexes and an ordered list of weaker binding complexes sequentially from the

global minimum. These complexes were then examined for predicted improvements in binding affinity.

Three peptides predicted to bind tighter to the inactivated protease resulted from computational techniques and were tested experimentally using isothermal titration calorimetry (ITC). One peptide, containing a valine mutation at the P2 position of the wild-type RT–RH substrate, bound ten-fold tighter. Thermodynamic measurements of binding showed that peptide–inactive protease association is entropically driven, and that mutations predicted to improve peptide binding tended to increase the entropic contribution even further. Crystal structures were also obtained for each of the three designed peptide complexes, showing good agreement with calculation and substrate peptide complexes. Overall, agreement between both charge optimization and protein design along with experimental validation reinforces the usefulness of these computational methods in improving binding interfaces.

4.2 Results and Discussion

4.2.1 Electrostatic optimization of peptide binding

Charge optimization was applied to each RT–RH substrate peptide side chain independently, keeping the charges of all other peptide or protease side chains and all backbone atoms at their parameterized values. Four different sets of constraints were applied to the resulting charge distributions; in separate calculations the optimized side chain total charge was required to be negative ($-1e$), neutral ($0e$), or positive ($+1e$), and, as a control, all partial atomic charges in the optimized side chain were set to zero, which in all-hydrogen model is equivalent to a shape conserving hydrophobic replacement. The results of these charge optimization studies are summarized in Table 4.1. Charge optimization produced new partial atomic charge values that satisfy the sum-of-charges constraint for the side chain without modifying the shape, which

is a useful guide for design studies.

Out of the nine peptide residues visible in the structure, the four between the P3 and P1' positions show computed, idealized improvements greater than 1.0 kcal/mol upon charge optimization. In addition, these four residues can achieve a significant portion of this improvement by selecting an all-zero charge distribution, creating a hydrophobic isostere. This signifies that there is little energetic advantage to having strong dipolar groups or net charge, as the electrostatic interaction gained upon binding is generally unable to offset the cost of desolvating the side chain. This finding is consistent with the dominantly hydrophobic nature of the side-chain binding pockets of the protease. Overall, most residues best optimum corresponded to the same net charge as their standard state at pH 7, with the exception of Glu-P3 which is relatively indifferent to the net charge constraint. This is due to the fact that Glu-P3 is relatively solvent exposed in the bound state, with the $-1e$ optimized and $0e$ optimized charge distributions having a difference of only 0.9 kcal/mol in desolvation penalty, compared to a corresponding value of 9.5 kcal/mol for Thr-P2.

Given the energetics of charge optimization and the resulting partial atomic charge distributions, mutations were considered that have a similar shape but an altered charge distribution that closely matches optimality. The charge optimization results for the P3 and P2 positions are indeed suggestive of mutations that have a similar shape but with a more optimal charge distribution. The glutamate side chain at P3 shows possible improvements of 3 kcal/mol or greater upon optimization to a neutral or negative net charge. In addition, the energy of the hydrophobic isostere is a 2.4 kcal/mol improvement over the parameterized charges, indicative of a desolvation penalty that can not be fully recovered through direct electrostatic interaction with the protease upon binding. This result can be explained by the environment of the Glu-P3 side chain in the bound complex (Figure 4-1A), where the partially solvent exposed side chain is not making any electrostatic contacts with the protease. In the

case of the neutrally optimized side chain, the partial atomic charges depolarize by moving towards zero, and in the negatively charged case, charge tends to accumulate near the C γ atom and its bonded hydrogen atoms. The clustering of negative charge on these atoms is reflective of their proximity to the positively charged Arg8B side chain on the protease (Figure 4-1B). These results are suggestive of three mutations. The first is conversion to glutamine, which eliminates the net charge, has a less polar charge distribution, is very shape conservative, and is suggested by the large improvements from depolarization. The second is leucine, which is hydrophobic and has less shape conservation due to the change in hybridization at the branched carbon, but has a charge distribution similar to a hydrophobic isostere. The third is conversion to aspartate, which although very different in shape, may mimic the increase of negative charge near the C γ atom of the glutamate side chain upon optimization to a $-1e$ total charge.

At the P2 position, the wild-type threonine has a similar optimization profile to the glutamate at P3. The P2 side chain is surrounded by a hydrophobic pocket (Figure 4-1B), and the threonine hydroxyl does not make hydrogen bonding interactions in the crystal structure. As a result, both the hydrophobic isostere and 0e optimization results show significant possible improvements because the desolvation penalty of the hydroxyl can not be recovered in direct electrostatic interaction with the protease. The obvious candidate mutation at P2 is valine, which has a similar shape to threonine yet is very hydrophobic.

For the other two residues that show possible improvement greater than 1 kcal/mol, Phe-P1 and Tyr-P1', there are no clear candidate mutations. In both cases, conversion to a hydrophobic isostere is favorable, and optimization to a neutral side chain results in a charge distribution that depolarizes the phenyl ring even further than the parameterized charges, which consist of $-0.125e$ and $+0.125e$ dipoles along each C-H bond. For Phe-P1, no obvious mutation to a naturally occurring

Table 4.1: Changes in Electrostatic Binding Energy upon Charge Optimization of Peptide Side Chains

Residue	Position	$\Delta\Delta G_{\text{h}\phi}$	$\Delta\Delta G_{\text{opt}}^{-1}$	$\Delta\Delta G_{\text{opt}}^0$	$\Delta\Delta G_{\text{opt}}^{+1}$	Mutation
Ala	P4	0.00	+4.39	-0.23	+0.89	-
Glu	P3	-2.39	-3.45	-2.94	-0.56	Gln, Leu, Asp
Thr	P2	-1.11	+6.99	-1.27	+11.74	Val
Phe	P1	-0.78	+1.99	-1.30	+7.18	-
Tyr	P1'	-0.83	+0.37	-1.41	+2.62	-
Val	P2'	0.00	+8.86	-0.13	+10.75	-
Asp	P3'	+0.51	-0.08	+0.03	+3.10	-
Gly	P4'	0.00	0.00	0.00	0.00	-
Ala	P5'	0.00	-0.24	-0.05	+0.40	-

Changes in the electrostatic component of the binding free energy after optimization of the side-chain partial atomic charges. Results are shown in kcal/mol for constraining all partial atomic charges to zero (hydrophobic isostere, $\Delta\Delta G_{\text{h}\phi}$), as well as constraining the total residue charge to $-1e$, $0e$, and $+1e$. Negative free energy changes correspond to improved binding. For several positions, mutations suggested by the optimization data are provided.

amino acid exists that has these properties. In the case of Tyr-P1', one might consider conversion to Phe; however, most of the 0.6 kcal/mol improvement going from a hydrophobic isostere to an optimal neutral residue involves increasing the dipole of the hydroxyl group. This is due to a hydrogen bonding interaction between the hydroxyl group of Tyr-P1' and the side chain of Arg8A in the protease. Unfortunately, there is no naturally occurring amino-acid side chain that can maintain this favorable hydrogen bonding interaction, have a similar shape to tyrosine, and be otherwise less polarized than an aromatic ring. Therefore, the optimization results for the P1 and P1' positions do not suggest obvious mutations.

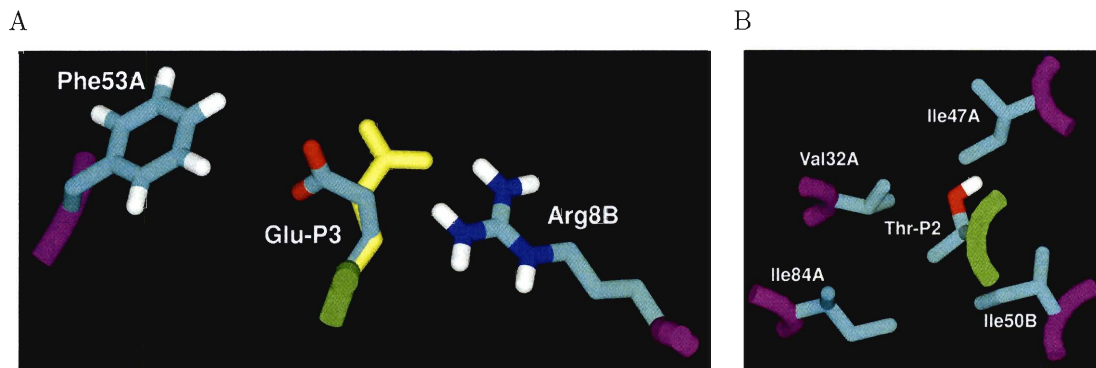


Figure 4-1: Molecular environments of the Glu-P3 (A) and Thr-P2 (B) peptide residues found to be suboptimal for electrostatic binding in the RT–RH crystal complex. The position of the P3 glutamate (A, atom colors, center) is pointed away from Arg8B and makes contact with Phe53A. Calculations suggest an alternative conformation (yellow) that makes better electrostatic interactions with Arg8B at the expense of packing. The wild-type P2 threonine residue (B, center) is situated in a pocket composed of four hydrophobic residues and makes no polar interactions.

4.2.2 Computed binding energetics of mutants suggested from charge optimization

Although the charge optimization methodology is extremely useful in identifying regions of poor electrostatic complementarity, it is limited in that it assumes the conformation of the side chains are rigid during the calculation, and any prediction based on the results is only accurate within this fixed shape assumption. Because all mutations involve some degree of shape change, the suggested mutations were explicitly modeled in order to further screen for improved binding candidates.

To this extent, the wild-type side chain as well as candidate mutations proposed in the previous section were built into the structure and their computed energetics were compared to that of the crystal, as summarized in Table 4.2. At the P3 position, rebuilding the wild-type glutamate using the same procedure applied to mutants led to an unexpectedly large energetic improvement over the crystal conformation, mainly due to its selection of an alternative conformation that made more favorable interactions with the side chain of Arg8B (Figure 4-1A). In this alternative conformation, improved electrostatic interactions were offset by worsened van der Waals

packing. The significance of this finding is that if this alternative conformation were truly populated, a mutation at this position would need to exceed this 1.5 kcal/mol improvement over the crystal structure in order to be a good candidate for observable tighter binding.

The mutation to glutamine at P3, irrespective of whether the side chain was built in its minimum energy conformation or directly on top of the crystal glutamate, led to a calculated improvement of about 1.5 kcal/mol over the crystal structure, mainly due to improved electrostatics as identified by charge optimization. However, if the wild-type glutamate can adopt the alternative conformation, mutation to glutamine is not predicted to be a significant binding improvement. Therefore, the mutation of glutamate to glutamine at P3 is a good choice for an experimental test, because if the binding energy were to remain the same, it would suggest that the wild-type glutamate residue may spend some of its time interacting more favorably with Arg8B in solution.

Additional mutations suggested by charge optimization at the P3 position include leucine and aspartate, which attempt to mimic the optimal charge distributions for a neutral and negatively charged side chain, respectively. Designing a leucine at P3 leads to a computed electrostatic benefit of 1.6 kcal/mol over the crystal structure, as expected from charge optimization, with nearly identical van der Waals interactions. This results in an energetic profile similar to that of the glutamine mutation, where a predicted improvement is dependent on the likelihood of the alternative wild-type glutamate conformation. Building and relaxing aspartate at this position led to greatly improved electrostatic interactions, due to proximity to Arg8B, but weakened van der Waals packing due to the reduced size of aspartate compared to glutamate. These effects canceled, leading to the same predicted binding energy as the crystal structure, and indicate that the aspartate mutation is not a good candidate for binding improvement.

At the P2 position, rebuilding of the wild-type threonine residue leads to similar energetics as the crystal, with a small tradeoff between electrostatics and packing. The suggested valine mutation at this position is computed to be favorable by 1.5–2.5 kcal/mol depending on whether the valine is rebuilt on top of the threonine or in an energy minimized geometry. This improvement is mainly accounted for by eliminating the desolvation penalty for the threonine hydroxyl group. Because valine has roughly the same shape as threonine, and is predicted to bind better in several built geometries, this mutation is an excellent candidate for improved binding.

Overall, building mutations into the structure is a useful tool to assess the feasibility of mutations suggested by charge optimization. In some cases, the building procedure highlights weaknesses of the rigid binding model used in charge optimization, as exemplified by the P3 Gln, P3 Leu, and P3 Asp suggested mutations, and in other cases it confirms its usefulness, such as the P2 Val mutation. Mutations that are good candidates for experimental validation are the shape conservative P3 Gln, computed to have improved binding if the crystal structure is the correct reference structure, serving as a good diagnostic for the interactions of the wild-type glutamate, and the P2 Val mutation, highly shape conservative and computed to be favorable in multiple geometries.

4.2.3 Improving binding through protein design

The charge optimization and building procedure presented above, although very useful for rapidly identifying side chains with poor electrostatic complementarity, is still ultimately limited by its fixed-shape approximation. Mutations initially suggested for building need to maintain shape similarity, and side chains on the protease and at peptide positions not being optimized and built must be kept in fixed conformation. To relax these requirements, we have applied a full computational protein design treatment that systematically considers all twenty common amino acids, ex-

Table 4.2: Binding Energies of Mutants Suggested From Charge Optimization

Residue	Position	Mutation	$\Delta\Delta G_{\text{elec}}^{\text{crystal}}$	$\Delta\Delta G_{\text{vdW}}^{\text{crystal}}$	$\Delta\Delta G_{\text{total}}^{\text{crystal}}$
Glu	P3	Glu _{min}	-3.2	+1.7	-1.5
		Gln _{min}	-0.8	-1.1	-1.9
		Gln _{crystal} ¹	-1.2	-0.2	-1.5
		Gln _{crystal} ²	-1.1	-0.5	-1.7
		Leu _{min}	-1.6	+0.1	-1.5
		Asp _{min}	-3.3	+3.2	0.0
		Thr	P2	Thr _{min}	+0.9
Val _{min}	-2.2	-0.1		-2.4	
Val _{crystal}	-0.7	-0.7		-1.5	

Estimated change in binding free energy for rebuilt wild-type and mutant structures suggested from charge optimization. All energetics are relative to the crystal structure, and are in kcal/mol. Negative free energies correspond to improved binding. Energy differences are shown for two rebuilding techniques, one where a minimized structure is obtained for the mutating residue (min), and an alternative method where the crystal structure dihedral angles are used to rebuild the mutant side chain (crystal). For the Glu to Gln mutation, there are two possible orientations using the crystal dihedrals. Energetics are shown for the electrostatic and van der Waals (vdW) components of the energy, as well as the total energy difference, including a surface area contribution.

cept proline, in multiple conformations for the central eight side chains (P4–P4′) of the peptide, while simultaneously considering active-site side chains in all rotamer combinations. Thus, this procedure permits side chain but not backbone relaxation to accommodate peptide mutations. Due to the large number of possible peptide sequences to consider, global minimal energy conformations were obtained for all single, double, and triple mutants from the wild-type RT–RH peptide sequence in a discrete conformational space, and their rigid binding energetics were computed to identify complexes with improved predicted binding properties.

When the wild-type RT–RH sequence was rebuilt using this protein design procedure, the same conformational heterogeneity observed at the P3 glutamate position in the charge optimization protocol was again identified. These two conformations for the P3 glutamate are separated by 0.5–2.0 kcal/mol depending on parameter set, and show the same energetic tradeoff between electrostatics and van der Waals packing upon the change of geometry to make closer interactions with Arg8B (Figure 4-1A). In addition, different parameter sets disagree about which state is more favorable, with van der Waals energies derived from PARAM19 favoring the crystal conformation, and those from united-atom AMBER favoring the alternative conformation. To factor out this uncertainty, the energetics for any sequence that does not involve a mutation at P3 was referenced against the wild-type structure with the same conformation for P3 Glu, and any structure involving a mutation at P3 was required to score better than both P3 conformations of the wild-type sequence. In all cases, sequences predicted to have improved binding were also required to score better than the reference in both electrostatics with solvation and van der Waals energies across multiple parameter sets. That is, given uncertainty in structure and some disagreement between models, we required all models to validate a prediction, which is a conservative criterion; because the balance between packing and electrostatic interactions is sometimes difficult to quantify, we further required improvements in both

terms simultaneously, which is again conservative.

The binding energetics for all single mutations to the RT–RH sequence are presented in Figure 4-2, showing the worst change in binding free energy across two different van der Waals parameter sets (Figure 4-2A) and two different electrostatics/solvation parameter sets (Figure 4-2B) relative to the appropriate wild-type reference structure. In terms of packing interactions, there exist many opportunities for improvement, especially at the P4 and P2 positions. Substitution for the solvent exposed wild-type alanine at the P4 position to larger amino acids was almost always an advantage for van der Waals' interactions due to their ability to pack against the side of the binding interface. The wild-type valine residue at P2 sits in a pocket that computations suggest can accommodate a slightly larger amino acid, and moderately sized amino acids provide a packing advantage. Other pockets such as P1, P1' and P2' appear also to accommodate larger amino acids for improved packing interactions.

The P3' and P4' positions show no opportunity for improvement; no side chain except the wild-type could be grown from the fixed backbone with a reasonable energy score. In the case of the P3' aspartate, mutation to any other side chain caused the energy of the bound complex to be at least 5 kcal/mol worse than that of the wild type, thereby excluding it from further consideration. At the P4' position, any amino acid besides the wild-type glycine caused a large van der Waals clash with another region of the fixed backbone.

Single mutations predicted to offer a binding advantage in electrostatics and solvation were much less numerous, with only one position, P2, offering any significant opportunities. Mutations to valine or isoleucine at P2 were predicted to be better electrostatically than the wild-type threonine, in excellent agreement with charge optimization calculations. Again, the valine and isoleucine mutations were computed to avoid the desolvation penalty associated with burying the threonine hydroxyl group in a hydrophobic pocket. Although no additional positions yielded significant solvation

improvements, there were several mutations that were computed to have similar electrostatics as wild type, for example, a serine mutation at P3, a glutamine mutation at P2, or a leucine mutation at P2'.

Van der Waals packing and electrostatics (with solvation) are important forces governing binding, but it is unclear that they contribute equally to successful molecular design. One may assign higher confidence to mutations that score well in both of these properties across multiple parameter sets, represented by substitutions that are white to blue in both Figures 4-2A and 4-2B. In Figure 4-2C, several of the sequences that scored best in electrostatics/solvation while still maintaining a favorable van der Waals binding free energy change are presented, along with their improvements over wild type in two parameters sets. Mutations at P2 to isoleucine and valine were predicted to be quite favorable, along with leucine at P2', while other mutations such as glutamine at P2 and asparagine at P4 were computed to be excellent for packing but slightly unfavorable for electrostatics/solvation.

Double and triple mutant sequences were also computationally screened for improved binding, and five of the top sequences in electrostatics/solvation that maintained good packing are also shown in Figure 4-2C. Overall, the double and triple mutants were additive in that their predicted improvements could be explained by adding together the benefit of their individual single mutations. This allowed less favorable mutations to be carried along into high ranking sequences by pairing them with the most favorable individual mutations, as was often observed in the triple mutant sequences. For example, one of the highest electrostatically ranked triple mutations, containing the favorable isoleucine mutation at P2 and leucine mutation at P2', permits the P3 aspartate mutation which, although deficient in van der Waals interaction, is electrostatically reasonable.

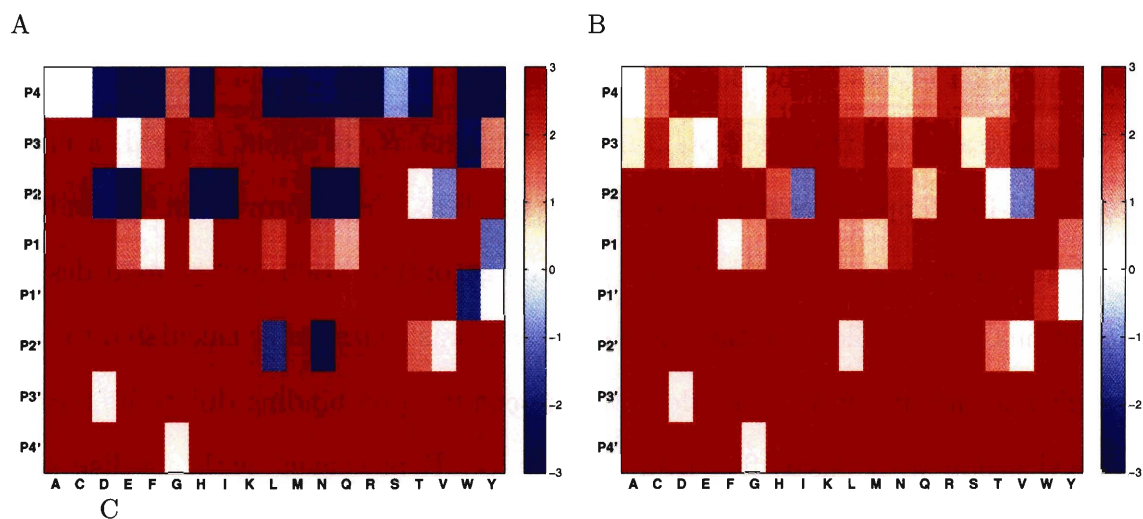
The mutations identified from protein design calculations served to support and extend those derived from charge optimization calculations. The protein design re-

sults recapitulated the ability to improve electrostatics at the P2 position by substituting the wild-type threonine with a valine mutation. In addition, they go beyond the fixed shape assumption and propose that isoleucine can achieve additional packing interactions at this position in addition to improving electrostatic complementarity. Although the protein design methods did not identify improvements at the P3 position, this was largely due to the difficulty in making predictions when the wild-type reference structure was in question, and the strict requirements placed on predicted improvements at this position. Protein design also identified several sites where improved packing interactions could be made, such as the P2' position, but did not discover any additional electrostatic improvements beyond those identified by charge optimization.

4.2.4 Experimental determination of binding energetics for designed peptides

In order to experimentally test the effectiveness of the computational design procedures, three designed peptides, one from protein design and two suggested from charge optimization, were tested along with wild-type RT–RH peptide for binding to the inactivated (D25N) protease using isothermal titration calorimetry (Figure 4-3A). These designed peptides include one of the high scoring triple mutants from protein design (Peptide 1), as well as the shape conserving T-P2-V mutation (Peptide 2) and diagnostic E-P3-Q mutation (Peptide 3) suggested from charge optimization. The triple-mutant peptide selected for synthesis and testing (Peptide 1) was computed to be less favorable for packing than other triple mutants (Figure 4-2C), most likely due to the aspartate mutation at P3. However, it was chosen because of the promising electrostatic profile of the Asp-P3 mutant (Table 4.2), and the computed optimality of the corresponding wild-type aspartate at P3'.

The wild-type RT–RH sequence has an observed dissociation constant of about



Energy Term	vdW	vdW	Elec/Solv	Elec/Solv
Parameter Set	PARAM19	AMBER UA	PARAM19	PARSE
P4–P4' Sequence				
<u>AE</u> <u>I</u> FYVDG	-4.0	-3.0	-1.4	-1.1
<u>AE</u> <u>V</u> FYVDG	-1.6	-0.8	-1.3	-1.0
<u>AET</u> FYLDG	-1.5	-2.2	-0.1	-0.2
<u>AEQ</u> FYVDG	-6.2	-4.4	+0.7	+0.2
<u>NET</u> FYVDG	-4.1	-2.3	+0.1	+0.6
<u>AE</u> <u>I</u> FY <u>L</u> DG	-5.5	-5.2	-1.5	-1.2
<u>AE</u> <u>V</u> FY <u>L</u> DG	-3.1	-3.0	-1.4	-1.1
<u>SE</u> <u>I</u> FYVDG	-5.2	-4.4	-1.7	-1.0
<u>SE</u> <u>I</u> FY <u>L</u> DG	-6.8	-6.5	-1.8	-1.1
<u>AD</u> <u>I</u> FY <u>L</u> DG	-1.2	-2.8	-1.0	-1.1
<u>TE</u> <u>I</u> FY <u>L</u> DG	-8.1	-7.5	-0.9	-0.9

Figure 4-2: Changes in binding free energy contributions predicted for mutants derived from protein design calculations. The worst-case changes across two van der Waals (vdW) (A) and two electrostatics/solvation (Elec/Solv) (B) parameter sets were computed for all single mutations (except proline) at each peptide position relative to the appropriate wild-type RT–RH sequence. Changes upon mutation in excess of +3 kcal/mol, or mutations ranked worse than +5 kcal/mol from the wild-type sequence in stability of the complex, were dropped from consideration in the protein design calculation and are represented as the darkest red. Sequences and energetics for several of the best electrostatically ranking single, double, and triple mutations are also presented (C), broken down by energy term and parameter set relative to wild type. Mutations to the sequence are underlined, relative energies are in kcal/mol, and negative numbers indicate predicted improvements to binding.

5 μM , and an overall poor experimental binding profile, marked by a lack of a clear binding transition (Figure 4-3B). In contrast, Peptide 2, corresponding to the T-P2-V mutation suggested by both charge optimization and protein design, shows a very sharp transition (Figure 4-3C), and an apparent K_d of about 0.5 μM , a more than a ten-fold improvement over wild-type RT–RH. The improvement in binding seen here supports the analysis from both charge optimization and protein design calculations, as the wild-type threonine side chain was consistently calculated to pay a significant and uncompensated desolvation penalty upon binding due to its buried hydroxyl group with unsatisfied hydrogen bonds. Replacement with a valine side chain eliminates this penalty while maintaining the same capability for packing interactions.

Peptide 1, a triple mutant derived from protein design, exhibited a more modest improvement of 2–3 fold over wild-type RT–RH. It contains a mutation to aspartate at P3, estimated to be better in electrostatics and worse in van der Waals than the crystal glutamate, and vice-versa for the alternative conformation, as computed from protein design. It also contains an isoleucine mutation at P2, which is functionally equivalent to the valine mutation in terms of electrostatics, but contains an additional methyl group expected to make better packing interactions. Finally, it contains a valine to leucine mutation at P2', which is computed to be equivalent electrostatically but to make better van der Waals interactions upon binding. Overall, it is difficult to attribute this modest binding improvement (less than expected) to any particular residue or combination of residues. Although the isoleucine and leucine mutations were consistently computed to be binding improvements, the aspartate's predictions were marginal and depended on the conformation of the wild-type residue at that position. Because aspartate is tolerated at the P3' position (as it is in the RT–RH peptide), it would be surprising for aspartate alone to be the explanation for only the small improvement in affinity.

Finally, Peptide 3, containing the diagnostic E-P3-Q mutation, has similar affinity to the wild-type peptide. The glutamate mutation was only predicted to be favorable by charge optimization if the structure could not relax to an alternative conformation where the wild-type glutamate makes better electrostatic interactions with Arg8B (Figure 4-1A). This experimental result suggests that the wild-type glutamate at P3 may interact more closely with Arg8 in solution than is observed in the crystal structure, making it more favorable than it would appear from looking at the crystal structure alone. This emphasizes the importance of building charge-optimization suggested side chains into the structure, because if a prediction were made from electrostatic optimization of the crystal structure alone, this mutation would have been expected to be a substantial binding improvement.

For all peptides studied, the thermodynamics of binding to the inactivated protease is dominated by a large favorable change in entropy, which is counterbalanced by an smaller but unfavorable change in enthalpy. In addition, all designed mutations served to increase the favorable contribution of entropy and worsen the enthalpic contributions to binding. This may not be surprising given that all of the suggested mutations serve to decrease the amount of polar surface area buried upon binding, and in some cases increase the total surface area buried. Because these trends tend to correlate with the hydrophobic effect and solvent release upon binding [87], it may not be surprising that solvent entropy plays a major role in the observed thermodynamics.

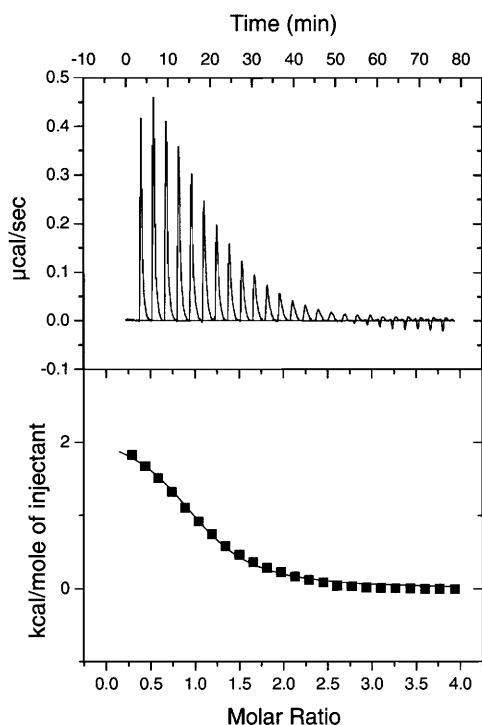
4.2.5 Crystal structures of designed complexes

After successfully designing peptides that bind tighter to inactivated HIV-1 protease, two important questions arise. The first is whether or not the structures predicted by the protein design methodology are an accurate representation of what is occurring experimentally. The second, more fundamental to the issue of resistance in HIV protease, is if there exist differences in the way that higher affinity peptides bind to

A

Name	P5-P5' Sequence	Suggested By	K_d (μM)	ΔG	ΔH	$-T\Delta S$
RT-RH	GAETFYVDGA	Wild-type	6.2 ± 2.1	-6.9 ± 0.19	2.4 ± 0.40	-9.3
Peptide 1	GAD <u>I</u> FYLDGA	Protein Design	2.2 ± 0.31	-7.4 ± 0.080	4.8 ± 0.048	-12.2
Peptide 2	GA <u>E</u> VFYVDGA	All Calculations	0.54 ± 0.20	-8.3 ± 0.22	4.7 ± 0.21	-12.9
Peptide 3	GAQTFYVDGA	Charge Optimization	6.4 ± 1.3	-6.8 ± 0.12	3.8 ± 0.49	-10.6

B (RT-RH)



C (Peptide 2)

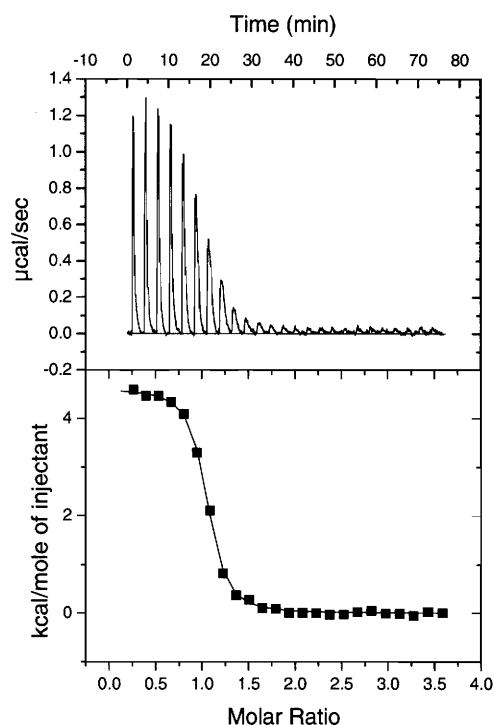


Figure 4-3: Isothermal titration calorimetry data for the binding of wild-type RT-RH and designed peptides to inactivated protease. In (A), the sequences of peptides tested (mutations underlined), their design origin, and their determined thermodynamic parameters of binding are shown. Units of the dissociation constant K_d are μM , and energies are in kcal/mol. Errors on thermodynamic parameters are derived from the fitting error after repeating the experiment at least three times. A comparison of the ITC traces for the wild-type RT-RH peptide (B) and Peptide 2 (C) shows that a sharp transition is only present for the tighter binding designed peptide.

the protease as compared to the natural substrates. To address these questions, the crystal structures of the three designed complexes were experimentally determined, with crystallographic statistics outlined in Table 4.3.

In order to determine whether there were fundamental differences between the way that the tightest-binding designed mutant (Peptide 2) and wild-type RT–RH peptides bound to the protease, a $C\alpha$ double difference plot was constructed to identify regions of the protein backbone that deviated between the two complexes (Figure 4-4). Overall, few regions exhibited significant deviations, and the largest differences were associated with surface protease residues such as Gly17, Gly51, and Cys67 where flips of the peptide bond geometry occurred. A closer examination of the residues lining the active site reveals that the backbone of Ile50 adopts slightly different relative positions in the two structures, most likely due to an alternative selection of side-chain rotamers. There is some slight positional changes between the A and B monomers and relative movement of the peptide N-terminus, however no evidence of large-scale, concerted conformational motion was found.

Although the backbone geometries of the protease and peptide remained fairly constant between the natural and tightest-binding complexes, there was a significant change in the geometry of the Glu-P3 side chain of the peptide. As described earlier for the wild-type RT–RH structure, the Glu-P3 side chain makes direct interactions with the side chain of residue Phe53A, even though computation suggested an alternative conformation that could make closer interactions with Arg8B (Figure 4-1A, yellow). However, in response to the Thr to Val mutation at P2 in the tightest-binding designed mutant (Peptide 2), the conformation of Glu-P3 adjusts to make a new hydrogen-bonding interaction with the side chain of Arg8B, in a similar fashion to the structure predicted computationally. These results suggest that the multiple conformations proposed by computation for the Glu-P3 residue were justified, and that the energetic barrier between these structures may be low. It is also possible that

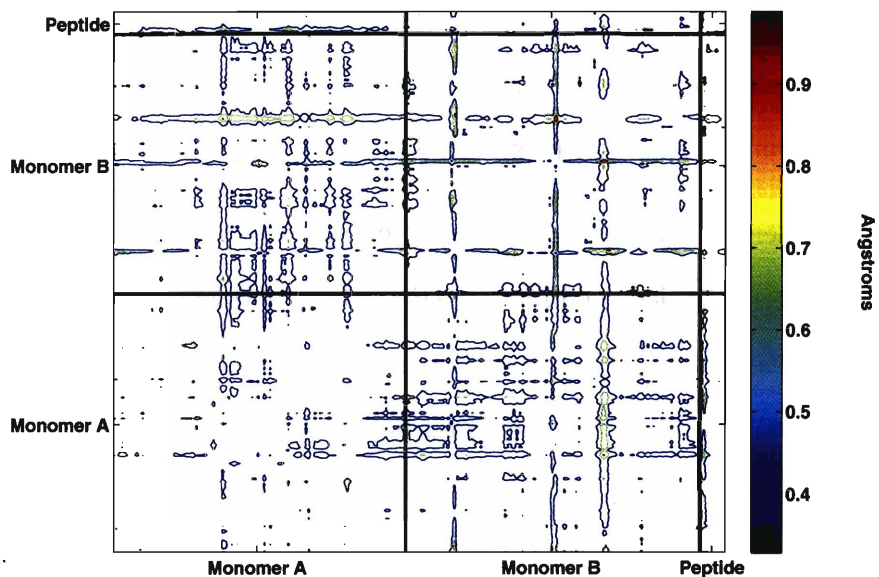


Figure 4-4: Double difference plot between $C\alpha$ atoms in the crystal structures of the wild-type RT-RH peptide complex and the tightest-binding designed complex (Peptide 2). The largest deviations correspond to the surface residues 17, 51, and 67 in monomer B.

the new hydrogen-bonding interaction plays a role in the improved affinity observed for Peptide 2, and may help to explain why the suggested Gln mutation at P3 might not be a binding improvement, as Gln-P3 is also capable of hydrogen bonding with Arg8B.

4.2.6 Comparison of predicted to experimental structures

The crystal structures for the designed complexes were also used to gauge the accuracy of the structural predictions from the protein design methodology. The structure of each designed complex was compared to the predicted structure by aligning the alpha carbons of the protease and examining the deviation between the predicted and experimental peptide side-chain geometries.

For reference, the crystal structure of the wild-type RT-RH peptide complex, from which all calculations were based, compares favorably with the wild-type structure as generated in protein design. For all but two of the side chains, the protein design

Table 4.3: Crystallographic statistics of the three RT–RH variant complexes. Statistics for the corresponding previously determined RT–RH complex [44] are also presented for comparison.

Parameter	Structure			
	WT RT–RH	Peptide 1	Peptide 2	Peptide 3
Peptide Sequence	GAETF*YVDGA	GADIF*YLDGA	GAEVF*YVDGA	GAQTF*YVDGA
<i>Data Collection</i>				
Resolution (Å)	2.0	2.0	2.0	2.25
Temperature	RT	Cryo	Cryo	Cryo
Space Group	P212121	P212121	P212121	P212121
<i>a</i> (Å)	51.3	51.2	50.9	51.0
<i>b</i> (Å)	58.8	57.6	57.5	58.5
<i>c</i> (Å)	62.0	61.5	61.9	61.7
Z	4	4	4	4
Total Reflections	60432	53438	54458	30202
Unique Reflections	13019	11909	12332	8523
R _{merge} (%)	7.0	3.4	7.1	6.5
Completeness (%)	98.6	92.7	95.8	91.0
I/σ _I	7.6	21.2	12.5	10.3
<i>Crystallographic Refinement</i>				
R value (%)	18.4	19.5	16.0	18.7
R _{free} (%)	22.6	25.7	21.6	24.4
Sigma Cutoff	None	None	None	None
RMSD in:				
Bond Lengths (Å)	0.006	0.007	0.007	0.008
Bond Angles (°)	1.3	1.2	1.5	1.3
PDB Code	1KJG	N/A	N/A	N/A

methodology selected the crystal conformation as optimal. For the two remaining peptide side chains, P3 Glu and P3' Asp, conformations were selected from the rotamer library that were close to the crystal geometry (Deviations — P3 Glu: χ_1 10°, χ_2 12°, χ_3 43° P3' Asp: χ_1 23°, χ_2 56°) (Figure 4-5A). In addition, the computed structure for the protease remained very close to that of the crystal. Overall, the total root-mean-square deviation (RMSD) for designed side chains between the crystallographic and designed complexes for the wild-type RT–RH sequence was 0.6 Å.

The computationally predicted structure for Peptide 2 agrees very well with the experimental structure, with a negligible difference in side-chain geometry. As expected, the valine mutation is directly superimposed on the wild-type threonine (Figure 4-5B), confirming the hypothesis that this is a shape-conserving mutation. Although Peptide 3 did not show an improvement in binding, it was not necessarily due to an error in modeling the bound state as the predicted and experimental conformations of the P3 glutamine mutation are very similar. The side-chain dihedral angles are virtually identical, and the only differences in their placement are due to small changes in the backbone geometry (Figure 4-5C).

A comparison of the predicted and experimental structures for Peptide 1 shows less agreement. Although much of the side-chain density is missing from the crystal structure of the Peptide 1 complex, two out of the three designed side chains are visible. The modeled isoleucine mutation at the P2 position differs from the crystal by approximately 30 degrees in the first χ angle for an overall side-chain RMSD of 0.8 Å (Figure 4-5D). Although a rotamer closer to the crystal structure did exist in the search space, its energy was predicted to be only 0.2 kcal/mol higher than that of the selected conformation, indicative of a relatively flat energy landscape. Mutation at the P2' position to leucine was also visible in the structure, and its predicted geometry was somewhat different from the experimental observation. The first χ angle deviates by 41 degrees as the crystal side chain adjusts to make closer

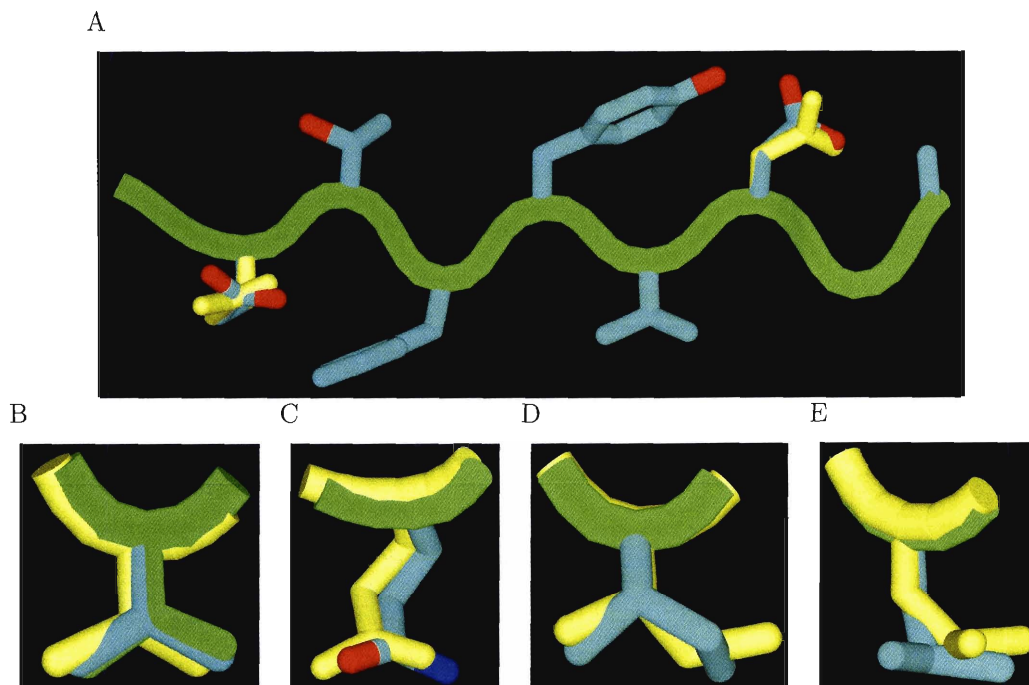


Figure 4-5: Comparison of predicted to experimentally determined structures. For reference, comparisons were made between the designed (atom colors) and crystal structure (yellow) for the wild-type RT–RH peptide sequence (A). Crystal rotamers were selected by design at all but two residues. Designed mutation to valine at P2 in Peptide 2 (atom colors) has good structural agreement with both the crystal structure of the mutant (yellow) and the wild-type threonine (green) in (B). The glutamate to glutamine structural prediction at P3 (atom colors) for Peptide 3 also agrees well with its experimental structure (yellow) (C). Mutations at P2 to isoleucine (D) and P2' to leucine (E) in Peptide 1 agree less favorably.

packing interactions with the backbone of Asp30B, rather than the Ile50A side chain as predicted by protein design (Figure 4-5E). In the rotamer library used for design, there did exist a leucine conformation slightly closer to the observed crystal structure (20 degrees closer in χ_1), but it was not selected because packing interactions with the Ile50A side chain were better in the chosen geometry.

4.3 Conclusions

Two computational techniques to improve binding affinity were successfully applied to design peptide sequences that bind tighter to inactivated (D25N) HIV-1 protease

in an effort to obtain thermodynamic and structural data to aid in the understanding of substrate specificity and drug resistance. Charge optimization techniques identified two residues in the tightest binding natural substrate peptide (RT–RH), Glu-P3 and Thr-P2, that provided opportunities to improve electrostatic complementarity, suggesting several mutations to enhance binding or serve as diagnostics for elucidating energetic contributions of the wild-type residues. Computational protein design techniques, which allow for side-chain flexibility and shape changing mutations, confirmed the predictions from charge optimization, demonstrated the limitations of charge optimization’s fixed geometry model, and suggested additional substitutions to further improve packing interactions at the interface.

Three designed peptides were tested for inactive protease binding through isothermal titration calorimetry. A single threonine to valine substitution at the P2 position, heavily suggested by calculation, led to more than a ten-fold improvement in binding, which is a considerable improvement given the small perturbation. In addition, a triple mutant sequence derived from protein design led to a 2–3 fold improvement. Substitution of glutamine for glutamate at the P3 position suggests that the wild-type residue may be more favorable than calculated from the crystal structure alone. It is important to note that the three designed peptides presented here were the only peptides tested experimentally, and that only the data presented in this work was used to select the designed peptides.

Thermodynamic measurements of peptide binding revealed that binding to the inactive protease was heavily entropically driven, most likely due to release of solvent from the active site upon binding. Mutations suggested by computation tended to replace polar groups with less polar groups and increase the total amount of buried hydrophobic surface area, resulting in a further increase in the entropic contribution. These results may be important in understanding resistance, as peptides that bind primarily through the hydrophobic effect without making strongly orientation depen-

dent interactions may be better suited to adapt to resistance mutations where less flexible small-molecule inhibitors cannot.

Crystal structures of the designed peptide complexes, in addition to validating structural predictions from computation, revealed only minor differences between complexes of the natural RT–RH substrate and the tightest-binding designed peptide (Peptide 2). These findings suggest that using the geometries of natural substrates is sufficient to define envelopes that represent the consensus volume of substrates, which may be useful in subsequent drug design efforts [44, 136–139].

An additional application for tighter binding peptides is in the understanding of a more recently identified mechanism of drug resistance against protease inhibitors, substrate co-evolution [167]. In cases of highly drug resistant protease, compensatory mutations can arise in the cleavage sites to improve catalytic efficiency. One of the most common of these substrate mutations occurs in the nucleocapsid–p1 site (NC–p1), where a wild-type alanine in the P2 position is converted to valine in response to a V82A drug resistance mutation in the protease. This mutation is analogous to the valine P2 mutation predicted and tested to be favorable in the RT–RH background, supporting a hypothesis that substrate co-evolution may serve to generally increase peptide–protease affinity.

Overall, computational design methods have been shown to be a useful tool in designing tighter binding peptides, providing sequences that are useful in helping to understand how peptides bind HIV-1 protease, their thermodynamic contributions to binding, and rationalizing the use of natural substrate structures as models for future drug development.

4.4 Materials and Methods

4.4.1 Protein structure preparation

The crystal structure of inactivated (D25N) HIV-1 protease complexed with RT–RH substrate peptide [44] (Protein Data Bank, ID 1KJG) was used as the starting point for all calculations. For residues with multiple occupancy, the first conformation was selected except for residues Glu 21A and Glu 35A where the second conformation was selected to improve the hydrogen-bond network. In addition, the ring of His69B was flipped 180° to improve hydrogen bonding. Because the substrate peptide used in the crystal structure had blocked termini, acetyl and methylamide groups were added to the N and C termini of the peptide respectively. All water molecules were removed from the structure except for the five waters conserved across all protease–substrate complexes [44]. Hydrogens were added to the structure using the HBUILD module [116] in the CHARMM computer program [117] with the PARAM22 all-atom parameter set [175]. All ionizable residues were left in their standard states at pH 7, with histidines neutral and δ -protonated.

4.4.2 Continuum electrostatic and charge optimization calculations

All continuum electrostatic calculations were performed using a locally-modified version of the DELPHI computer program [25–27, 121] to solve the linearized Poisson–Boltzmann equation. A dielectric constant of 4 was used for the protein and 80 for water, as well as an ionic strength of 145 mM in the solvent region. The boundary between protein and solvent was represented by a molecular surface computed with a probe radius of 1.4 Å, and an ion-exclusion layer of 2.0 Å surrounded all molecules. Partial atomic charges and radii for continuum electrostatic calculations were taken from the PARSE parameter set [22], and structures derived from protein design calcula-

tions were additionally screened using charges and radii from the PARAM19 parameter set [176,177]. Electrostatic binding and solvation contributions for protease-peptide complexes designed from charge optimization were calculated using a 257x257x257 cubic grid and a two-stage boundary condition focusing scheme. The complex was placed on the grid such that it occupied first 23% and then 92% of the grid size, where the solved potentials from the lower percent fill were used as boundary conditions for the higher percent fill calculation. This resulted in a final grid spacing of 0.23 Å. In addition, each result was averaged over 10 translations of the complex relative to the grid in order to reduce artifacts from the grid-based representation of atomic charges and the molecular surface. Complexes from protein design calculations were evaluated similarly, except a 129x129x129 grid and only one translation were used due to the large number of structures that needed evaluation. The matrix elements for electrostatic optimization were also computed in a similar fashion, but using a 129x129x129 grid, 10 translations and a three-stage over-focusing scheme of 23%, 92% and 184% percent where the 184% calculation was centered on the single atom charged when computing a row of the ligand desolvation matrix.

Charge optimization was performed as previously described [23,24,30,31,101,102,169,171,173], using locally written software. Each side chain of the peptide had its partial atomic charges independently optimized to maximize the rigid electrostatic binding free energy of the peptide to the inactivated protease. Each peptide side chain was constrained to net charges of $-1e$, $0e$ and $+1e$, as well as requiring that no atom exceed a charge magnitude of $0.85e$. In addition, each side chain was constrained to have all zero partial atomic charges, thus creating a hydrophobic isostere. For each constraint set, the energy difference between the optimal and parameterized charge distribution was determined by using the resulting charges with the optimization objective function [24,102,173]. All constrained optimizations were carried out using the software package LOQO [178,179].

4.4.3 Modeling of peptide mutations from charge optimization

Peptide mutations suggested from charge optimization were built into the RT–RH peptide complex structure using the CHARMM computer program [117] and the PARAM22 all-atom parameter set [175], using a distance-dependent dielectric constant of $4r$. Two procedures were used to model each mutant structure. The first aimed to find a low-energy structure for each mutation, and all mutant side-chain dihedral angles were enumerated every 30 degrees. Each enumerated side-chain conformation was then minimized until convergence, with the rest of the protease and peptide held fixed. The lowest energy minimized structure was selected to represent the mutation. Alternatively, because charge optimization suggests shape-conserving mutations, the mutant side chain was built on top of the wild-type crystal structure using the dihedral angles common to both side chains.

4.4.4 Protein design calculations

Through the use of locally developed protein design software, the central eight positions of the peptide (P4–P4′) were allowed to mutate to and select a conformation for any of the naturally occurring amino acids except proline. Histidines were modeled with all three possible protonation states. In addition, the side chains of active-site protease residues 8, 23, 25, 29, 30, 32, 45, 47, 50, 53, 82 and 84 in both monomers were allowed to change their conformation but not their identity. In all cases, the protein backbone was held fixed. Side-chain conformations were modeled discretely from the backbone independent rotamer library of Dunbrack and Karplus (May 2000) [180], but with additional rotamers ± 10 degrees about the first two χ angles. In addition, the exact structure of the crystal side chain was permitted as a choice in the design. The objective function minimized in the protein design proce-

ture was the energy of the bound complex. A model of the peptide unfolded state was not used in these calculations because short peptides tend not to form stable structures [181]. The energy of the bound complex was estimated using an energy function pairwise additive in side-chain conformations, consisting of an electrostatics term with a distance-dependent dielectric constant of $4r$, a van der Waals interaction term, and internal dihedral energy as computed by CHARMM [117] using a modified version of the PARAM19 parameter set [176,177] where aromatic and sulfhydryl hydrogens are included. For each single, double, and triple mutant peptide sequence, the global minimum energy conformation (GMEC) was identified for the complex using dead-end elimination (DEE) [11,14,70] and A* [13]. These structures were then screened for predicted improvements in binding as described below.

4.4.5 Calculation of binding energetics

The binding free energy of peptide-inactivated protease complexes was estimated using a rigid binding model, where the difference between solvation energy in the bound and unbound state was added to the vacuum ($\epsilon = 4$) interaction energy in the bound state. Vacuum binding energies were computed with CHARMM [117], using the PARAM22 parameter set [175], for mutants derived from charge optimization. For mutants suggested from protein design, a modified version of the united-atom PARAM19 parameter set [176,177] where aromatic and sulfhydryl hydrogens were added was used for vacuum energies. In addition, mutants from protein design were also screened with the AMBER united-atom parameter set [182]. Solvation energies were computed using a Poisson-Boltzmann/Surface Area (PBSA) model [22]. The linearized Poisson-Boltzmann equation was solved with a locally-modified version of DELPHI as described above, and CHARMM was used to compute the analytical surface area. PARSE [22] radii and charges were used for all structures, and designed complexes from protein design were additionally screened using PARAM19 radii and

charges. To compute the hydrophobic contribution to solvation, the surface area was multiplied by 5 cal/mol/Å² [22]. In order to avoid predicted binding improvements from depending on possible scaling problems between the electrostatic/solvation and van der Waals components of the estimated binding energy, mutations considered as improvements were required to be better than wild-type in both of these components of the energy function.

4.4.6 Substrate peptides

Decameric peptide variants of the HIV-1 reverse transcriptase-RNase H substrate (RT–RH) with sequences GADIF*YLDGA (Peptide 1), GAEVF*YVDGA (Peptide 2), and GAQTF*YVDGA (Peptide 3) were chosen as previously described [44]. The peptides were purchased from 21st Century Biochemicals, Marboro, MA.

4.4.7 Isothermal titration calorimetry

Thermodynamic parameters of peptide binding were determined using an isothermal titration calorimeter, VP-ITC (MicroCal Inc., Northampton, MA). The buffer used for all protease and peptide solutions consisted of 10 mM sodium acetate pH 5.0, 2% DMSO, and 2 mM Tris(2-carboxyethyl) phosphine (TCEP). A 1 mM solution of each peptide was directly titrated into a solution of 30–70 μM D25N HIV-1 protease in 10 μl aliquots. The experiments were performed at 15° C. Each experiment was repeated at least in triplicate. Data were processed using the Origin 7 software package from MicroCal.

4.4.8 Mutagenesis, protein purification, and crystallization

The previously described method [167,183] was followed throughout all wet chemistry experiments. A synthetic gene of HIV-1 protease, optimized for *Escherichia coli*

codon usage was used as the starting template for mutagenesis to introduce D25N substitution. This isosteric mutation inactivates the protease, and it was made using the QuickChange Site-Directed Mutagenesis Kit (Stratagene, La Jolla, CA). The purified protease, in 50% acetic acid was then refolded by rapid 10-fold dilution into a mixture of 0.05 M sodium acetate (pH 5.5), 10% glycerol, 5% ethylene glycol, and 5 mM dithiothreitol (refolding buffer) kept over ice. The diluted protein was concentrated and dialyzed to remove residual acetic acid. Protease used for crystallization was further purified using a Pharmacia Superdex 75 FPLC column equilibrated with refolding buffer. Crystals were grown by the hanging drop, vapor diffusion method. Stock solutions (25 mM) of substrate peptides used for co-crystallization were prepared in dimethyl sulfoxide. The protein concentration was approximately 1 mg/ml. Small crystals started appearing after 3–7 days with the longest length between 0.1 and 0.2 mm.

4.4.9 Crystallographic data collection

The crystals were flash frozen over a nitrogen stream and intensity data were collected on an in-house Rigaku X-ray generator with a R-axis IV image plate. Frames were indexed using DENZO and scaled using SCALEPACK [184,185]. Complete data collection statistics are listed in Table 4.3.

4.4.10 Structure solution and crystallographic refinement

The crystal structures were solved and refined using the programs within the CCP4i interface [186]. Structure solution was carried out by molecular replacement program using AMORE [187]. The crystal structure of wild-type protease variant complexed to the inhibitor TMC-114 (PDB accession code 1T3R) [136] was used as the starting model. The molecular replacement phases were further improved by using ARP/wARP [188] by building solvent molecules into the unaccounted regions

of electron density. Subsequently, interactive model building was carried out using O [189]. Initial **2Fo-Fc** and **Fo-Fc** maps indicated the positions of RT–RH substrate variants. Conjugate gradient refinement using REFMAC5 [190] was performed by incorporating Schomaker and Trueblood tensor formulation of TLS (translation, libration, screw-rotation) parameters [191–193]. The working R (R_{work}) and its cross validation (R_{free}) were assessed using PROCHECK [194] at the end of each refinement round. The refinement statistics are also shown in Table 4.3.

4.4.11 Analysis of and comparison to experimental structures

The double difference plot comparing the structures of the tightest-binding mutant and wild-type peptide complexes was computed by determining the interatomic distances between all pairs of $C\alpha$ atoms in each structure. The absolute value of the difference between these two distance matrices was subsequently calculated.

Crystal structures for the designed complexes were compared to predicted structures by performing a best RMSD fit between all $C\alpha$ atoms of just the protease using the program PROFIT [157]. The resulting RMS difference between peptide side-chain conformations was subsequently calculated.

Chapter 5

FFTSVD: A Fast Multiscale Boundary Element Method Solver Suitable for BioMEMS and Biomolecule Simulation ¹

Abstract

We present a fast boundary element method (BEM) algorithm that is well-suited for solving electrostatics problems that arise in traditional and Bio-MEMS design. The algorithm, FFTSVD, is Green's function independent for low-frequency kernels and efficient for inhomogeneous problems. FFTSVD is a multiscale algorithm that decomposes the problem domain using an octree and uses sampling to calculate low-rank approximations to dominant source distributions and responses. Long-range interactions at each length scale are computed using the FFT. Computational results illustrate that the FFTSVD algorithm performs better than precorrected-FFT style algorithms or the multipole style algorithms in FastCap.

¹This chapter has been previously published as:

M. D. Altman, J. P. Bardhan, B. Tidor, and J. K. White. FFTSVD: A fast multiscale boundary-element method solver suitable for bio-MEMS and biomolecule simulation. *IEEE T. Comput. Aid. D.*, 25:277–284, 2006.

Copyright 2006, IEEE. Published in the IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works, must be obtained from the IEEE.

5.1 Introduction

Microelectromechanical systems (MEMS) have recently become a popular platform for biological experiments because they offer new avenues for investigating the structure and function of biological systems. Their chief advantages over traditional *in vitro* methods are reduced sample requirements, potentially improved detection sensitivity, and structures of approximately the same dimensions as the systems under investigation [195]. Devices have been presented for sorting cells [196], separating and sequencing DNA [197], and biomolecule detection [198]. Furthermore, because arrays of sensors can be batch fabricated on a single device, parallel experiments and high-throughput analysis are readily performed. However, since microfabrication is relatively slow and expensive, numerical simulation of MEMS devices is an essential component of the design process [199, 200]. Design tools for integrated circuits cannot address multiphysics problems, and this has motivated the development of several computer-aided MEMS design software packages, most of which are based on the finite element method (FEM) and the boundary element method (BEM) [201].

BioMEMS, when applied to such problems as biomolecule detection, are often functionalized with receptor molecules that bind targets of interest [202]. Molecular labels can also be used to aid in the detection process [203]. However, the interactions between these molecules, the MEMS device, and the solvent environment are often neglected during computational prototyping. In other fields, such as computational chemistry and chemical engineering, continuum models of solvation are often used to study the electrostatic component of these interactions [29]. These mean-field models permit the efficient calculation of many useful properties, including solvation energies and electrostatic fields [25, 204], and have been shown to correlate well with more expensive calculations that include explicit solvent [205]. However, continuum models are unable to resolve specific molecular interactions between solvent molecules and the solute. A variety of numerical techniques can be used to simulate the continuum

models, including the finite difference method (FDM), the finite element method (FEM), and the boundary element method (BEM) [26, 206, 207].

The boundary element method has a number of advantages relative to FDM and FEM, such as requiring only surface discretizations and exactly treating boundary conditions at infinity. However, discretizing boundary integral equations produces dense linear systems whose memory costs scale as $O(n^2)$ and solution costs scale with $O(n^3)$, where n is the number of discretization unknowns. This rapid rise in cost with increasing problem complexity has motivated the development of accelerated BEM solvers. Preconditioned Krylov subspace techniques, combined with fast algorithms for computing matrix–vector (MV) products, can require as little as $O(n)$ memory and time to solve BEM problems [208]. Many such algorithms have been presented, including the fast multipole method (FMM) [209] \mathcal{H} -matrices [210–212], the precorrected-FFT method [49], wavelet techniques [213, 214], FFT on multipoles [215, 216], kernel-independent multipole methods [217, 218], the hierarchical SVD method [50, 219], plane-wave expansion based approaches [220], and the PILOT algorithm [221]. Some algorithms, such as the original FMM, exploit the decay of the integral equation kernel; the precorrected-FFT method makes use of kernel shift-invariance. This paper introduces an algorithm that combines the benefits of both of these approaches, leading to a method that has excellent memory and time efficiency even on highly inhomogeneous problems.

Fast BEM algorithms whose structures depend on kernel decay suffer from a common, well-known problem: computing medium- and long-range interactions is still expensive, even when their numerical low rank is exploited. For instance, in the fast multipole method, computing the M2L (multipole to local) products dominates the matrix–vector product time, since each cube can have as many as 124 or 189 interacting cubes, depending on the interaction list definition, and the work per M2L multiplication scales as $O(p^4)$, where p is the expansion order and is related

to accuracy [47, 209]. Much work has focused on reducing this cost; for the FMM, plane-wave expansions [220, 222] diagonalize the M2L translation, but are typically only efficient for large p . The precorrected-FFT (pFFT) algorithm [49] relies on not the kernel’s decay but rather its translation invariance to achieve high efficiency. The pFFT method is Green’s function independent, even for highly oscillatory kernels. Consequently, the method has been applied in a number of different fields, including wide-band impedance extraction [223], microfluidics [224–226] and biomolecule electrostatics [227]. One weakness of the precorrected-FFT method is that its efficiency decreases as the problem domain becomes increasingly inhomogeneous [49].

In this paper, we introduce a fast BEM algorithm called FFTSVD. The method is well-suited to MEMS device simulation because it is Green’s function independent and maintains high efficiency when solving inhomogeneous problems. The FFTSVD algorithm is similar to the PILOT algorithm introduced by Gope and Jandhyala [221], in that our algorithm is multiscale and based on an octree decomposition of the problem domain. Similar to PILOT and IES³, our algorithm uses sampling and QR decomposition to calculate reduced representations for long-range interactions. The FFT is used to efficiently compute the interactions, as in the kernel-independent multipole method [218]. Numerical results from capacitance extraction problems demonstrate that FFTSVD is more memory efficient than FastCap or pFFT and that the algorithm does not have the homogeneity problem. In addition, we illustrate electrostatic force analysis by simulating a MEMS comb drive [226]. Finally, we demonstrate the method’s kernel-independence by calculating the electrostatic free energy of transferring a small fluorescent molecule from the gas phase to aqueous solution, using an integral formulation of a popular continuum electrostatics model [207, 227].

The following section briefly describes a representative MEMS electrostatics problem, a boundary element method used to solve the problem, and a more complicated surface formulation for calculating the electrostatic component of the solvation en-

ergy of a biomolecule. Section 3 presents the FFTSVD algorithm. Computational results and performance comparisons appear in Section 4. Section 5 describes several algorithm variants and summarizes the paper.

5.2 Background examples

In this section we describe two electrostatics problems that arise in BioMEMS design and describe how they can be addressed using BEM.

5.2.1 MEMS electrostatic force calculation

Consider the electrostatically actuated MEMS comb drive illustrated in Figure 5-1. Two interdigitated polysilicon combs form the drive; one comb is fixed to the substrate and the other is attached to a flexible tether. Applying a voltage difference to the two combs results in an electrostatic force between the two structures, and the tethered comb moves in response [226]. The electrostatic response of the system to an applied voltage difference can be calculated by solving the first kind integral equation

$$\int_S \sigma(r') G(r; r') dr' = V(r), \quad (5.1)$$

where S is the union of the comb surfaces, $V(r)$ is the applied potential on the comb surfaces, $G(r; r') = 1/||r - r'||$ is the free-space Green's function, and $\sigma(r)$ is the charge density on the comb surfaces. Note that this is a standard capacitance extraction problem.

We can compute the axial electrostatic force between the combs by the relation

$$F(s) = -\frac{d}{ds} E = -\frac{d}{ds} \frac{1}{2} V^T C(s) V, \quad (5.2)$$

where $F(s)$ is the force in the axial direction, s is the separation between the combs,

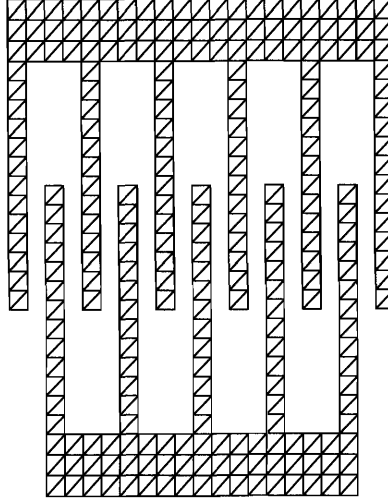


Figure 5-1: An electrostatically actuated MEMS comb drive.

E is the electrostatic energy of the system, V is the vector of conductor potentials, and $C(s)$ is the capacitance matrix, written as a function of the comb separation.

To solve (5.1) numerically, we discretize the surfaces into n_p panels and represent $\sigma(r)$, the charge density on the surface as a weighted combination of compactly supported basis functions defined on the panels:

$$\sigma(r) = \sum_{i=1}^{n_p} x_i f_i(r). \quad (5.3)$$

Here, $f_i(r)$ is the i^{th} basis function and x_i the corresponding weight. Forcing the integral over the discretized surface to match the known potential at a set of collocation points, we form the dense linear system

$$Gx = b. \quad (5.4)$$

The Green's function matrix G is defined by

$$G_{ij} = \int f_j(r') G(r_i, r') da', \quad (5.5)$$

where r_i is the i^{th} collocation point and $b_i = V(r_i)$. Alternatively, one can use a Galerkin method, in which case

$$G_{ij} = \int \int f_i(r) f_j(r') G(r; r') dr dr' \quad (5.6)$$

and

$$b_i = \int f_i(r) \psi(r) dr. \quad (5.7)$$

The linear system of equations (5.4) is solved using preconditioned GMRES [228].

5.2.2 BEM simulation of biomolecule electrostatics

Electrostatic solvation energy, the cost of transferring a molecule from a nonpolar low dielectric medium to an aqueous solution with mobile ions, plays an important role in understanding molecular interactions and properties. To calculate solvation energy, continuum electrostatic models are commonly employed. Figure 5-2 illustrates one such model. The Richards molecular surface [229] is taken to define the boundary a that separates the biomolecule interior and the solvent exterior. The interior is modeled as a homogeneous region of low permittivity ϵ_I , where the potential $\varphi(r)$ is governed by the Poisson equation, and partial atomic charges on the biomolecule atoms are modeled as discrete point charges at the atom centers:

$$\nabla^2 \varphi(r) = - \sum_{i=1}^{n_c} \frac{q_i}{\epsilon_I} \delta(r - r_i), \quad (5.8)$$

where n_c is the number of discrete point charges and q_i and r_i are the i^{th} charge's magnitude and location, respectively. In the solvent region, the linearized Poisson–Boltzmann equation

$$\nabla^2 \varphi(r) = \kappa^2 \varphi(r) \quad (5.9)$$

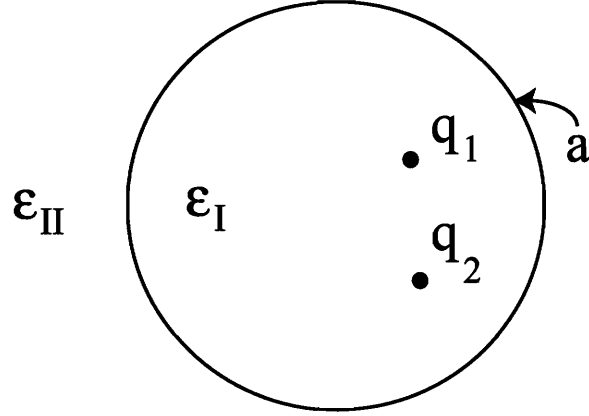


Figure 5-2: Continuum model for calculating biomolecule solvation.

governs the potential, where κ , the inverse Debye screening length, depends on the concentration of ions in the solution and a higher permittivity ϵ_{II} . We write Green's theorem in the interior and exterior regions and then enforce continuity conditions at the boundary to produce a pair of coupled integral equations,

$$\frac{1}{2}\varphi(r_a) + \int_a dr' \varphi(r') \frac{\partial G_1}{\partial n}(r_a; r') - \int_a dr' \frac{\partial \varphi}{\partial n}(r') G_1(r_a; r') = \sum_{i=1}^{n_c} \frac{q_i}{\epsilon_I} G_1(r_a; r_i) \quad (5.10)$$

$$\frac{1}{2}\varphi(r_a) - \int_a dr' \varphi(r') \frac{\partial G_2}{\partial n}(r_a; r') + \frac{\epsilon_I}{\epsilon_{II}} \int_a dr' \frac{\partial \varphi}{\partial n}(r') G_2(r_a; r') = 0, \quad (5.11)$$

where r_a is a point on the surface, f denotes the Cauchy principal value integral, G_1 is the Laplace Green's function, G_2 is the real Helmholtz Green's function, $\frac{\partial G_i}{\partial n}$ denotes the appropriate double layer Green's function, $\varphi(r)$ is the potential on the surface, and $\frac{\partial \varphi}{\partial n}(r)$ is the normal derivative of the potential on the surface. Readers are referred to [207, 227] for detailed derivations of the formulation. To solve (5.10, 5.11) numerically we define a set of basis functions on the discretized surface and represent the surface potential and its normal derivative as weighted combinations of these basis

functions:

$$\varphi(r) \approx \sum_i x_i f_i(r) \quad (5.12)$$

$$\frac{\partial \varphi}{\partial n}(r) \approx \sum_i y_i f_i(r). \quad (5.13)$$

We force the discretized integrals to exactly match the known surface conditions at the panel centroids; this produces the dense linear system

$$\begin{bmatrix} \frac{1}{2}I + \frac{\partial G_1}{\partial n} & -G_1 \\ \frac{1}{2}I - \frac{\partial G_2}{\partial n} & +\frac{\epsilon_I}{\epsilon_{II}}G_2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \sum_k \frac{q_k}{\epsilon_I} G_1(r; r_k) \\ 0 \end{bmatrix}, \quad (5.14)$$

where, denoting the i^{th} panel centroid as r_i , the block matrix entries are

$$G_{1,ij} = \int f_j(r') G_1(r_i; r') dr' \quad (5.15)$$

$$\left(\frac{\partial G_1}{\partial n} \right)_{ij} = \int f_j(r') \frac{\partial G_1}{\partial n(r')} (r_i; r') dr' \quad (5.16)$$

and the block matrices G_2 and $\frac{\partial G_2}{\partial n}$ are similarly defined. Note that boundary element method solution of this problem requires a Green's function independent fast algorithm.

5.3 The FFTSVD algorithm

The FFTSVD is a multiscale algorithm like most fast algorithms for low frequency applications: to compute the total action of the integral operator on a vector, we separate its actions at different length scales and compute them separately, combining them only at the end. In describing the FFTSVD algorithm, it is helpful to think of the basis functions as sources, $\int f_i(r') G(r; r') dr'$ as the potential produced by source i , and the collocation points r_i as destinations. Multiplying x by G in Equation (5.4)

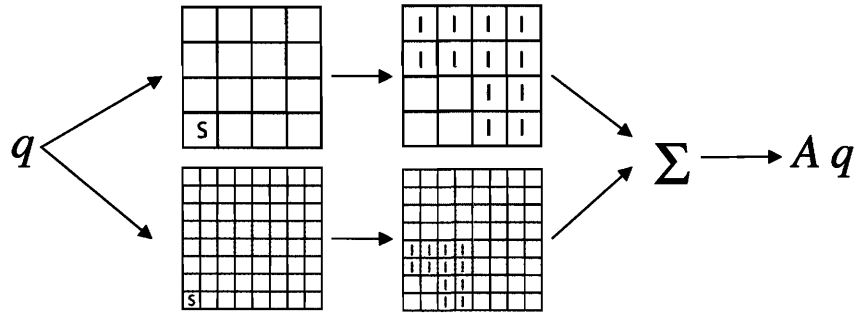


Figure 5-3: The multiscale approach to fast matrix multiplication.

is then computing potentials at all the destinations due to all sources. Figure 5-3 illustrates the multiscale approach to fast matrix multiplication: the square S denotes a source, and the squares denoted I represent destinations.

5.3.1 Notation

Let d and s denote two sets of panels: then $G_{d,s}$ is the submatrix of G that maps sources in s to responses in d . The number of panels in set i is denoted by n_i .

5.3.2 Octree decomposition

We first define the problem domain to be the union of all the sets of panels that comprise the discretized surfaces. We then place a bounding cube around the domain and recursively decompose the cube using octrees. Given a cube s at level i , the *nearest neighbors* N_s are those cubes at level i that share a face, edge, or vertex with s . The *interaction list* for s is denoted as I_s and defined to be the set of cubes at level i that are not nearest neighbors to s and not descended from any cube in an interaction list of an ancestor of s [230]. Figure 5-4 illustrates the exclusion process for a 2-D domain. At every level, each panel is assigned to the cube that contains its centroid. Where ambiguity will not result, s denotes either the cube itself or the set of panels assigned to it. This assignment rule ensures that each panel-panel interaction is treated exactly once.

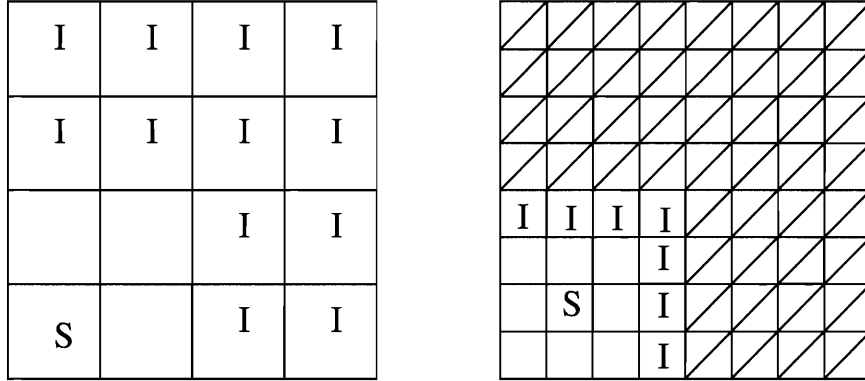


Figure 5-4: Interacting squares at two levels of decomposition.

The coarsest decomposition is termed level 0 and has 4^3 cubes; coarser decompositions have null interaction lists. We continue decomposing the domain until we reach a level l at which no cube is assigned more than $n_{p,max}$ destinations. At each level i , every cube s has a set of interacting cubes I_s that are well-separated from s with respect to the current cube size. Note that the definition of an interaction list is symmetric: $d \in I_s \rightarrow s \in I_d$.

5.3.3 Sampling dominant sources and responses

One can compute the potential response φ_{I_s} in I_s due to a source q_s in s by the dense matrix-vector product

$$\begin{aligned} \varphi_{I_s} &= G_{I_s,s} q_s & (5.17) \\ G_{I_s,s} &\in \mathfrak{R}^{n_{I_s} \times n_s}. \end{aligned}$$

However, the separation between s and I_s motivates the approximation

$$\begin{aligned}
G_{I_s,s} &\approx U_{I_s} V_{s,\text{src}}^T & (5.18) \\
U_{I_s} &\in \mathfrak{R}^{n_{I_s} \times k} \\
V_{s,\text{src}}^T &\in \mathfrak{R}^{k \times n_s} \\
k &\ll n_{I_s}
\end{aligned}$$

where $V_{s,\text{src}}$ has orthogonal columns [50]. The matrix $V_{s,\text{src}}$ is small and represents the k source distributions in s that produce dominant effects in I_s . It is a reduced row basis for $G_{I_s,s}$. The projection of q_s onto $V_{s,\text{src}}$ loosely parallels the fast multipole method's calculation of multipoles from sources, in the sense that both the multipole expansion and the product $V_{s,\text{src}}^T q_s$ capture the important pieces of q_s when calculating far-field interactions. We call $V_{s,\text{src}}$ the source compression matrix.

A similar low-rank approximation can be made to find the response in a cube d given a source distribution in I_d :

$$\begin{aligned}
\varphi_d &= G_{d,I_d} q_{I_d} & (5.19) \\
&\approx U_{d,\text{dest}} V_{I_d}^T q_{I_d} \\
U_{d,\text{dest}} &\in \mathfrak{R}^{n_d \times k} \\
V_{I_d}^T &\in \mathfrak{R}^{k \times n_{I_d}} \\
k &\ll n_{I_d}.
\end{aligned}$$

Here, $U_{d,\text{dest}}$ is small and represents the k dominant potential responses in d , the destination cube, due to source distributions in I_d . We call $U_{d,\text{dest}}$ the destination compression matrix; $U_{d,\text{dest}}$ is a reduced column basis for G_{d,I_d} .

Since it is impractical to compute $G_{I_s,s}$ and G_{s,I_s} for each cube s , we use a sampling procedure inspired by the Kapur and Long hierarchical SVD method [50]. Figures 5-

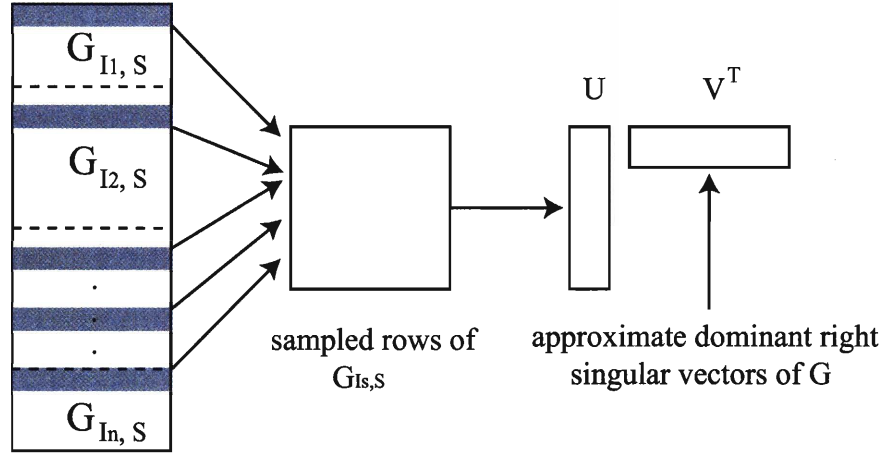
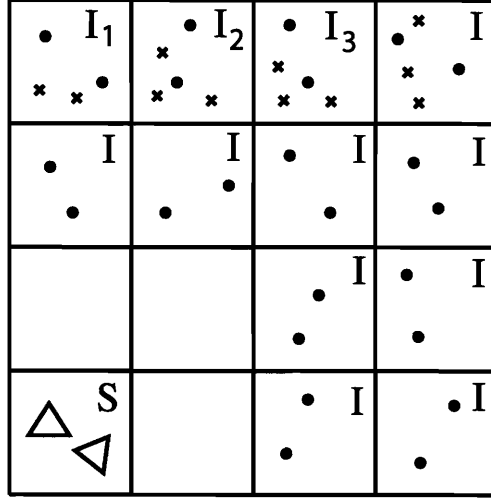


Figure 5-5: Computing dominant row basis for $G_{I_s, s}$ using sampling.

5 and 5-6 illustrate the process of finding a reduced row basis $V_{s, \text{src}}$. To determine the row basis, we begin by selecting one destination per interacting cube, computing the corresponding rows of $G_{I_s, s}$, and performing rank-revealing QR factorization with reorthogonalization on the transpose of the submatrix. If the submatrix rank is less than half the number of sampled destinations, the QR-determined row basis is considered to be adequate. Otherwise, an additional destination is sampled for each interacting cube; the extra destination is chosen to be well-separated from the originally chosen destination. The transpose of the new submatrix is factorized and again required to have rank less than half the total number of samples. The process of resampling is continued until the required rank threshold is met.

To compute the reduced column basis $U_{d, \text{dest}}$ for the matrix G_{d, I_d} , we select a set of well separated panels in I_d , compute the corresponding columns of G_{d, I_d} , and QR factorize the submatrix.



- * Collocation points
- Sampled collocation points
- △ Basis function support

Figure 5-6: Sampling a small set of long-range interactions.

5.3.4 Computing long-range interactions

Consider two well separated cubes s and d . Because the cubes are well separated, we could find a low-rank approximation to $G_{d,s}$ by truncating its SVD:

$$\varphi_d = G_{d,s} q_s \tag{5.20}$$

$$= U_{d,s} \Sigma_{d,s} V_{d,s}^T q_s \tag{5.21}$$

$$\approx \hat{U}_{d,s} \hat{\Sigma}_{d,s} \hat{V}_{d,s}^T q_s \tag{5.22}$$

where the hat denotes truncation to k columns, $k < n_s$. Since the source compression matrix $V_{s,\text{src}}$ finds an approximation to the dominant row space of $G_{I_s,s}$, we expect that it also approximates the dominant row space of $G_{d,s}$, which is a submatrix of $G_{I_s,s}$. Similarly, we expect that $U_{d,\text{dest}}$ approximates the dominant column space of $G_{d,s}$. A small matrix $K_{d,s}$ maps source distributions in the reduced basis $V_{s,\text{src}}$ to

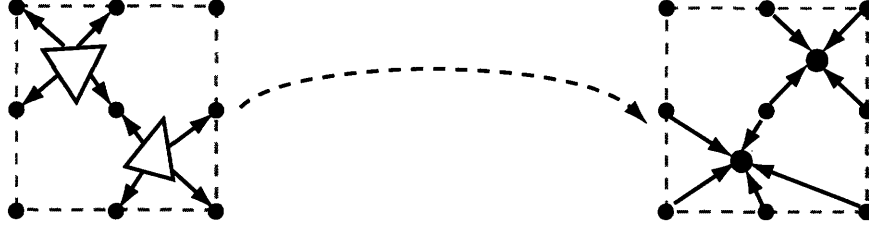


Figure 5-7: Schematic of the FFTSVD method for computing long-range interactions.

responses in the reduced basis $U_{d,\text{dest}}$:

$$\varphi_d = U_{d,\text{dest}} K_{d,s} V_{s,\text{src}}^T q_s, \quad (5.23)$$

and it is easy to see that

$$K_{d,s} = U_{d,\text{dest}}^T G_{d,s} V_{s,\text{src}}. \quad (5.24)$$

Note that $K_{d,s}$ is not diagonal because $U_{d,\text{dest}}$ and $V_{s,\text{src}}$ only approximate the singular vectors of $G_{d,s}$. If $V_{s,\text{src}} \in \mathfrak{R}^{n_s \times k_s}$ and $U_{d,\text{dest}} \in \mathfrak{R}^{n_d \times k_d}$, then $K_{d,s} \in \mathfrak{R}^{k_d \times k_s}$.

The action of the K matrices can be computed in a number of different ways: they can be computed explicitly, via multipoles, or via an FFT. Explicit storage is memory intensive, and multipole representations are Green's function dependent. We have therefore chosen to implement the memory-efficient, Green's function independent FFT translation method presented by Ying *et al.* [218].

5.3.5 Diagonalizing long-range interactions with the FFT

Our method projects sources to a grid, uses an FFT convolution to accomplish translation between source and destination, and interpolates results back from the grid. Figure 5-7 illustrates the approach. We introduce two matrices: $P_{g,j}$ projects sources in cube j to the cube grid, and $I_{j,g}$ interpolates from the grid in cube j to the evaluation points in j . We use an equivalent density scheme similar to those used by Phillips and White [49] and Biros *et al.* [217] to determine the projection and interpolation

matrices.

Projection matrix calculation

Given a cube s and the basis function weights q_s for panels in s , we wish to find a set of grid charges $q_{g,s}$ that reproduce the potential field far from s . We accomplish this by defining a sphere Γ bounding s and picking a set of quadrature points [231] on the sphere. Denoting quadrature point i on Γ by $r_{\Gamma,i}$, the mapping between q_s and the responses at the quadrature points can be written as $G_{\Gamma,s}$, where

$$G_{\Gamma,s,ij} = \int_{\text{panel } j} G(r_{\Gamma,i}; r') dr'. \quad (5.25)$$

The mapping between grid charges and responses at the quadrature points can be written as

$$G_{\Gamma,g,ij} = G(r_{\Gamma,i}, r_{g,j}) \quad (5.26)$$

where $r_{g,j}$ is the position of the j^{th} grid point. If more quadrature points than grid points are used for the matching, solving a least squares problem gives the desired projection $P_{g,s}$:

$$P_{g,s} = G_{\Gamma,g}^{-1} G_{\Gamma,s}. \quad (5.27)$$

In practice, one uses the singular value decomposition to solve for $P_{g,s}$.

Interpolation matrix calculation

Given grid potentials q_d in a cube d , we find the potentials φ_d at the panel centroids in d by interpolation. For problems in which centroid collocation is used to generate a linear system of equations, the interpolation matrix is calculated as

$$I_{d,g} = (G_{\Gamma,g}^{-1} G_{\Gamma,d})^T \quad (5.28)$$

where $G_{\Gamma,d}$ denotes the Green’s function matrix from the quadrature points on Γ to the panel centroids in d . If Galerkin methods are used rather than centroid collocation, the interpolation matrix is the transpose of the projection matrix.

Diagonal translation

Once the grid charges in s are known, a spatial convolution with the Green’s function produces the potentials at the grid points in the destination cube d . This spatial convolution is diagonalized by the Fourier transform; we write the transform matrix as \mathcal{F} , its inverse by \mathcal{F}^{-1} , and the transform of the Green’s function matrix by $\tilde{G}_{d,s}$. After calculating the grid potentials in d , interpolation produces the potentials at the desired evaluation points. The matrix $G_{d,s}$ is therefore written as

$$G_{d,s} = I_{d,g}\mathcal{F}^{-1}\tilde{G}_{d,s}\mathcal{F}P_{g,s}. \quad (5.29)$$

The products $I_{d,g}\mathcal{F}^{-1}$ and $\mathcal{F}P_{g,s}$ could be stored, but in our experience this precomputation only marginally improves the matrix–vector product time while increasing memory use since \mathcal{F} and \mathcal{F}^{-1} are padded and complex.

In addition to diagonalizing the translation operation between cubes, the FFT significantly decreases memory requirements. Using explicit K matrices requires storing a small dense matrix for each pair of cubes; using FFT translation eliminates the expensive per-pair matrix cost. Instead, each cube has its own P_g and I_g matrices, which are used for all long-range interactions. In addition, because the Green’s function is translationally invariant, we only need to store a small number of \tilde{G} matrices for each octree level; each one represents a particular relative translation between source and destination cubes. Because these matrices are diagonal, storage requirements are minimal.

Since translation is the dominant cost in the FFTSVD matrix–vector product,

efficient implementation of the translation procedure is essential to maximizing performance. The translation operation is simply an element-wise multiplication of two complex vectors, therefore, for g_p grid points per cube side, each translation vector is $(2g_p - 1)^2[(2g_p - 1)/2 + 1]$ complex numbers long when using the FFTW library [232]. This number takes into account padding and symmetry. For example, with $g_p = 3$, 75 complex numbers are required, resulting in 250 individual multiplies during the translation operation. This number has been reduced by taking advantage of vectorization. Many modern CPUs include instructions that can assist in multiplying complex numbers within a register, effectively halving the number of required multiplies. For comparison, standard fast multipole method translations require more multiplications since they are not diagonal, and cannot be vectorized as easily since they involve matrix-vector products. In addition, we have yet to exploit additional ways to accelerate the FFTSVD translation operation. These include using symmetries between related translation vectors (\tilde{G}), such as those that translate in opposite directions, and exploiting the fact that for axial translations, many \tilde{G} elements are purely real.

5.3.6 Local interactions

At the finest level of the decomposition, interactions between nearest neighbor cubes are computed directly by calculating the corresponding dense submatrices of G . These submatrices are denoted by $D_{i,j}$ where j is the source cube and i the destination. We bound the complexity of the local interaction computation by continuing the octree decomposition until each cube has fewer than $n_{p,\max}$ panels.

5.3.7 Algorithm detail

The mapping from source cube s to destination cube d can thus be written as

$$\varphi_d = U_d \left(U_d^T I_{d,g} \right) \mathcal{F}^{-1} \tilde{G} \mathcal{F} (P_{g,s} V_s) V_s^T q_s \quad (5.30)$$

The computations are grouped to eliminate redundant multiplications; the matrix products $U_d^T I_{d,g}$ and $P_{g,s} V_s$ are stored for each cube rather than recomputed at every iteration. Below, we introduce the restriction operator $M_j^{(i)}$ that restricts a global vector to a local vector associated with cube j at level i ; let the inverse operator map a local vector to the global by inserting appropriate zeros. Let L^i denote the set of cubes at level i . Given a charge vector q , the matrix–vector product is computed by the following procedure:

1. DOWNWARD PASS FOR LONG-RANGE INTERACTIONS: For levels $i = 0, 1, \dots, l$:

- (a) PROJECT INTO DOMINANT SOURCE SPACE: For each cube $j \in L^i$, compute

$$\zeta_j = \mathcal{F}(P_{g,j} V_{j,\text{src}}) V_{j,\text{src}}^T M_j^{(i)} q. \quad (5.31)$$

- (b) COMPUTE LONG-RANGE INTERACTIONS: For each cube $j \in L^i$, compute

$$\nu_j = \sum_{s \in I_j} \tilde{G} \zeta_s. \quad (5.32)$$

- (c) DETERMINE TOTAL DOMINANT RESPONSE: For each cube $j \in L^i$, compute

$$\varphi = \varphi + M_j^{(i),-1} U_{j,\text{dest}} (U_{j,\text{dest}}^T I_{j,g}) \mathcal{F}^{-1} \nu_j. \quad (5.33)$$

2. SUM DIRECT INTERACTIONS: For each cube d at level l , add the contribu-

tions from neighboring cubes N_d :

$$\varphi = \varphi + M_d^{(l),-1} \sum_{s \in N_d} D_{d,s} M_s^{(l)} q. \quad (5.34)$$

5.4 Computational results

To demonstrate the accuracy, speed, and memory efficiency of the FFTSVD algorithm, we have used FFTSVD to solve for self and mutual capacitances in various geometries. A MEMS comb drive example [226] illustrates electrostatic force calculation using FFTSVD. In addition, to show Green’s function independence and use of double layer kernels, we have used FFTSVD to solve for the electrostatics of solvation for the highly charged dye molecule fluorescein. Fluorescein is often used as a fluorescent label in BioMEMS applications [233, 234], and its electrostatic properties in aqueous solution modulate its interaction with other molecules and surfaces.

The FFTSVD algorithm has several adjustable parameters: ϵ_{QR} is the reduced basis tolerance; g_p is the number of FFT grid points on each side of a finest-level cube; $n_{p,\max}$ is the maximum number of panels in a finest-level cube; n_{quad} is the number of quadrature points used on the equivalent density sphere, tol_{GMRES} is the tolerance on the relative residual that the resulting linear equations are solved to. At the two finest levels, g_p FFT grid points per cube edge are used, and the number of grid points per edge increases by one for each successively coarser level; experience has shown that using different numbers of grid points per edge provides significant accuracy improvements for marginal memory and time costs. The parameters used for the following results are 10^{-4} for ϵ_{QR} , 3 for g_p , 32 for $n_{p,\max}$, 25 for n_{quad} , and 10^{-4} for tol_{GMRES} unless otherwise specified.

For capacitance calculations, we compare performance to FastCap, based on the fast multipole method [47], and `fftcap++`, based on the `pFFT++` implementation

of the precorrected-FFT method [235]. All programs were compiled with full optimizations using the Intel C++ compiler version 8.1 and benchmarked on an Intel Pentium 4 3.0-GHz desktop computer with 2 GB of RAM. All parameter settings in FastCap and fftcap++ were left at their defaults, except for the tolerance on solving the resulting linear equations, which was set to 10^{-4} unless otherwise specified.

5.4.1 Self-capacitance of a sphere

In order to test the accuracy of the FFTSVD method, we have applied it to solving for the self-capacitance of a unit 1-m radius sphere, a quantity known analytically. Figure 5-8 shows the improvement in accuracy with increasing sphere discretization for FFTSVD with values of 3 and 5 for g_p , 2nd and 4th order multipoles in FastCap, and default settings for fftcap++. A tolerance of 10^{-6} for the relative residual when solving the BEM equations was used in all programs. The analytical value for the self-capacitance of a 1-m radius sphere is 0.111265 nF as computed by Gauss' law. The results show that FFTSVD with a value of 3 for g_p tends to be more accurate than 2nd order multipoles in FastCap. In addition, FFTSVD with low values of g_p tends to overshoot the analytical solution while FastCap tends to undershoot with truncation of multipole order. These findings are consistent across many geometries when examining convergence behavior.

5.4.2 Woven bus example (homogeneous problem)

As stated previously, one of the advantages of the FFTSVD method is its use of diagonal translation operators. This advantage becomes apparent in cases of homogeneous geometry, since a large number of translation operations are required. To examine performance in a problem with homogeneous geometry, we have applied FFTSVD to solving for the mutual capacitances between woven bus conductors as in Figure 5-9. Table 5.1 summarizes the results for several woven bus capacitance

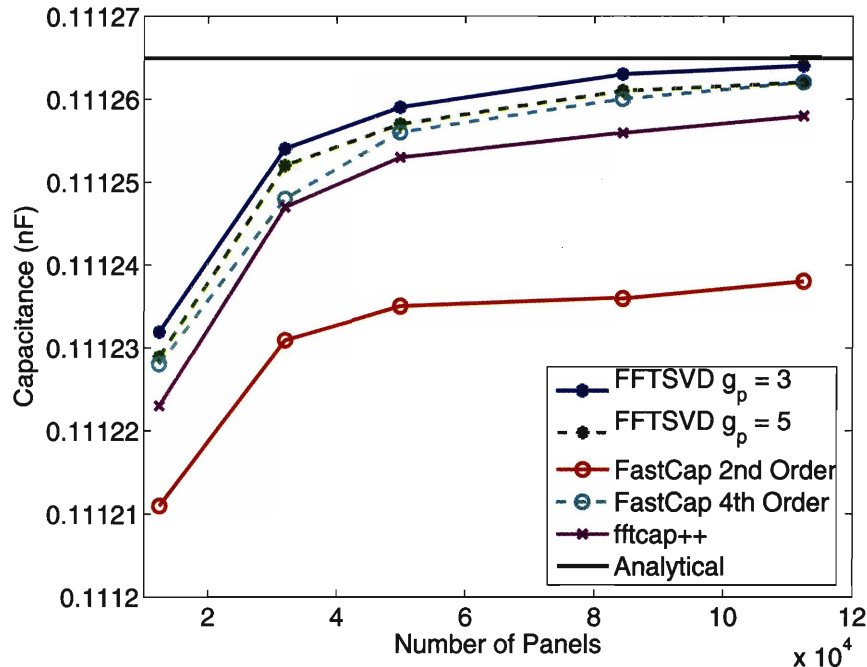


Figure 5-8: Accuracy versus number of panels for FFTSVD, FastCap and fftcap++ solving the unit sphere self-capacitance problem.

Table 5.1: Comparison of FastCap (FC), fftcap++ (FFT++) and FFTSVD (FS) performance in terms of matrix–vector product time (MV) and memory usage (MEM) on homogeneous woven bus capacitance problems with 2, 5 and 10 crossings (woven02n03, woven05n03, woven10n03) and 10 crossings with lower discretization (woven10n01).

Problem	Panels	FC MV (s)	FC MEM (MB)	FFT++ MV (s)	FFT++ MEM (MB)	FS MV (s)	FS MEM (MB)
woven02n03	3168	0.03	30	0.02	23	0.01	11
woven05n03	18720	0.17	205	0.22	411	0.09	110
woven10n01	8160	0.08	89	0.04	69	0.04	41
woven10n03	73440	0.73	901	0.51	818	0.41	466

problems. FFTSVD can achieve slightly better speed and memory performance than precorrected-FFT, which is expected to excel at problems with uniform distribution, and significantly better performance as compared to FastCap.

5.4.3 Inhomogeneous capacitance problem

One of the disadvantages of the precorrected-FFT method is that it lays down a uniform grid over the entire problem domain, and the simulation time grows roughly in proportion to the number of grid points. For simulations in which most of the domain is empty, therefore, the precorrected-FFT algorithm is inefficient. We have

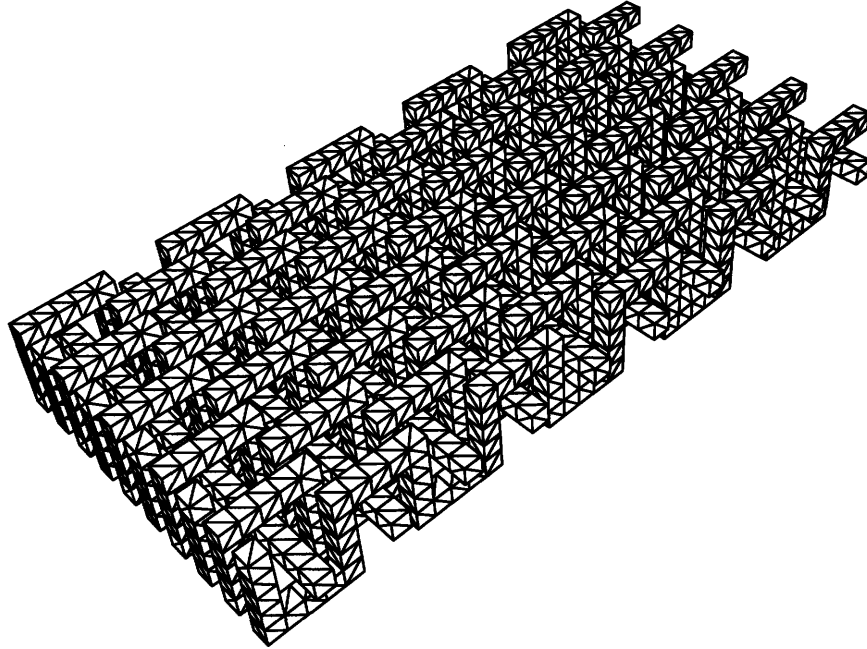


Figure 5-9: Homogeneous woven bus capacitance problem (woven10n01).

demonstrated this inefficiency, and FFTSVD's relative advantage, by configuring a set of conductors as shown in Figure 5-10. Almost all of the panels in this system are at the edges of a cube bounding the domain. Figure 5-11 plots the matrix-vector product times for the FFTSVD, FastCap and `ftcap++` codes, and Figure 5-12 plots the memory requirements. As expected, the precorrected-FFT based `ftcap++` code has poor performance, especially for fine discretizations of the inhomogeneous problem. FFTSVD performs consistently better than `ftcap++` and generally better than FastCap. The sharp jumps in FFTSVD and `ftcap++` matrix-vector product time with increasing panel count are due to a change in selection of the optimal octree decomposition depth or FFT grid size, respectively.

5.4.4 MEMS comb drive

We have simulated the MEMS comb drive illustrated in Figure 5-1 [226]. We applied a voltage difference of 1 V to the two structures and used a fourth-order finite difference scheme to approximate the derivative in Equation (5.2). Because the

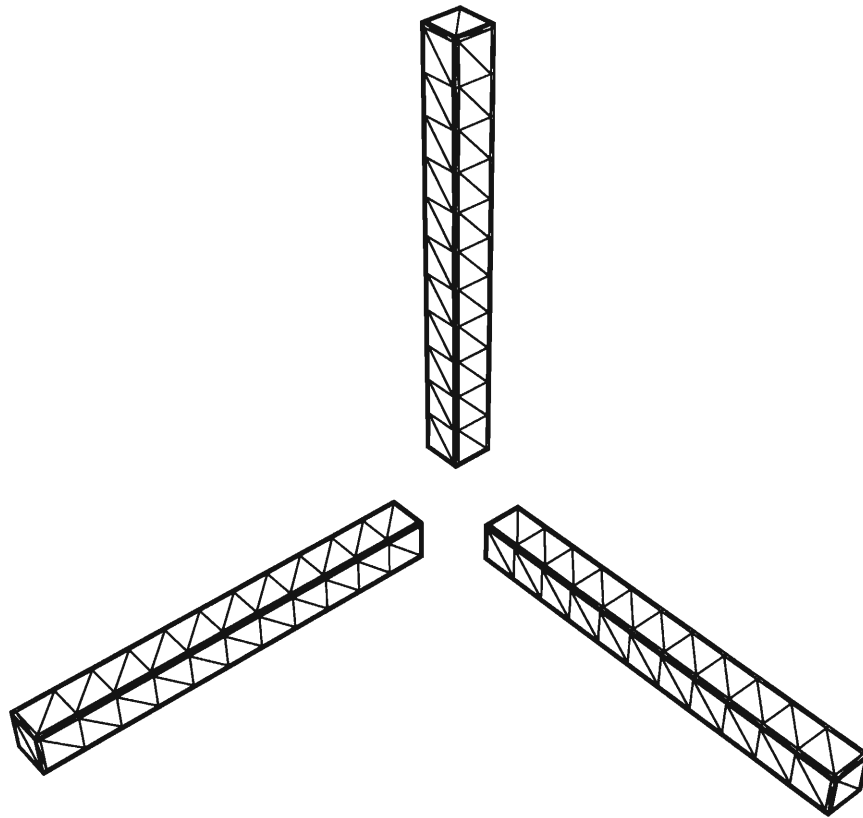


Figure 5-10: Inhomogeneous capacitance problem.

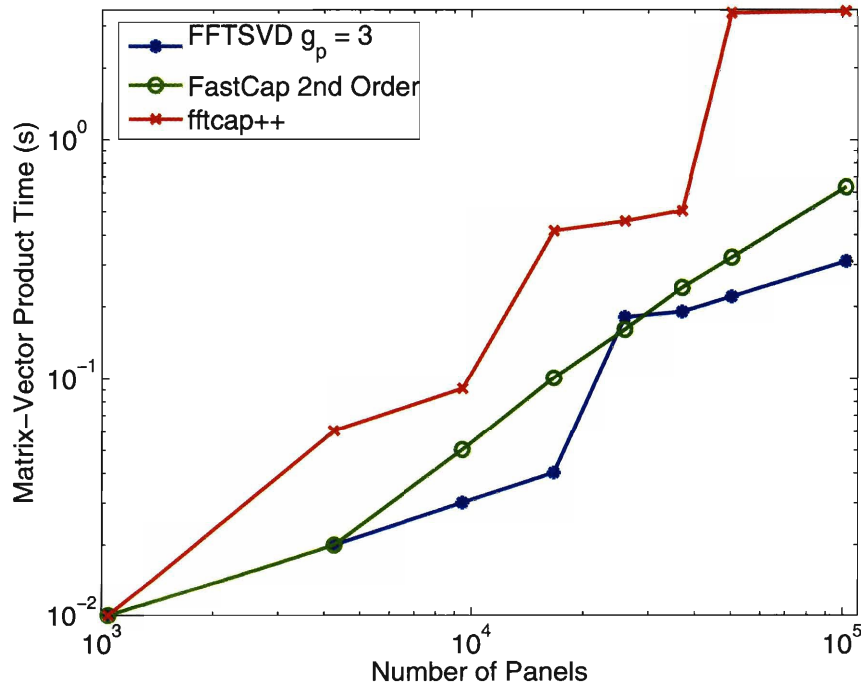


Figure 5-11: Matrix-vector product times for FFTSVD, FastCap and fftcap++ codes solving the inhomogeneous capacitance problem.

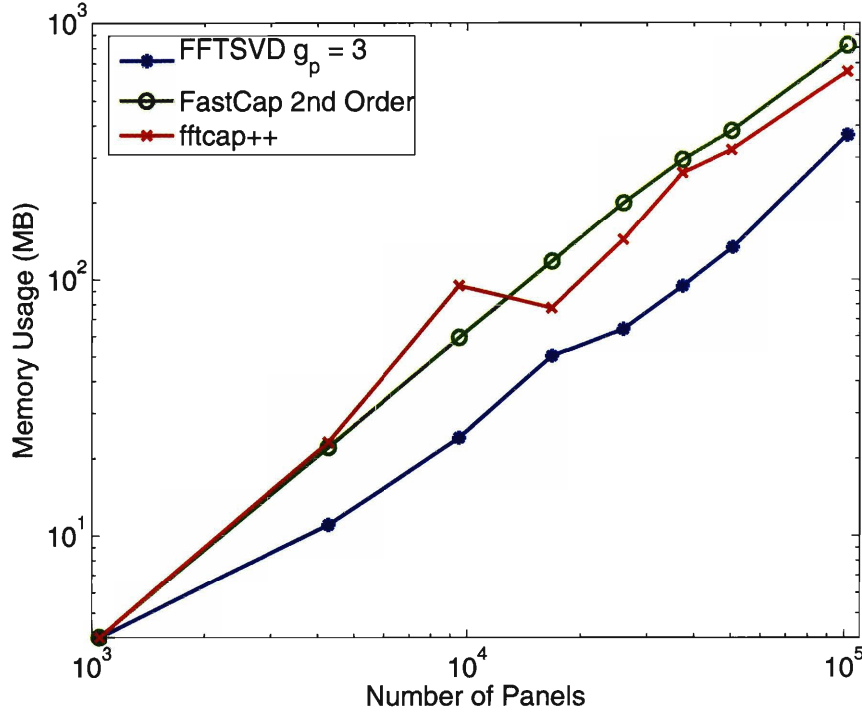


Figure 5-12: Memory requirements for FFTSVD, FastCap and fftcap++ codes solving the inhomogeneous capacitance problem.

finite-difference scheme for force calculation requires high accuracy in the capacitance calculations, more stringent parameters are required for these simulations. We have used $tol_{GMRES} = 10^{-6}$, $\epsilon_{QR} = 10^{-6}$, $g_p = 5$, $n_{QUAD} = 64$, and for each discretization we have fixed $n_{p,max}$ such that the octree decomposition depth is equal for each of the four geometries.

The contribution of each panel to the axial force is plotted in Figure 5-13 and the total axial electrostatic force is plotted in Figure 5-14 as a function of the number of panels used to discretize the comb drive. We have used general triangles and note that the discretization scheme is poorly tuned for the calculation of electrostatic forces; nonuniform meshes achieve superior accuracy at reduced panel counts [236]. The force can also be calculated by integrating the squared charge density over the conductor surface, but this approach requires specialized treatment because the charge density becomes infinite at the edges and corners of the conductors [237, 238].

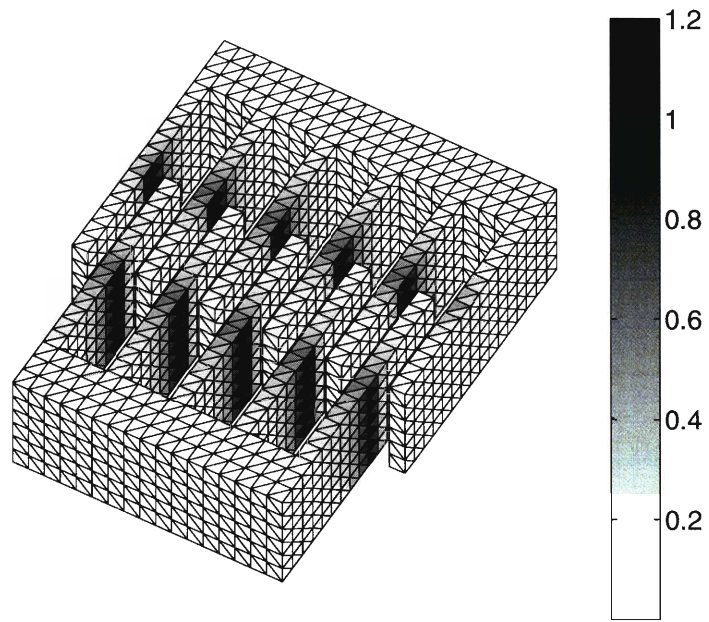


Figure 5-13: Magnitudes of panel contributions to the axial electrostatic force. Units are pN.

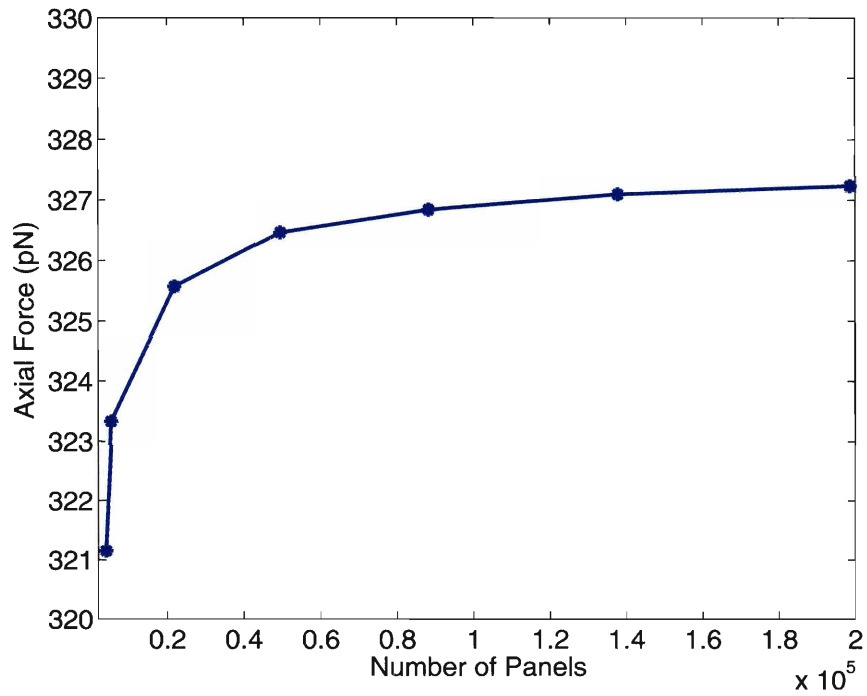


Figure 5-14: Calculated total axial electrostatic force on one comb.

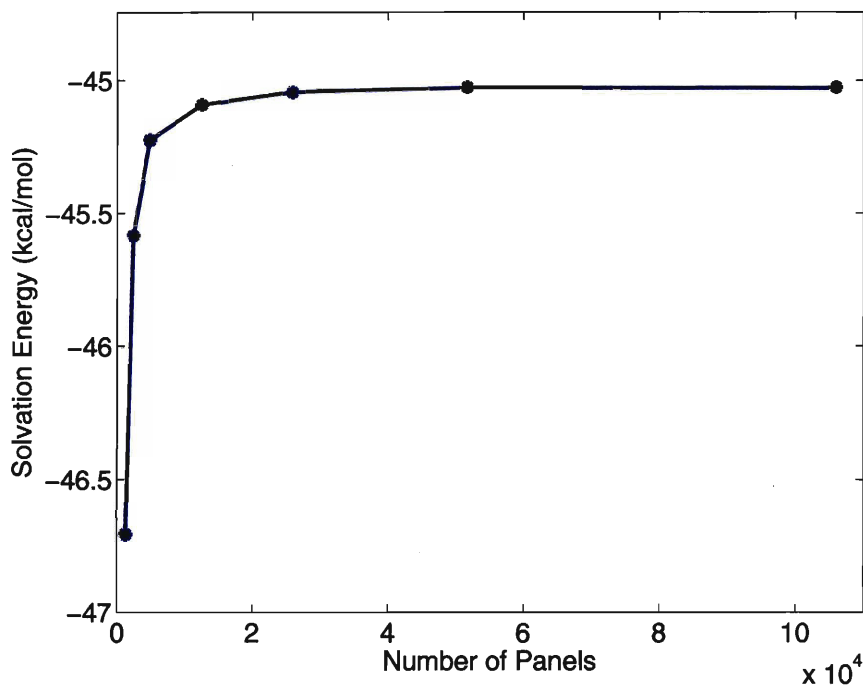


Figure 5-15: Computed electrostatic solvation energy of fluorescein with increasing problem discretization.

5.4.5 Solvation of fluorescein

We have used the integral formulation in Equations (5.10) and (5.11) to calculate the solvation energy of fluorescein. To prepare a model for solvation calculations, its structure and partial atomic charges were determined from quantum mechanical calculations. Radii were assigned to each atom and used to generate a triangulation of the molecular surface. The interior of the fluorescein molecule was assigned a dielectric constant of 4, and the exterior was assigned a dielectric constant of 80 (for water) with an ionic strength of 0.145 M ($\kappa = 0.124 \text{ \AA}^{-1}$). FFTSVD was used to solve for both the electrostatic solvation energy (Figure 5-15), as well as the total electrostatic potential on the surface of the fluorescein molecule (Figure 5-16). We note that the long-range single and double layer integrals can be computed using only one set of translation operations. Different projection operators are used to find the corresponding grid charges due to monopole and dipole distributions, and the grid charges can then be summed for translation.

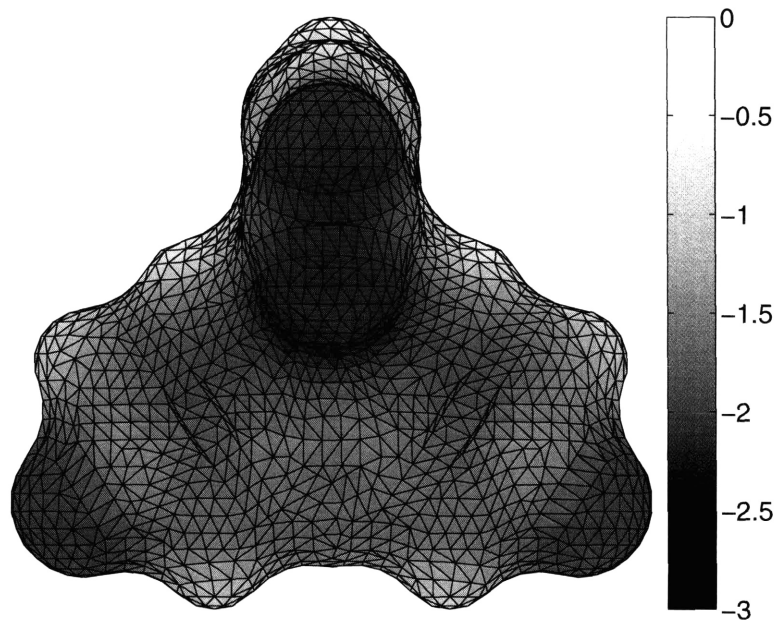


Figure 5-16: Electrostatic solvation potentials on the molecular surface of fluorescein. Units are $\text{kcal mol}^{-1} e^{-1}$.

5.5 Discussion

5.5.1 Algorithm variants

For problems with a small number of integral operators, memory constraints may not be a significant consideration. In these cases, the matrices $K_{d,s}$ can be stored explicitly. These $K_{d,s}$ matrices are computed using Equation (5.24), but instead of computing $G_{d,s}$ explicitly, we project, translate and interpolate an identity matrix using the methodology outlined in Section 5.3.5. Although setup time and memory use increase when explicit K-matrices are used, the matrix-vector product time is significantly reduced. We have also implemented a parameter that allows a tradeoff between speed and memory use through K-matrices. Pairs of interacting octree cubes that contain fewer panels than the parameter are handled with explicit K-matrices, while all other cubes use the FFT-based translation. In this manner, the balance between speed and memory can be fine-tuned for the given application.

It is also straightforward to create an FFTSVD variant that runs in linear time;

the same method used to generate the projection and interpolation matrices can be used to create “upward pass” and “downward pass” operators such as those found in multipole algorithms. This variant algorithm is essentially equivalent to the kernel-independent method by Ying *et al.* [218], except that we allow all the grid charges to be nonzero. The Ying method, in contrast, uses only grid charges on the surface of the cube.

The linear-time FFTSVD method requires a greater number of grid points per cube, due to the loss of degrees of freedom during each upward pass from child to parent cube. In addition, the SVD based compression of dominant sources and responses is no longer computed, since these bases are now taken directly from child cubes. This method is extremely memory efficient since dominant source and response bases are no longer stored, but it trades off performance to achieve it due to the larger required grid sizes.

Finally, the multilevel structure of FFTSVD allows easy parallelization. Each processor can be assigned responsibility for a set of cubes on coarse levels, and the computation can proceed independently until the final potential responses are summed. We have implemented parallel FFTSVD using both OpenMP and MPI libraries with good results.

5.5.2 Summary

We have developed a fast algorithm for computing the dense matrix-vector products required to solve boundary element problems using Krylov subspace iterative methods. The FFTSVD method is a multiscale algorithm; an octree decomposes the matrix action into different length scales. For each length scale, we use sampling to calculate reduced bases for the interactions between well-separated groups of panels. The FFT is used to diagonalize the translation operation that computes the long-range interactions. The method described here relies on both kernel decay and

translation invariance.

Numerical results illustrate that FFTSVD is much more memory-efficient than FastCap or precorrected-FFT, and that it is generally faster than either technique on a variety of problems. In addition, FFTSVD is Green's function independent, unlike FastCap, and the method performs well even when the problem domain is sparsely populated, unlike precorrected-FFT. Our implementation is well-suited to solve problems with multiple dielectric regions. Finally, we note that the structure of the algorithm permits treatment of kernels that are not translation-invariant; for such problems, the K -matrix algorithm variant should be used rather than the FFT. Together, the algorithm's performance and flexibility make FFTSVD an excellent candidate for fast BEM solvers for microfluidic and microelectromechanical problems that appear in BioMEMS design.

Chapter 6

Accurate Solution of Multi-region Continuum Electrostatic Problems Using the Linearized Poisson–Boltzmann Equation and Curved Boundary Elements ¹

Abstract

We present a boundary-element method (BEM) implementation for solving problems in biomolecular electrostatics using the linearized Poisson–Boltzmann equation. The motivating factor behind this implementation was the desire to create an efficient and accurate solver capable of precisely describing the molecular topologies prevalent in continuum models. Underlying this implementation are three key features that address many of the well-known practical challenges associated with the boundary-element method. First, we present a general boundary-integral approach capable of modeling an arbitrary number of embedded homogeneous dielectric regions with differing dielectric constants, possible salt treatment, and point charges. Second, molecular and accessible surfaces used to describe dielectric and ion-exclusion boundaries are discretized with curved boundary elements that faithfully reproduce even complicated geometries. Robust numerical integration methods are employed to accurately evaluate singular and near-singular integrals over the curved boundary elements. Third, we avoid explicitly forming the dense BEM matrix, and instead solve the linear system with preconditioned GMRES, using the FFTSVD algorithm to accelerate matrix–vector multiplication. A comparison of the presented BEM implementation and standard finite-difference techniques demonstrates that for certain classes of electrostatic calculations, the improved convergence properties of the BEM approach can have a significant impact on computed energetics. These results suggest that solvers with improved accuracy may be important to ensure that predictions based on continuum models are limited by the models themselves rather than by errors in the models’ numerical evaluation.

¹All work presented in this chapter was performed in collaboration with J. P. Bardhan and J. K. White at the Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science.

6.1 Introduction

Continuum theories of solvation have become common tools for molecular modeling, and have led to an improved understanding of electrostatic interactions in biomolecular systems [27, 29]. One of the most popular models of continuum solvation treats a molecule and its solvent environment as homogeneous regions of low and high dielectric constants respectively, with embedded point charges representing the molecular charge distribution and Debye–Hückel theory modeling the effect of salt. The linearized Poisson–Boltzmann equation governs this continuum model, and this equation has received much attention in recent years [239–241]. The linearized Poisson–Boltzmann equation (LPBE), an elliptic partial differential equation (PDE) [242], is well understood theoretically and can be solved numerically using a variety of techniques including finite-difference methods (FDM) [25, 26, 28, 122, 243–246] finite-element methods (FEM) [206, 247–249], and boundary-element methods (BEM) [250–261].

Boundary-element methods offer several inherent advantages over volume-based methods for solving the LPBE with regions of homogeneous dielectric [262]. For example, the BEM only requires discretization of problem boundaries rather than the entire infinite domain, and inherently captures the correct zero-potential boundary condition at infinity. In comparison to finite-difference methods, the BEM has the ability to model point charges exactly, rather than requiring grid projection.

Unfortunately, boundary-element methods require sophisticated numerical techniques in order to be competitive with the flexibility and performance of volume-based methods. Several challenges complicate the implementation of BEM techniques for biomolecule electrostatics. The first challenge arises from the surface-based analysis of the problem. Some Poisson–Boltzmann modeling problems require treatment of multiple embedded or disconnected regions with differing dielectric constants and screening parameters [263, 264]. These features allow the simulation of solvent-filled cavities

within macromolecules, salt-filled regions in large cavities, and an ion-exclusion layer surrounding the molecule with solvent permittivity but no salt. Multiple regions are easily modeled in volume methods like FDM and FEM because the dielectric constant and the presence of salt can be assigned to each grid point or volume element independently. Implementing these features using BEM requires the discretization of every interface between dielectric regions and between those governed by differing PDEs. In contrast, volume-based methods need no additional degrees of freedom. Previous BEM approaches have addressed these limitations by developing specific formulations to treat multiple embedded dielectric regions without salt [263], multiple disconnected dielectric bodies with salt [264], and hybrid boundary-element/finite-difference methods to treat ion-exclusion layers [265].

A second important challenge for biomolecule BEM is the strong dependence of solution accuracy on the quality of the surface representation. In this work and in most others, the dielectric and ion-exclusion surfaces are described according to one of two definitions. Accessible surfaces [266] are defined as a union of spheres, where the atomic radii are expanded by a probe's radius. Molecular surfaces [229, 267, 268] represent the surface of closest approach of a probe sphere rolled over a union of spheres representing a molecule. These curved surfaces, which consist of portions of spheres and torii, are analytically defined but often difficult to discretize because the surfaces have cusps and singularities. Most boundary-element methods for solving the LPBE represent these surfaces approximately using large numbers of planar triangular elements, or panels, that can never truly capture the curved geometries. The importance of using curved elements has already been discussed [269, 270], but previous implementations have introduced other approximations. For example, other work has modified the molecular surface definition to avoid singularities and thin regions, used elements with low-order curvature that cannot accurately represent spheres or torii, or discretized surfaces using standard spherical triangles that cannot exactly

represent the intersections between atoms.

A third challenge for BEM is that discretization of surface integral equations gives rise to dense linear systems of equations. As a result, memory costs scale quadratically in the number of unknowns. In contrast, the FDM and FEM generate sparse matrices that reflect the local nature of the differential operators. Solving the BEM linear system by matrix factorization requires $O(n^3)$ time, where n is the number of unknowns. Computational costs rapidly become prohibitive for systems with more than 10^4 unknowns, which is currently insufficient to accurately model large macromolecules such as proteins. The quadratic memory and cubic time costs can be reduced to linear or near-linear complexity by combining two approximation schemes. First, the linear systems are solved approximately, rather than exactly, using Krylov subspace iterative methods such as the conjugate gradient method (CG) or the generalized minimum residual algorithm (GMRES) [228]. Every iteration of a dense Krylov subspace method requires the multiplication of a vector by the BEM matrix, costing a prohibitive $O(n^2)$ memory and time. A second approximation reduces the matrix–vector product cost by interpreting the formation of the product as an n -body potential calculation [47]. This interpretation enables the use of techniques such as multipole methods [47, 209, 255, 257, 259, 260], or multiscale methods [258], to reduce the solution costs to $O(n)$ or $O(n \log n)$. Multipole methods require specialized expansions for every governing equation, and expansions for the LPBE have been developed in recent years [271]. One disadvantage of the fast multipole method (FMM) in particular is that the computational costs grow rapidly when improving accuracy [272] due to dense translation operations between multipole and local expansions, motivating the development of more efficient techniques [272–274].

Another challenge for the BEM is that the computation of elements in the dense systems of linear equations requires the integration of possibly singular functions over the panels used to discretize the boundary surfaces. These integrations can be inter-

interpreted as the calculation of the potential at an evaluation point due to a charge distribution defined on a boundary element. In contrast, the matrix elements for FDM and FEM problems are relatively easily computed. Although analytical expressions exist for the integral of the Laplace (Poisson) kernel over flat triangular panels [275,276], integration of the the LPBE kernel, or integration over general curved domains, require numerical approximation. When the evaluation point is sufficiently far from the panel, quadrature rules can be used to perform numerical integration, even over curved panels [252,254]. However, when the evaluation point is near or on the panel, even high-order quadrature rules do not suffice to capture the singularity. The evaluation of near-singular and singular integrals has been noted to be a limiting factor in the accuracy of BEM implementations for molecular electrostatics [256], and a variety of techniques have been developed to either avoid computing these integrals [256] or to approximate them with specialized quadrature rules [277].

In this paper we present a boundary-element method implementation for solving the linearized Poisson–Boltzmann equation (LPBE) that addresses all of these challenges, with the ultimate goal of achieving high accuracy given reasonable computational resources. Three key features underlie the implementation. First, we have developed a general boundary-integral approach that can easily treat an arbitrary number of embedded regions of homogeneous dielectric with different dielectric constants and possibly salt. Second, the accessible and molecular surfaces are discretized using curved boundary elements that accurately capture the problem geometry, employing robust methods to compute self- and near-field integrals. Third, the dense linear systems are solved using preconditioned Krylov subspace methods and the FFTSVD algorithm [274].

Our Green’s-theorem-based integral-equation formalism allows for ion-exclusion layers, solvent-filled cavities in the solute, and multiple homogeneous dielectric regions. Finite-difference and finite-element simulations have long been capable of mod-

eling problems with these features, but this paper presents the first detailed derivation for BEM treatment. The accessible and molecular surfaces are represented essentially exactly using curved boundary-element discretizations that accurately reproduce singularities, cusps, and thin regions. Accurate numerical integration techniques for the singular Laplace (Poisson) and LPBE Green's functions [270] allow the BEM to achieve exceptional accuracy. The FFTSVD algorithm [274] efficiently sparsifies the dense BEM matrix, and memory and time requirements scale effectively linearly in the number of boundary elements. This fast BEM technique can be applied without modification to compress all of the integral operators in biomolecule electrostatics. Furthermore, the dense translation operations that dominate the FMM computational cost are replaced in the FFTSVD method with more efficient diagonal translations, allowing for a better tradeoff between computational expense and accuracy.

After describing the boundary-element implementation, we present a set of computational experiments in order to assess the relative accuracy and computational cost of finite-difference and boundary-element method simulations for several categories of calculations. We calculate the electrostatic contributions to free energies of solvation for an analytically solvable sphere geometry, a short peptide derived from an HIV-1 substrate site [43], and the barnase–barstar protein complex [278]. We also compute rigid and non-rigid electrostatic binding free energies for the wild-type barnase–barstar complex as well as three single mutants. Solvation calculations demonstrate that the BEM presented here provides better convergence as a function of compute time. Rigid-binding results also suggest that the BEM converges more rapidly than FDM. However, when comparing differential rigid binding energies between wild-type and mutant protein complexes, where the structure remains the same except at the site of mutation, even low-resolution finite-difference simulations seem to accurately capture this difference. The curved BEM regains an accuracy advantage for differential non-rigid binding calculations, suggesting that the accuracy of

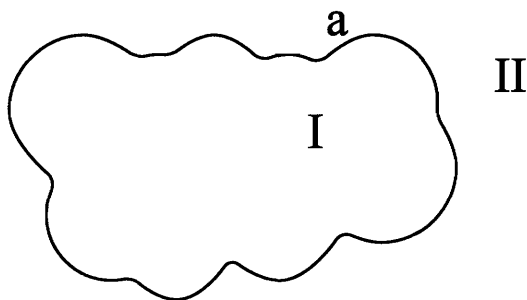


Figure 6-1: A one-surface problem in molecular electrostatics. The molecular interior (Region *I*) is surrounded by a salt solution with high dielectric constant and inverse Debye length κ (Region *II*).

finite-difference rigid binding may result from fortuitous cancellation of error. Finally, we demonstrate that the BEM implementation offers a clear advantage in accuracy and comparable simulation time for calculations that require repeated solution of the same problem geometry with different sets of atomic charges. Electrostatic component analysis [173, 279, 280] and charge optimization [23, 24] are types of calculations that fall into this category.

6.2 Theory

6.2.1 Green's theorem integral formulation

We begin our presentation of the multi-region integral formulation by deriving the one-surface Green's-theorem-based integral formulation described by Yoon and Lenhoff [207]. This method is also known as the non-derivative Green's theorem formulation [261, 269]. Figure 6-1 illustrates the problem and notation.

A single boundary a divides space into two regions. The molecular interior, labeled region *I*, has a uniform dielectric constant ϵ_I and contains n_c discrete point charges. The i^{th} point charge, located at r_i , is of value q_i . In region *I*, the electrostatic potential

$\phi_I(r)$ is governed by a Poisson equation

$$\nabla^2 \phi_I(r) = - \sum_{i=1}^{n_c} \frac{q_i}{\epsilon_I} \delta(r - r_i), \quad (6.1)$$

where $\delta(r - r_i)$ is the Dirac delta function translated by r_i .

The solvent region II exterior to a represents solvent with mobile ions; we model the region as having a uniform dielectric constant ϵ_{II} and an inverse Debye length κ . In this region, the electrostatic potential $\phi_{II}(r)$ is assumed to obey the linearized Poisson–Boltzmann equation:

$$\nabla^2 \phi_{II}(r) = \kappa^2 \phi_{II}(r). \quad (6.2)$$

The free-space Green's functions for the Poisson and linearized Poisson–Boltzmann equations are

$$G_I(r; r') = \frac{1}{4\pi \|r - r'\|} \quad \text{Region I} \quad (6.3)$$

$$G_{II}(r; r') = \frac{e^{-\kappa \|r - r'\|}}{4\pi \|r - r'\|} \quad \text{Region II} \quad (6.4)$$

respectively. Across the boundary surface a , the electrostatic potential and the normal displacement are continuous [281]. Using the relation $D = \epsilon E$, where the electric field E satisfies $E = -\nabla\phi$, we can write the continuity conditions for a point r_a on the surface a as

$$\phi_I(r_a) = \phi_{II}(r_a) \quad (6.5)$$

$$\epsilon_I \frac{\partial \phi_I}{\partial n}(r_a) = \epsilon_{II} \frac{\partial \phi_{II}}{\partial n}(r_a). \quad (6.6)$$

In Equation 6.6, the normal direction is defined to point into the solvent region.

After specifying the problem domains and boundary conditions, one applies

Green's theorem in both regions. Green's Theorem,

$$\int_V [\Psi \nabla^2 \Phi - \Phi \nabla^2 \Psi] dV = \int_\Omega \left[\Psi \frac{\partial \Phi}{\partial n} - \Phi \frac{\partial \Psi}{\partial n} \right] d\Omega, \quad (6.7)$$

where $\Psi(r)$ and $\Phi(r)$ are two scalar fields, allows the determination of the potential at a point in a volume V given the free-space Green's function for the governing equation in V as well as the potential and its normal derivative at the bounding surface Ω .

We first apply Green's theorem to find the potential at a point r_I in region I , which has the bounding surface $\Omega = a$. Using the Green's function (Equation 6.3) and substituting $\Psi(r') = G_I(r_I; r')$, $\Phi(r') = \phi_I(r')$, and Equation 6.1, we have

$$\begin{aligned} \int_V \left[G_I(r_I; r') \left(- \sum_{i=1}^{n_c} \frac{q_i}{\epsilon_I} \delta(r - r_i) \right) - \phi_I(r') \nabla^2 G_I(r_I; r') \right] dV' = \\ \int_a \left[G_I(r_I; r') \frac{\partial \phi_I}{\partial n}(r') - \phi_I(r') \frac{\partial G_I}{\partial n}(r_I; r') \right] dA'. \end{aligned} \quad (6.8)$$

In Equation 6.8 and throughout this section, the normal derivative of G_I is taken with respect to the integration variable r' : that is, $\frac{\partial G_I}{\partial n}(r_I; r')$ denotes the potential at r_I induced by a normally-oriented dipole at r' . Simplifying the left-hand side using the definition of the Green's function,

$$\nabla^2 G_I(r_I; r') = -\delta(r_I - r'), \quad (6.9)$$

eliminates the volume integral in Equation 6.8, and by rearranging terms one obtains an expression for the potential at r_I as a function of the solute charge distribution and the boundary conditions:

$$\phi_I(r_I) = \sum_{i=1}^{n_c} \frac{q_i}{\epsilon_I} G_I(r_I; r_i) + \int_a \left[G_I(r_I; r') \frac{\partial \phi_I}{\partial n}(r') - \phi_I(r') \frac{\partial G_I}{\partial n}(r_I; r') \right] dA'. \quad (6.10)$$

To apply Green's theorem in region II , one must first bound the region by in-

roducing a hypothetical surface Γ at infinity, and using the substitutions $\Psi(r') = G_{II}(r_{II}; r')$, $\Phi(r') = \phi_{II}(r')$, Equation 6.4, and the LPBE Green's function definition. Assuming the potential obeys regularity conditions at infinity [253], the surface integrals over Γ vanish, and we can write the potential at a point r_{II} in region II as

$$\phi_{II}(r_{II}) = \int_a \left[G_{II}(r_{II}; r') \frac{\partial \phi_{II}(r')}{\partial n} - \phi_{II}(r') \frac{\partial G_{II}(r_{II}; r')}{\partial n} \right] dA', \quad (6.11)$$

and here, as in Equation 6.10, the normal direction is defined to point into region II .

We derive a pair of coupled integral equations by letting the points r_I and r_{II} approach a point r_a on the surface. Using Equation 6.10,

$$\begin{aligned} \phi_I(r_a) &= \lim_{r_I \rightarrow r_a} \phi_I(r_I) & (6.12) \\ &= \int_a G_I(r_a, r') \frac{\partial \phi_I(r')}{\partial n} dA' - \lim_{r_I \rightarrow r_a} \left[\int_a \phi(r') \frac{\partial G_I(r_I; r')}{\partial n} dA' \right] \\ &\quad + \sum_{i=1}^{n_c} \frac{q_i}{\epsilon_I} G_I(r_a; r_i). & (6.13) \end{aligned}$$

The second term in Equation 6.13 can be interpreted as the potential induced by a dipole layer of charge on the surface. Such a potential is discontinuous as the evaluation point crosses the surface and must be handled with care. We write

$$\phi_I(r_a) = \int_a \left[G_I(r_a; r') \frac{\partial \phi_I(r')}{\partial n} - \phi_I(r') \frac{\partial G_I(r_a; r')}{\partial n} \right] dA' + \frac{1}{2} \phi_I(r_a) + \sum_{i=1}^{n_c} \frac{q_i}{\epsilon_I} G_I(r_a; r_i), \quad (6.14)$$

where f represents a Cauchy principal value integral, and we assume that the limit as $r_I \rightarrow r_a$ has been taken from the direction opposite the normal. A similar limiting process applied to Equation 6.11, in which we let $r_{II} \rightarrow r_a$, yields

$$\phi_{II}(r_a) = \int_a \left[-G_{II}(r_a; r') \frac{\partial \phi_{II}(r')}{\partial n} + \phi_{II}(r') \frac{\partial G_{II}(r_a; r')}{\partial n} \right] dA' + \frac{1}{2} \phi_{II}(r_a). \quad (6.15)$$

Finally, we eliminate the unknowns $\phi_{II}(r_a)$ and $\frac{\partial \phi_{II}(r_a)}{\partial n}$ using the continuity con-

ditions (Equations 6.5 and 6.6). Two coupled integral equations result:

$$\frac{1}{2}\phi_I(r_a) + \int_a \phi_I(r') \frac{\partial G_I}{\partial n}(r_a; r') dA' - \int_a \frac{\partial \phi_I}{\partial n}(r') G_I(r_a; r') dA' = \sum_{i=1}^{n_c} \frac{q_i}{\epsilon_I} G_I(r_a; r_i) \quad (6.16)$$

$$\frac{1}{2}\phi_I(r_a) - \int_a \phi_I(r') \frac{\partial G_{II}}{\partial n}(r_a; r') dA' + \frac{\epsilon_I}{\epsilon_{II}} \int_a \frac{\partial \phi_I}{\partial n}(r') G_{II}(r_a; r') dA' = 0. \quad (6.17)$$

Introducing an abbreviated notation allows the equations to be written as

$$\begin{bmatrix} \frac{1}{2}I + D_{I,a}^a & -S_{I,a}^a \\ \frac{1}{2}I - D_{II,a}^a & \epsilon_{I,II} S_{II,a}^a \end{bmatrix} \begin{bmatrix} \phi_a \\ \frac{\partial \phi_a}{\partial n} \end{bmatrix} = \begin{bmatrix} \sum_i \frac{q_i}{\epsilon_I} G_{I,i}^a \\ 0 \end{bmatrix}, \quad (6.18)$$

where ϕ_a and $\frac{\partial \phi_a}{\partial n}$ denote the surface potential and normal displacement on a , I denotes the identity operator, $\epsilon_{I,II}$ abbreviates $\frac{\epsilon_I}{\epsilon_{II}}$, and $S_{I,v}^u$ and $D_{I,v}^u$ denote the single- and double-layer operators that compute potential at the surface u due to a monopole or dipole charge density on surface v , given the Green's function $G_I(r; r')$. The operator $S_{I,v}^u$ is defined such that:

$$S_{I,v}^u \frac{\partial \phi_v}{\partial n} = \int_v G_I(r_u; r') \frac{\partial \phi_v}{\partial n(r')} (r') dA'; \quad (6.19)$$

similarly,

$$D_{I,v}^u \phi_v = \int_v \frac{\partial G_I}{\partial n(r')} (r_u; r') \phi_v(r') dA'. \quad (6.20)$$

In Equation 6.18, we have also defined $G_{I,i}^a = G_I(r_a; r_i)$.

6.2.2 Numerical solution using the boundary-element method

To simultaneously solve Equations 6.16 and 6.17 using the boundary-element method (BEM), we first approximate the surface variables $\phi_I(r_a)$ and $\frac{\partial \phi_I}{\partial n}(r_a)$ as weighted

combinations of a set of n basis functions $\chi_1(r), \chi_2(r), \dots, \chi_n(r)$ on the surface:

$$\phi_I(r_a) \approx \sum_{k=1}^n u_k \chi_k(r_a) \quad (6.21)$$

$$\frac{\partial \phi_I}{\partial n}(r_a) \approx \sum_{k=1}^n v_k \chi_k(r_a). \quad (6.22)$$

The unknown weights u_k and v_k are then found by forcing the integral equation to be satisfied as closely as possible in some choice of metric.

In this work, we discretize the surfaces into a discrete set of n_p non-overlapping curved boundary elements and use piecewise-constant basis functions that have a value of one on a single panel and are zero everywhere else:

$$\chi_k(r_a) = \begin{cases} 1 & \text{if } r_a \text{ is on panel } k \\ 0 & \text{otherwise.} \end{cases} \quad (6.23)$$

Defining the integral equation residual to be the difference between the known condition on the surface and the integral operator applied to the approximate solution, one can form a square linear system by forcing the residual to equal zero at the boundary-element centroids, a technique known as centroid collocation [282]. Using the piecewise-constant basis functions and denoting the centroid of panel i as r_{c_i} , the discretized (matrix) form of the operator $S_{I,a}^a$ from Equation 6.19 has entries

$$S_{i,j} = \int_{\text{panel } j} G_I(r_{c_i}; r') dA'_j, \quad (6.24)$$

and the double-layer discretized operator $D_{I,a}^a$ similarly has entries

$$D_{i,j} = \int_{\text{panel } j} \frac{\partial G_I}{\partial n(r')} (r_{c_i}; r') dA'_j. \quad (6.25)$$

The total matrix equation representing the discretized form of Equation 6.18 therefore has dimension $2n_p$. Once this equation is solved, the potential anywhere in space may

be calculated using the discretized forms of Equations 6.10 and 6.11.

6.2.3 Extension to multiple dielectrics, solvent cavities, and ion-exclusion layers

Continuum electrostatics models of biomolecular systems can be defined by multiple embedded regions of differing homogeneous dielectric constant and salt treatment. Integral-equation formulations that can solve these problems often possess a complicated block structure because there exist numerous operators that couple variables on one surface to conditions on other surfaces. To illustrate this block structure, we next present Green's theorem formulations for two-surface and three-surface example problems. We then describe how a tree-based representation of the enclosed regions facilitates the determination of the appropriate Green's-theorem-based integral operator for arbitrary multi-region problems.

Two-surface formulation

Figure 6-2 is a schematic of a two-surface problem in molecular electrostatics; salt ions are not permitted to directly reach the molecular surface a , but instead are bounded by an accessible surface b a specified distance outside the molecule. The enclosed volume between the surfaces is termed the ion-exclusion layer. Region I , again representing the molecular interior, has dielectric constant ϵ_I and n_c point charges. The ion-exclusion layer, region II , has dielectric constant ϵ_{II} , and in this region, the Laplace equation governs the electrostatic potential. Region III represents solvent with mobile ions and has dielectric constant ϵ_{III} (usually the same as ϵ_{II}) but contains a Debye-Hückel salt treatment; the potential in this region is governed by the linearized Poisson-Boltzmann equation. This problem has continuity conditions at

both a and b :

$$\phi_I(r_a) = \phi_{II}(r_a) \quad (6.26)$$

$$\epsilon_I \frac{\partial \phi_I}{\partial n}(r_a) = \epsilon_{II} \frac{\partial \phi_{II}}{\partial n}(r_a) \quad (6.27)$$

$$\phi_{II}(r_b) = \phi_{III}(r_b) \quad (6.28)$$

$$\epsilon_{II} \frac{\partial \phi_{II}}{\partial n}(r_b) = \epsilon_{III} \frac{\partial \phi_{III}}{\partial n}(r_b). \quad (6.29)$$

The associated integral equations have four surface variables, which are the potential and normal derivative on both surfaces: ϕ_a , $\frac{\partial \phi_a}{\partial n}$, ϕ_b , $\frac{\partial \phi_b}{\partial n}$. The free-space Green's functions in each region are again denoted by G with the region label as subscript: $G_{II}(r; r')$, for instance, denotes the free-space Laplace Green's function. As in the one-surface derivation, we apply Green's theorem in each region using the appropriate substitutions, let the field points approach the bounding surfaces, and eliminate redundant variables using the continuity conditions. The resulting operator takes the form

$$\left[\begin{array}{cc|cc} \frac{1}{2}I + D_{I,a}^a & -S_{I,a}^a & & \\ \frac{1}{2}I - D_{II,a}^a & +\epsilon_{I,II}S_{II,a}^a & +D_{II,b}^a & -S_{II,b}^a \\ \hline -D_{II,a}^b & +\epsilon_{I,II}S_{II,a}^b & \frac{1}{2}I + D_{II,b}^b & -S_{II,b}^b \\ & & \frac{1}{2}I - D_{III,b}^b & +\epsilon_{II,III}S_{II,b}^b \end{array} \right] \begin{bmatrix} \phi_a \\ \frac{\partial \phi_a}{\partial n} \\ \phi_b \\ \frac{\partial \phi_b}{\partial n} \end{bmatrix} = \begin{bmatrix} \sum_i \frac{q_i}{\epsilon_I} G_{I,i}^a \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad (6.30)$$

which can be solved with the boundary-element method described above.

Note that the integral operator contains several zero blocks. These blocks arise from the application of Green's theorem in regions for which one or more surfaces do not form part of that region's bounding surface. For instance, surface b forms no portion of the bounding surface for region I , and consequently variables on surface b contribute nothing to the integral equation derived by applying Green's theorem in region I . Note also that two of the integral equations derive from the application of

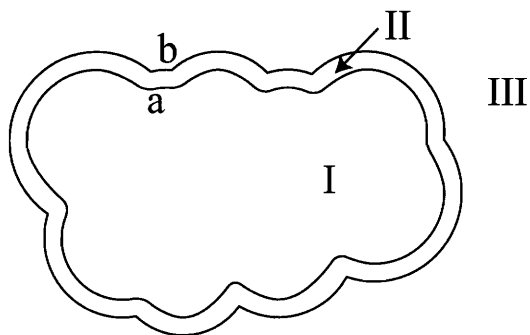


Figure 6-2: A two-surface problem in molecular electrostatics. The molecular interior (Region *I*) is surrounded by an ion-exclusion layer with solvent dielectric and no salt (Region *II*), which in turn is surrounded by solvent with a salt treatment (Region *III*).

Green's theorem in region *II*.

Three surface formulation

To identify more general trends in the construction of multi-boundary integral operators, we extend the two surface formulation by adding a solvent-filled cavity inside the protein interior (Figure 6-3). In this problem and for the remainder of this section, we will follow the convention that region *I* is the outermost solvent region. The additional region *IV* has dielectric constant ϵ_{IV} (generally equal to ϵ_I and ϵ_{II}), and is not large enough to contain an ion-exclusion surface. Again, we apply Green's theorem in every region, take limits on the surface integrals as the field points approach the boundaries, and enforce continuity conditions. The resulting operator takes the form

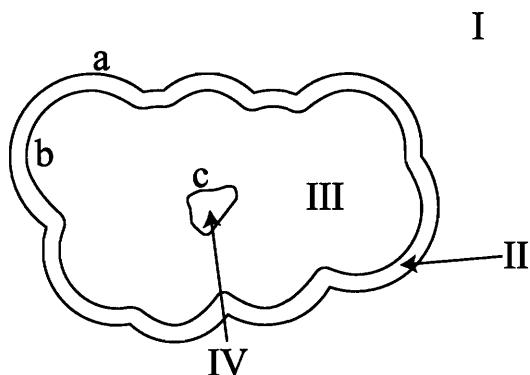


Figure 6-3: A three-surface problem in molecular electrostatics. This geometry is analogous to the two-surface problem (Figure 6-2) except that a solvent-filled cavity has been added within the molecular interior (Region *IV*). Note that in contrast to previous examples, the regions and surfaces have been labeled in reverse order.

erator taking into account all necessary interactions. Each node of the tree represents one region, and is associated with a dielectric constant and possibly salt treatment or point charges. The tree is constructed such that the node for a given region *X* is assigned to be the child of the node corresponding to the region surrounding *X*. Region *I*, which is bounded only by a hypothetical surface at infinity, is defined to be the root node. Furthermore, we associate with each node the exterior bounding surface of the corresponding region. Figure 6-4B is a tree diagram constructed to describe the system shown in Figure 6-4A.

The example geometry used here may be representative of an encounter complex in protein-protein binding, where two nearly associated binding partners (Regions *III_a* and *III_b*) are surrounded by a single ion-exclusion layer (Region *II*). There are also several solvent-filled cavities present in both binding partners (Regions *IV_{a-c}*), and one cavity is large enough to contain a small ion-exclusion layer (Region *V*).

Applying the multi-surface integral operator

A multi-region electrostatics problem with *n* surfaces generates a system of $2n$ coupled integral equations. For each surface, one writes Green's theorem for the regions exterior and interior to the surface and takes the appropriate limits as the evaluation

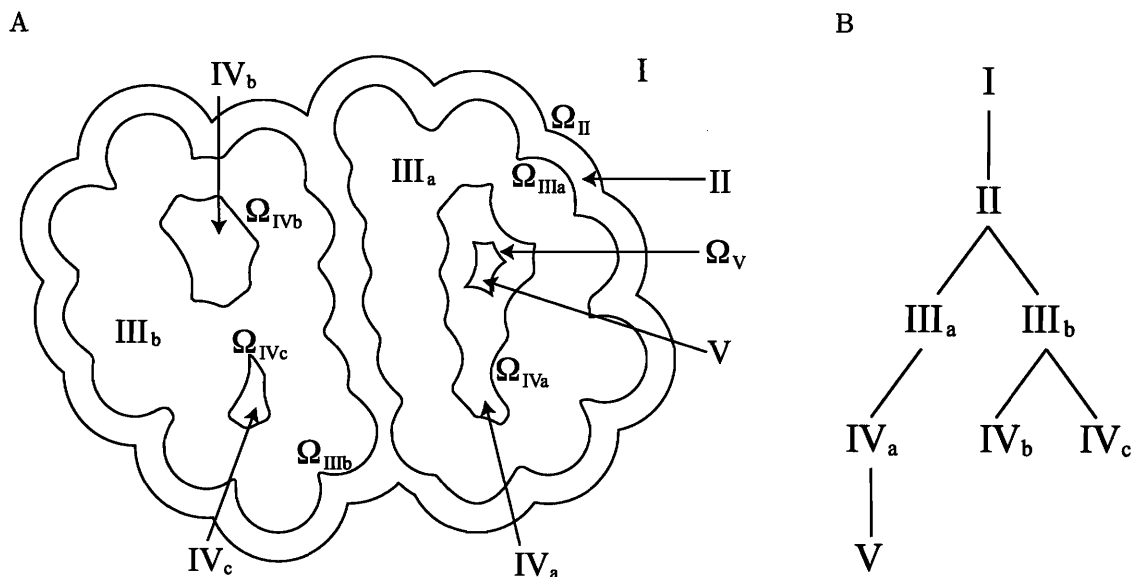


Figure 6-4: Tree representation of a general surface problem. The example molecular geometry shown in (A) might correspond to an encounter complex between two associating proteins (Regions III_a and III_b), surrounded by a single ion-exclusion layer (Region II), which in turn is surrounded by solvent with salt (Region I). The binding partners contain several solvent filled cavities (Regions IV_{a-c}), and one cavity is large enough to contain a small ion-exclusion layer (Region V). The tree representation for this example multi-surface geometry is shown in (B).

points approach the surface. Accordingly, one may refer to the resulting integral equations as the exterior and interior equations corresponding to the surface.

An integral equation derived from an application of Green's theorem contains contributions from the surfaces bounding the region. As an example, consider the interior equation for surface Ω_{IIIb} . Applying Green's theorem in region III_b defines the potential at a point in this region as a function of the surface potential and its normal derivative on Ω_{IIIb} , Ω_{IVb} , and Ω_{IVc} . Taking the limit of the Green's theorem expression as the field point approaches Ω_{IIIb} , we obtain the interior equation. Clearly, a surface's interior equation contains contributions from the surface as well as its children. Similarly, a surface's exterior equation contains contributions from the surface, its parent, and its siblings. This can be seen by letting the field point approach any of the cavity surfaces.

Multi-surface problems demand that careful attention be paid to the definition of

the surface normal. In this work we follow the mathematical convention that a normal always points outward from the finite volume enclosed by the surface. To apply the entire multi-surface operator for an arbitrary problem, we first define a tree such as shown in Figure 6-4B. The tree is traversed depth first, and at each node we apply several integral operators, which in the discretized problem correspond to dense block matrix–vector multiplications. Because each block multiplication may be interpreted as the computation of the potential at a surface due to a distribution of monopole or dipole charge on another surface, we refer to the two surfaces as the *source surface* and the *destination surface*. The set of block multiplications is determined by the topology of the surfaces, and is defined such that every non-zero block in the integral operator is applied exactly once.

We define four types of block integral operators: the self-surface interior operator, the self-surface exterior operator, the non-self interior operators, and the non-self exterior operators. As previously discussed, each operator represents an interaction between two surfaces. The labels *interior* and *exterior* specify whether the integral operator arises from an application of Green’s theorem to the region interior or exterior to the source surface. The self and non-self operators are distinguished because the discontinuity in the self operator double-layer calculation requires specific treatment.

For every node, the following block matrix-vector multiplications are performed. Let the current node correspond to the region X . Denote its parent region by W , sibling regions by S_i , and child regions by Y_i . Lowercase letters correspond to the outer bounding surfaces for these regions. Every dense block is applied to the vector $\left(\phi_x, \frac{\partial \phi_x}{\partial n}\right)^T$.

1. Apply the self-surface interior operator

$$\left[\begin{array}{cc} \frac{1}{2}I + D_{X,x}^x & -S_{X,x}^x \end{array} \right] \quad (6.32)$$

and add the result to the node's interior equation.

2. Apply the self-surface exterior operator

$$\left[\frac{1}{2}I - D_{W,x}^x + \epsilon_{X,W} S_{W,x}^x \right] \quad (6.33)$$

and add the result to the node's exterior equation.

3. Apply the appropriate non-self exterior operator

$$\left[-D_{W,x}^w + \epsilon_{X,W} S_{W,x}^w \right] \quad (6.34)$$

and add the result to the interior equation of the *parent* node.

4. For each *sibling* node S_i , apply the appropriate non-self exterior operator

$$\left[-D_{W,x}^{s_i} + \epsilon_{X,W} S_{W,x}^{s_i} \right] \quad (6.35)$$

and add the result to the exterior equation of the sibling node.

5. For each *child* node Y_i , apply the appropriate non-self interior operator

$$\left[+D_{X,x}^{y_i} - S_{X,x}^{y_i} \right] \quad (6.36)$$

and add the result to the exterior equation of the child node.

6.2.4 Matrix compression with the FFTSVD algorithm

As discussed in the Introduction, boundary-element methods give rise to dense matrix equations whose solution by LU factorization or Gaussian elimination requires $O(n^3)$ time and $O(n^2)$ memory for a system with n unknowns. Combining Krylov-subspace iterative methods with fast-solver algorithms reduces these costs to nearly $O(n)$. The

Krylov method requires only a way to apply the matrix A to a vector; in contrast, LU factorization and Gauss elimination require explicit access to every entry of A . In this work, we use the FFTSVD algorithm [274] to rapidly apply the dense integral operators.

FFTSVD, like multipole methods, exploits the smooth decay of the Green's functions as the distance between source and evaluation point increases. Both types of methods use a spatial decomposition of the set of boundary elements to separate near-field interactions, which are computed exactly, from far-field or long-range interactions, which can be accurately approximated. The long-range interactions are approximated by projecting the dominant panel source distributions, computed using an approximate singular value decomposition (SVD), onto a grid. Grid-grid interactions are computed via the fast Fourier transform (FFT), and the dominant responses are interpolated back to the destination integral equation collocation points. An overview of the FFTSVD method is presented in Figure 6-5, and a fully detailed description of the algorithm can be found in reference [274].

For the general multi-boundary Green's theorem formulation, each node in the tree contains an FFTSVD-compressed operator that simultaneously stores both the single- and double-layer interactions between all panels that bound the region.

6.2.5 Preconditioning

It has been previously noted in the literature that the non-derivative Green's theorem formulation can lead to ill-conditioned systems of linear equations, especially with decreasing boundary-element size [259]. To address this issue, we have implemented preconditioning in order to efficiently solve these systems with iterative methods. By definition, a preconditioner is any matrix P such that the equation $PAx = Pb$ has better convergence properties than the original system $Ax = b$ when the systems are solved iteratively. In general, Krylov iterative methods are most efficient at solving

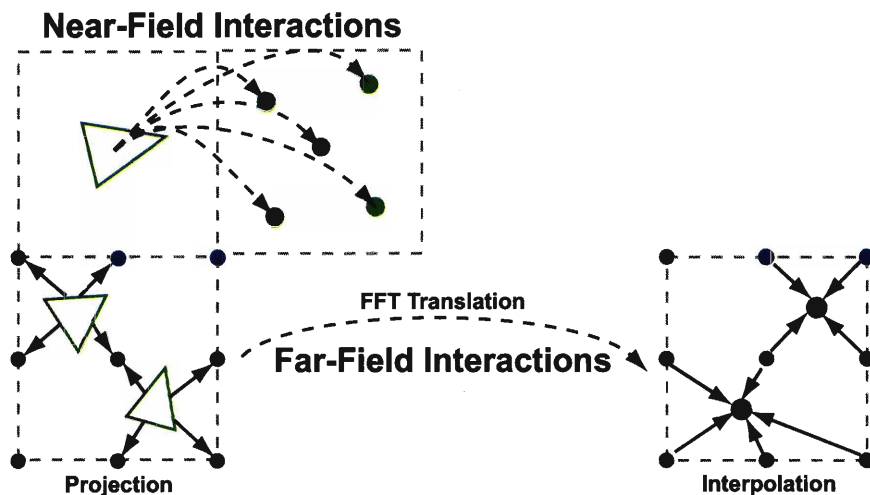


Figure 6-5: An overview of the FFTSVD matrix compression algorithm. FFTSVD uses a multi-level octree spatial decomposition to separate panel–evaluation point interactions into near- and far-field components at multiple length scales. When two cubes at the finest length scale are nearby, interactions are computed through direct integration. However, when two interacting cubes are well separated, dominant sources are projected onto a cubic grid and translated to a grid surrounding the recipient cube. The FFT is used to accelerate this translation operation. Finally, the grid potentials can be interpolated back onto the dominant responses of the panel centroids. This Figure has been adapted from [274].

linear systems with clustered eigenvalues [283]. Because the identity matrix I (or multiples) has an optimal clustering, P is generally selected such that $P \approx A^{-1}$ but is inexpensive to form and apply.

For the discretized integral operator matrices that arise from the Green’s theorem formulation, the dominant entries tend to be the self-influence terms, for which the evaluation point is on the element over which the integral is performed. Consequently, a reasonable choice for P is the inverse of a sparse matrix that contains only these self-term entries. As an examination of Equations 6.18, 6.30, and 6.31 should make clear, the sparse matrix that includes just the self-influence terms is not diagonal, but no row has more than two non-zero off-diagonals.

6.2.6 Curved panel discretization

In order to generate the basis functions used in the boundary-element method, we discretize the molecular and accessible surfaces that define the problem into curved elements that can exactly represent the underlying geometry [270]. Accessible surfaces [266], also called expanded van der Waals surfaces, are generally used to model the ion-exclusion layer and can be completely described by convex spherical patches bounded by circular arcs. These circular arcs are not necessarily geodesic arcs, and thus we use the concept of a generalized spherical triangle (GST) (Figure 6-6A) [270,284]. A GST is a three-sided curved element that lies on the surface of a sphere, where each edge is a portion of a circular arc. If the arc center for all three edges happens to be the center of the sphere, a traditional spherical triangle is recovered. A spherical patch can be discretized into a set of GSTs by starting with a flat element triangulation, and then assigning the appropriate circular arc to each element edge. Edges that lie along the interface between atoms are assigned non-geodesic arcs that follow the curve of intersection, while all other edges are assigned geodesic arcs.

Molecular surfaces [229, 267, 268], used here to model dielectric interfaces, are the surfaces of closest approach for the surface of a probe sphere that is rolled over a molecule. They can be described by three types of surface patches [267]. Convex spherical patches are defined where the probe sphere is in contact with only one atom, and can be described by portions of a sphere bounded by circular arcs and discretized with GSTs. Concave spherical re-entrant patches are formed when the probe touches three or more atoms simultaneously, and are also described by GSTs. When the probe simultaneously touches two atoms, a portion of a torus is generated. Toroidal regions are discretized into four-sided curved torus panels (Figure 6-6B) that are isomorphic to a rectangle. A fully meshed curved panel discretization for the barnase–barstar complex molecular surface is shown in Figure 6-7.

Techniques for integrating singular Green’s functions over these curved GST and

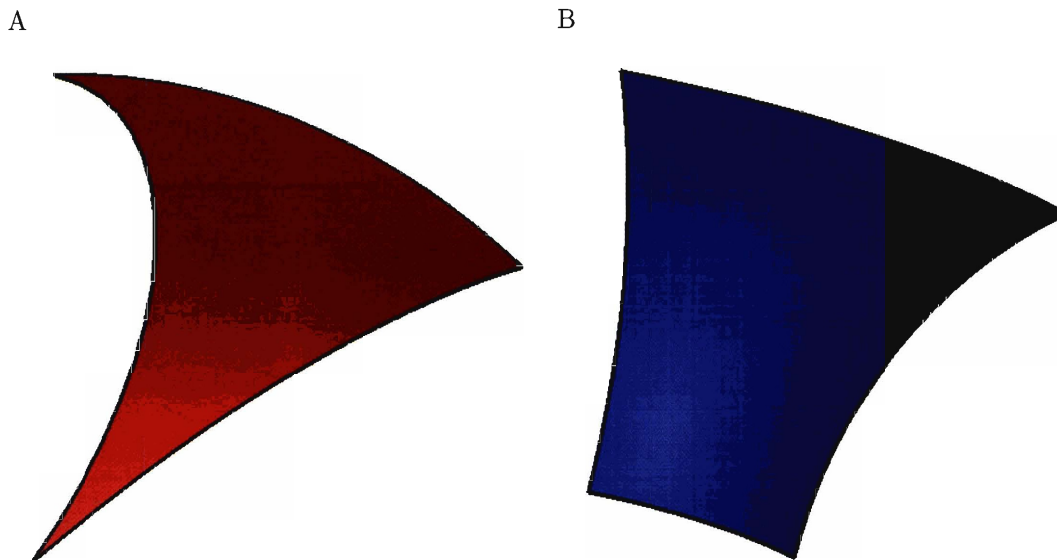


Figure 6-6: The two types of curved panels used to discretize accessible and molecular surfaces. A generalized spherical triangle (GST) (A), is a three-sided region on the surface of a sphere bounded by three circular arcs. These arcs are not necessarily geodesic arcs. Torus patches on molecular surfaces are discretized using toroidal panels (B), which are isomorphic to a rectangle.

torus panels have been developed, and are discussed in detail in [270]. Briefly, when the evaluation point in the integrand is far away from the panel, low-order quadrature rules are used to perform numerical integration. These quadrature rules are generated by creating a smooth mapping between a reference flat triangle or rectangle (for GSTs and torus panels respectively) that relates a known quadrature rule on these simple domains [285] to those applicable on the curved panels. When the evaluation point is near or on the curved panel, even high-order quadrature rules do not suffice to capture the singularity. As a result, we adopt specialized methods for each panel type and Green's function. For the single-layer Laplace (Poisson) kernel, we integrate over GSTs using a technique that reproduces the effect of panel curvature using a higher-order distribution on a reference flat triangle [286]. Single-layer Laplace integrals over torus panels are evaluated using a panel-splitting approach, which avoids integration near the singularity using recursive subdivision. When integrating the double-layer Laplace kernel in the near-field over both GST and torus panels, we exploit the fact

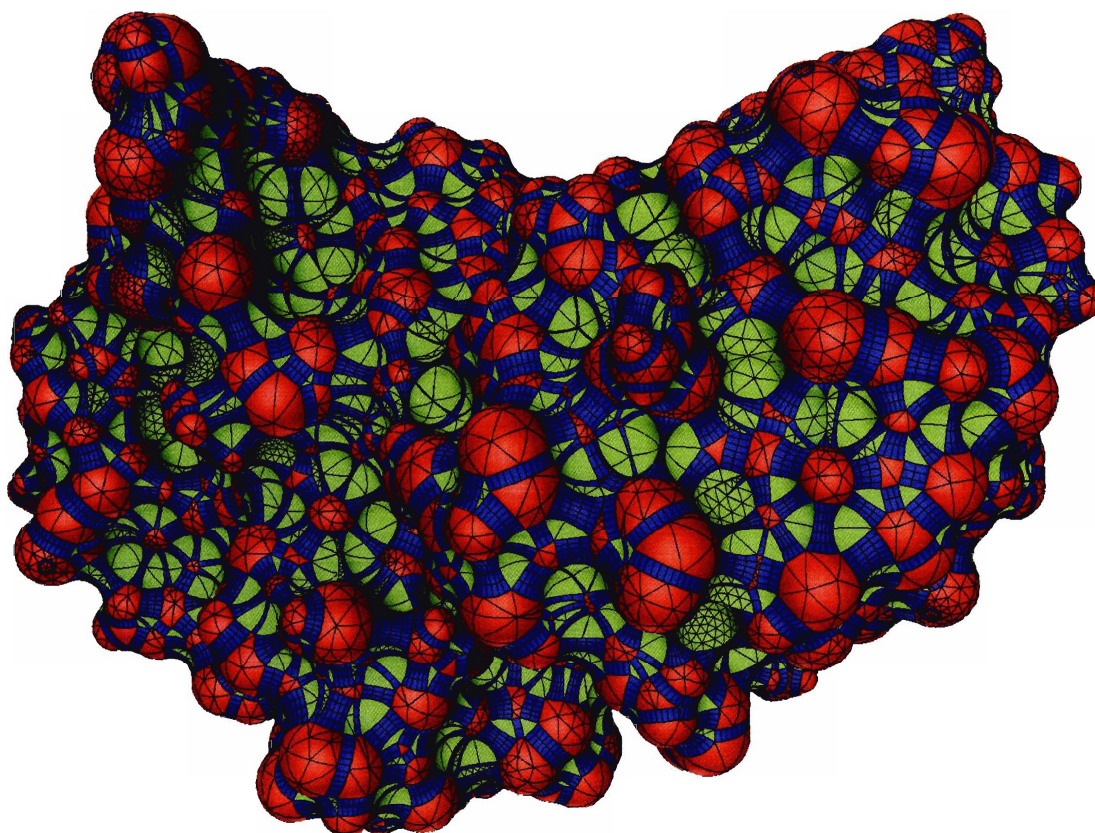


Figure 6-7: A rendering of a curved panel discretization for the molecular surface of the barnase–barstar protein complex. Red regions indicate convex spherical patches, green regions are re-entrant spherical patches, and blue regions are toroidal patches. Black lines indicate the boundaries between panels. The graphic depicts an approximation to the discretized geometry used for calculation. Every GST and torus panel has been approximated by a very large number of flat triangles for the purpose of visualization only, and the true surface normal in conjunction with Phong shading have been used to render the image.

that the double-layer potential is equal to the solid angle subtended by the curved panel when observed from the evaluation point [281, 287]. In order to integrate the linearized Poisson–Boltzmann kernel or its normal derivative in the near field, we adopt a previously presented desingularization technique [259]. This method divides the integral into a singular Laplace component that can be integrated as described above, and a smooth component that can be integrated using quadrature.

6.3 Computational details

6.3.1 Peptide and protein structure preparation

The structure of a peptide derived from an HIV-1 protease cleavage site was obtained from the Protein Data Bank (PDB) with accession code 1F7A [43]. This structure contains nine visible residues of a decameric peptide bound to an inactivated mutant of HIV-1 protease; only the peptide was considered in further calculations. An N-terminal acetyl blocking group and a C-terminal methylamide blocking group were added to the peptide. The wild-type structure for the barnase–barstar protein complex was also obtained from the PDB using accession code 1BRS [278]. To prepare this structure for calculation, we followed a previous protocol [31] where all but a set of 12 interfacial water molecules were removed. For both the peptide and barnase–barstar structures, hydrogen atoms were added using the HBUILD module [116] in the CHARMM computer program [117] using the PARAM22 parameter set [175] and a distance-dependent dielectric constant of 4. In addition, side-chain atoms that were missing from the crystal structures were rebuilt using CHARMM and the default PARAM22 geometry. All ionizable residues were left in their standard states at pH 7.

6.3.2 Modeling of barnase–barstar mutations

Three point mutations (E73Q in barnase, D39A in barstar, and T42A in barstar) were built into the barnase–barstar complex for subsequent analysis. The alanine mutations were created by cutting back the wild-type residue to the β -carbon. The E73Q mutation was built by sampling glutamine side-chain dihedral angles in 30-degree increments using CHARMM [117] and the PARAM22 parameter set [175]. For each sampled conformation, the side chain was energy minimized until convergence keeping all other atoms in the structure fixed. The lowest energy minimized geometry was taken to represent the E73Q mutation.

6.3.3 BEM and FFTSVD parameters

Parameters used in the FFTSVD algorithm included a drop tolerance of 10^{-5} for SVD compression, spatial decomposition until each cube contained no more than 32 panels, and a grid size of $4 \times 4 \times 4$ in each finest-level cube to represent dominant sources and responses during FFT translation. The boundary-element matrix equations were solved using the Krylov subspace method GMRES [228] to a relative residual of 10^{-6} . All curved BEM calculations were performed on a 2-way dual-core 2.0 GHz Opteron machine running a parallel version of the FFTSVD library. All presented timings are the sum of CPU usage across all four processors.

6.3.4 Finite-difference solver and parameters

In order to compare our curved-panel boundary-element solver to finite-difference methods (FDM), we have implemented a FDM solver using previously described techniques [28] and an analytical surface representation. This implementation uses successive over-relaxation (SOR) with an optimized acceleration factor to solve the finite-difference equations to a relative residual of 10^{-6} . In order to handle truncation

of the boundary condition at infinity, a focusing scheme [26] was employed in all FDM calculations where the molecule of interest occupied first 23% and then 92% of the finite-difference grid. For the low-percent fill run, a Debye–Hückel screened potential in solvent dielectric was used to assign potentials to the boundary of the cubic grid. For the high-percent fill run, boundary potentials were taken by interpolation from the low-percent fill solution. Although it is common to average results from multiple translations of the molecule relative to the grid in order to reduce error due to the grid representation [26], only one placement was used here to make a fair comparison to the curved BEM, which is insensitive to translations or rotations of the geometry. Cubic grids used to discretize molecular geometries in the FDM spanned 129 to 481 grid points per side in increments of 32, which are all solvable within 4 GB of computer memory. These sizes correspond to grid resolutions of approximately 2.3 to 8.6 grid points per Angstrom for the barnase–barstar complex. All FDM calculations were performed in serial on a 2-way dual-core 2.0 GHz Opteron machine.

6.3.5 Electrostatic solvation and binding calculations

All continuum electrostatics calculations were performed using a molecular dielectric constant of 4, a solvent dielectric constant of 80, a molecular surface with probe radius 1.4 Å for dielectric interfaces, an accessible surface with probe radius 2.0 Å for ion-exclusion layers, and an ionic strength of 145 mM. In order to compute the solvation free energy of a molecule, we take the difference between the energy of the solvated state and a reference state where the solvent dielectric constant is equal to the molecular dielectric constant and no salt is present. The BEM calculates this energy difference directly, and an explicit reference state is not needed. In the FDM implementation, the energy of the reference state is explicitly computed to cancel grid energy.

For rigid-binding calculations, the electrostatic component of the free energy of

binding was computed as the sum of Coulombic interactions in the bound state and the differential solvation energy between the bound complex and infinitely separated individual binding partners. For the FDM, proper grid placement was used to accelerate the calculation by canceling the grid energy in the complex with grid energies for the individual binding partners. Because the BEM only computes the reaction potential rather than the total electrostatic potential, the Coulombic interactions between the binding partners must be explicitly added.

Non-rigid electrostatic binding energies were computed by first energy minimizing the geometry of the complex and each of the isolated binding partners separately. The minimization was performed using CHARMM and the PARAM22 parameter set, relaxing all atoms with 1,000 steps of adapted basis Newton–Raphson (ABNR) minimization using a distance-dependent dielectric constant of 4. The binding energy was then computed using a thermodynamic cycle where the two isolated binding partners were first desolvated to a vacuum with the molecular dielectric constant. In vacuum, the partners were deformed to their bound-state structures and then rigidly bound, computing all electrostatic changes with Coulomb’s law in molecular dielectric. Finally, the complex was re-solvated. The sum of the energetic changes in these three steps was taken as the non-rigid electrostatic binding free energy. Due to the change in geometry between the bound and unbound states in non-rigid binding, the FDM grid energy cancellation technique could not be used, and explicit reference states were employed for all FDM solvation calculations.

6.3.6 Generating curved panel discretizations

Molecular and accessible surfaces were discretized into curved panels starting with high-quality flat triangular meshes for spherical regions from the program NETGEN [288]. These panels were then converted, along with torus patches, to curved panels using previously described methods [270]. Curved-panel discretizations for

molecular geometries were generated such that memory requirements did not exceed 4 GB. For the sphere test case, discretizations were obtained between roughly 80 and 58,000 curved panels including ion-exclusion and dielectric interface surfaces. For the peptide example, panel counts spanned approximately 5,200 to 128,000, and for the various barnase–barstar complexes, the span was roughly 92,000 to 310,000 total curved panels.

6.4 Results and Discussion

For all calculations, we compared our boundary element results to those generated using finite-difference methods. Although geometric measures can be defined for such comparisons [255, 258], we chose to use compute time as our metric, to determine which method can achieve superior convergence properties given a certain amount of time. We could not guarantee that the geometry of the problem being solved was exactly the same in both methods because different algorithms were used to generate molecular boundaries. Therefore, for systems without closed-form solutions, the level of convergence for a particular method was assessed solely on how little the solution changed as the compute time increased.

6.4.1 Electrostatic solvation free energies

One of the simplest linearized Poisson-Boltzmann calculations is the computation of the electrostatic component of the free energy associated with the transfer of a molecule from low- to high-dielectric medium, where the high-dielectric region contains an ion-exclusion layer with salt outside. This quantity, known as the electrostatic solvation free energy, is useful in many calculations and forms the basis for computing more complex quantities such as electrostatic binding energies. We first validated the multi-surface formulation by computing the solvation free energy for a simple spheri-

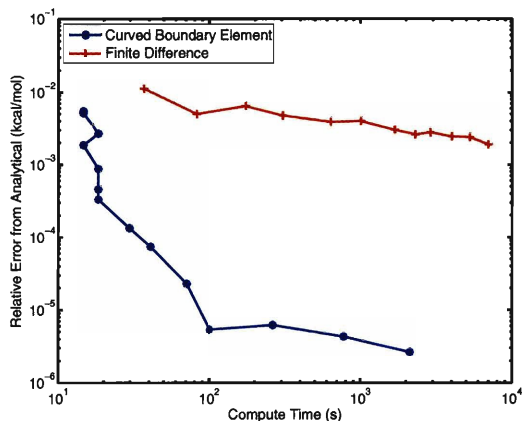


Figure 6-8: Convergence plot for the solvation free energy for a sphere with an eccentric charge and ion-exclusion layer. The relative error from the analytical solution is plotted as a function of compute time. Results are compared between the curved BEM and FDM implementations. The curved BEM accuracy is limited to 5–6 digits given the settings used in the FFTSVD matrix compression.

cal test case, which has a closed-form solution. Then, we gauged the accuracy of the solver by examining more complicated geometries including a peptide derived from an HIV-1 substrate site and the barnase–barstar protein–protein complex.

Sphere with ion-exclusion layer

In order to test the correctness of the multiple surface formulation, the electrostatic solvation free energy for a sphere of radius 1.0 Å with a charge of $+1e$ placed 0.5 Å away from the center was computed. An ion-exclusion layer was added 2.0 Å outside the sphere surface, creating a two boundary problem. BEM and FDM solutions were compared to the analytical solvation energy for this geometry [24] to generate the convergence plot shown in Figure 6-8.

From the sphere convergence results, it is clear that the curved BEM method is able to achieve superior accuracy given the same amount of compute time as the finite difference method. For this problem, the FDM is limited to 2–3 digits of accuracy, even when using resolutions greater than 50 grid points per Angstrom. The limited ability of finite-difference methods to achieve high accuracy has been noted previously

in the literature [289], although we obtain better than 1% accuracy on this sphere example. The accuracy of the curved BEM is limited to 5–6 digits given the settings selected in the FFTSVD matrix compression procedure. Additional accuracy can be achieved by increasing the size of the grids used to represent long-range interactions, at the expense of additional computational cost.

HIV-1 protease substrate peptide

To evaluate the method on a more complex example, the electrostatic solvation free energy for a peptide derived from an HIV-1 substrate site was computed using BEM and FDM including salt and an ion-exclusion layer. The computed solvation energy was plotted as a function of compute time (Figure 6-9A). It is clear from examining Figure 6-9A that the solutions provided by the curved BEM implementation seem more converged than those obtained from the FDM. Although it is unclear whether the two methods will converge to the same answer for this complex geometry, the solution at the finest discretization levels for the curved BEM are changing by as little as 10^{-3} kcal/mol, while those from FDM are still changing on the order of tenths of kcal/mol.

Barnase–barstar complex

In order to be competitive with finite-difference methods, the curved boundary-element method presented here must be able to achieve high accuracy per unit compute time on large macromolecules, where the number of curved panels required to discretize the geometry can be large. To test the solver on a moderately sized protein system, we computed the solvation free energy of the barnase–barstar protein complex [278, 290, 291], a model binding system for which electrostatic interactions have been shown to be important [31, 291–294]. In addition to an ion-exclusion layer, the problem geometry included four solvent-filled cavities inside the main dielectric

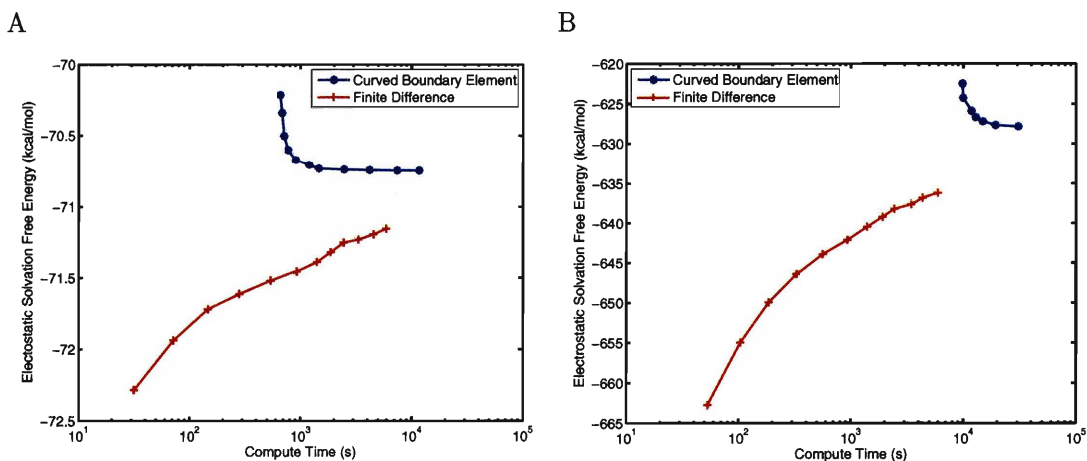


Figure 6-9: Computed solvation free energies, using curved BEM and FDM, for an HIV-1 substrate peptide (A) and the barnase–barstar complex (B). The absolute electrostatic solvation free energy is plotted as a function of compute time, and the selected discretizations used up to 4 GB of computer memory.

boundary. A comparison between the BEM and FDM for computing the absolute solvation energy of this complex is shown in Figure 6-9B. Even the finest BEM and FDM discretizations that can be solved on a computer with 4 GB of memory give answers that differ by 8–9 kcal/mol. Furthermore, it is difficult to establish whether the two methods will converge to the same answer. However, the curved BEM profile does appear to be relatively flat, even though the solution changed by approximately 0.2 kcal/mol between the two highest-resolution calculations.

As can be seen in Figure 6-9B, even the lowest BEM discretization obtained for the barnase–barstar complex requires more compute time than the highest discretization used for the FDM. The timings for the FDM remain relatively constant across the presented problems because they depend primarily on the grid size. In contrast, the BEM requires more curved panels to discretize a larger molecular surface, resulting in significantly increased simulation cost.

The accuracy of the BEM scales with the panel density; accordingly, the larger barnase–barstar complex cannot be discretized at the same level as was feasible for the peptide example. The BEM-calculation solvation energies in Figures 6-9A and

6-9B exhibit similar curvature, and the “knees” of the two curves are separated by approximately a factor of ten in compute time. This difference is as expected considering the ratio of the surface areas for the peptide and barnase–barstar complex (952 Å² and 8019 Å² respectively). The level of FDM convergence might also be expected to suffer for larger problems due to decreasing grid resolution given the same number of grid points. Surprisingly, the FDM appeared to lose less relative accuracy with increasing problem size as compared to the BEM. For the peptide and barnase–barstar solvation energies, the highest resolution FDM calculations were still changing by approximately 0.05 and 0.5 kcal/mol respectively. In the curved BEM results, they were changing by 0.001 and 0.2 kcal/mol, indicating a larger fold loss in convergence.

6.4.2 Importance of preconditioning

To demonstrate how effectively the block-diagonal preconditioner accelerates convergence of the iterative solution of the BEM equations, we repeated the solvation energy calculation for one discretization of the peptide example using several preconditioners. Specifically, we performed the calculation without preconditioning, with a purely diagonal preconditioner, and with the presented block-diagonal preconditioner. As shown in Figure 6-10, the number of GMRES iterations required to achieve a relative residual of 10^{-6} without preconditioning was 422. The purely diagonal preconditioner required 198 iterations, and the full block-diagonal preconditioner reduced this even further to 40 iterations. The block-diagonal preconditioner generally allows even complex geometries such as proteins to be solved to a relative residual of 10^{-6} in approximately 100 GMRES iterations or less.

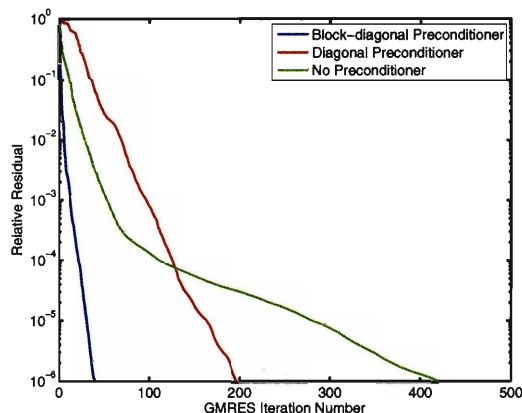


Figure 6-10: Comparison of preconditioning strategies when solving for the electrostatic solvation free energy of an HIV-1 protease substrate peptide discretized with 18,657 and 7,089 panels on the dielectric and ion-exclusion surfaces respectively. The block-diagonal preconditioner significantly reduces the number of GMRES iterations required to solve the linear system of BEM equations to a relative residual of 10^{-6} .

6.4.3 Rigid electrostatic binding free energies

Another useful quantity often calculated using the LPBE model is the rigid electrostatic binding free energy between a pair of interacting molecules. One component of this quantity is the difference in solvation energy between the bound state and two unbound states where the binding partners are rigidly separated to infinity. This differential electrostatic solvation is added to the direct Coulombic interactions made between the partners in the bound state. To measure the role that LPBE solver accuracy plays in this class of calculations, as well as compare the curved BEM to FDM, we computed the rigid electrostatic binding free energies for the wild-type barnase–barstar complex and three experimentally characterized single mutants (E73Q in barnase, T42A and D39A in barstar) [291, 295, 296] that have been previously shown to have a significant effect on electrostatic binding calculations [292, 297–299]. These mutations were built into the wild-type barnase–barstar complex with minimal perturbation, where all atoms remained in the same position except at the site of mutation.

The results of these rigid electrostatic binding calculations are shown in Figure 6-

11. For the wild-type barnase–barstar structure as well as the mutant complexes, the BEM calculations showed smaller changes in the computed energies with increasing problem discretization.

6.4.4 Differential rigid electrostatic binding free energies between mutants and wild type

Often, when comparing a set of protein mutations to identify those with improved electrostatic properties, one is more interested in the relative electrostatic rigid binding free energies as compared to wild type than the absolute binding energies themselves. To gauge the effect of solver accuracy on relative binding free energies, we calculated the difference in rigid electrostatic binding free energy between each mutant and the wild type at every level of problem discretization (Figure 6-12).

For all mutants studied, both methods appear to be converged to tenths of kcal/mol or better, and give very similar relative binding energies. Low discretizations of the FDM provide solutions very close to the final answer in a very short amount of time. This may be due to error cancellation because the mutant structures differ little from the wild type. For problems in which electrostatic energies are being compared between structures for which most atoms are located at identical positions, finite-difference methods may be a better choice than the boundary-element method presented here. Minimal-perturbation relative-binding calculations are often used when making predictions to improve protein binding or stability, especially in the field of protein design [67–69].

6.4.5 Non-rigid electrostatic binding free energies

The rigid binding model, although a useful approximation, is deficient in that it does not allow structural relaxation in the bound and unbound states. Consequently, a

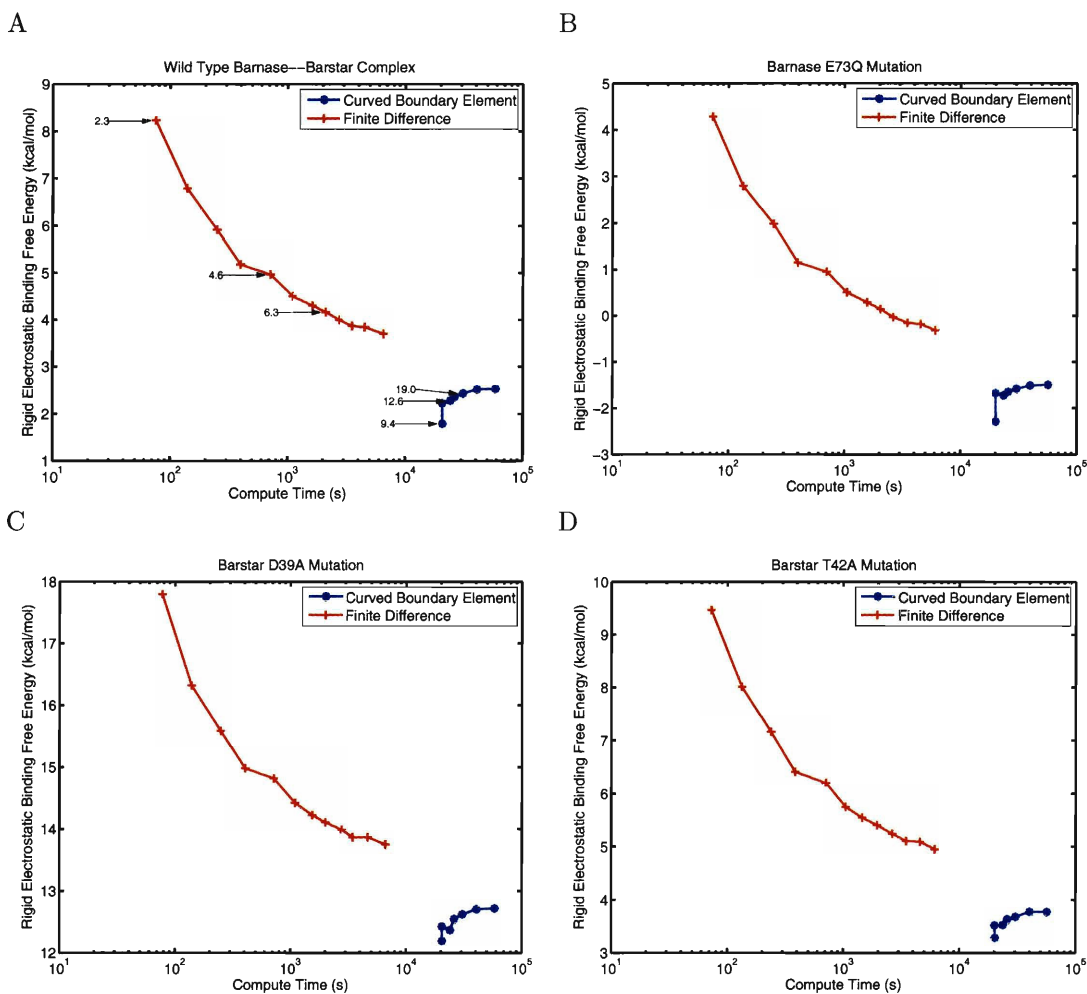


Figure 6-11: Comparison between curved BEM and FDM for computing the electrostatic component of the rigid binding free energy between the wild-type barnase–barstar complex (A), and three mutant complexes, E73Q in barstar (B), D39A in barnase (C), and T42A in barnase (D). The binding energy obtained is plotted as a function of the compute time required. In (A), several FDM and BEM results are labeled with their discretization level (grid points per Angstrom or panels per Angstrom², respectively).

variety of methods have been presented in the literature for treating non-rigid effects in protein–protein binding using continuum electrostatics [300, 301]. One feature most techniques share is that there is no longer a direct correspondence between the majority of atomic coordinates in the bound and unbound states. As a result, we hypothesized that the FDM would no longer be able to take advantage of cancellation of error when computing non-rigid binding effects, and that the accuracy of the overall calculation would depend strongly on the ability to independently converge the solvation energy for each state. To test this idea, we implemented a crude non-rigid binding scheme involving independent minimization of the complex and unbound binding partners and a thermodynamic cycle to compute electrostatic energies. The non-rigid electrostatic binding energies for mutants were subtracted from those for the wild-type barnase–barstar complex to generate non-rigid relative binding energies.

As shown in Figure 6-13, the curved BEM method regains an accuracy advantage in non-rigid binding calculations. The curves in this plot resemble those from absolute binding energy calculations (Figure 6-11). The finite-difference solution does not appear to be well converged at low resolution, and seems to gradually approach the boundary-element solution.

Because grid cancellation could not be exploited in non-rigid binding to avoid reference state calculations in the FDM, we computed the solvation of each state independently allowing the protein complex or binding partners to fill the entire finite-difference grid. Therefore, when subtracting the solvation energies of binding partners from the bound complex, we were subtracting calculations solved at very different grid resolutions. To determine if this was responsible for the inability of FDM to converge relative non-rigid electrostatic binding energies, we repeated the calculation using fixed grid placement to ensure that the solvation energy of each state was computed at roughly the same number of grid points per Angstrom. However, this modification did not improve the ability of FDM to converge relative non-rigid

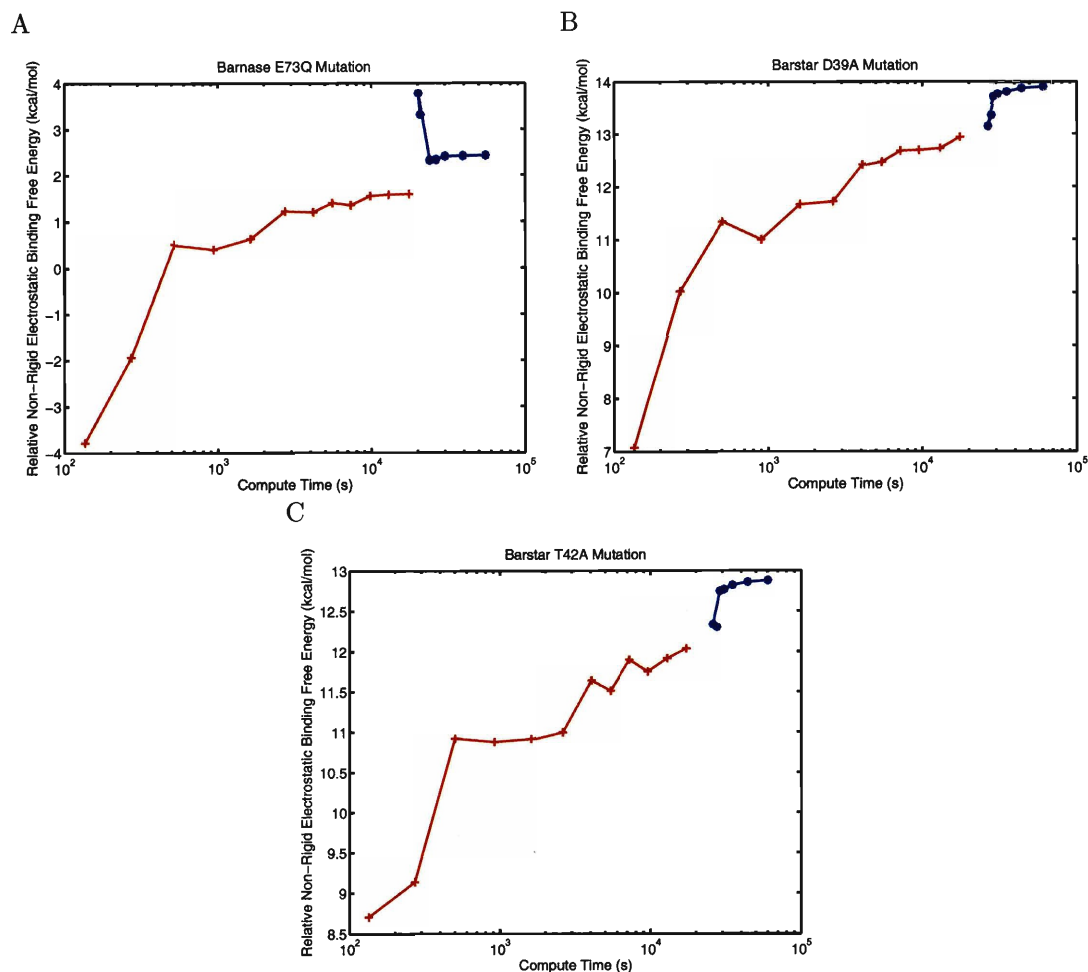


Figure 6-13: Comparison between curved BEM and FDM for computing relative non-rigid electrostatic binding energies between mutant and wild-type barnase–barstar complexes. Results are shown for the mutations E73Q on barstar (A), D39A on barnase (B), and T42A on barnase (C) The relative binding energy is plotted as a function of the compute time for the mutant complex non-rigid binding energy.

binding energies (data not shown).

6.4.6 Multiple electrostatic solves for the same problem geometry

As shown in the previous Results sections, the curved BEM, although offering better convergence properties, is quite time consuming on large geometries such as proteins. The dominant computational cost in our implementation is compressing the integral operators using the FFTSVD algorithm, which primarily involves computing costly

integrals over nearby curved panels. In contrast, the FDM requires very little time to initialize the system of linear equations and spends almost all compute time solving them. However, there exist several types of useful electrostatic calculations that involve multiple simulations of the same problem geometry; for these problems, the expensive BEM “setup” time can be amortized over all calculations.

One such example is charge optimization [23,24,31], which determines the optimal partial atomic charges for a ligand that minimize the electrostatic component of its binding free energy with a receptor molecule. In charge optimization, two geometries for the ligand are considered: the bound state, where it is complexed with the receptor molecule, and the unbound state, where it is isolated in solution. Each ligand charge is set to $+1e$ independently, leaving all others at zero, and one determines the difference in solvation potential at the ligand charge locations between the bound and unbound states by solving the LPBE twice. This produces the ligand desolvation matrix, an important component of the charge optimization equation [23,24]. Overall, $2n$ solves of the LPBE are required, where n is the number of atoms in the ligand. When using the BEM, each state’s integral operator only needs to be compressed once, and the compressed operator can be used to solve the n right-hand sides that only depend on the atomic charges.

To compare the performance of the curved BEM and FDM on a charge optimization problem, we computed the ligand desolvation matrix for barstar in the wild-type barnase–barstar complex. In total, 1403 simulations were performed in each of the bound and unbound states. In Table 6.1 we report the time required to compute the ligand desolvation matrix for three discretization levels of the finite-difference and curved boundary-element methods. The panel densities and grid spacings mentioned in Table 6.1 may be compared to the labeled points on the absolute binding free energy plot shown in Figure 6-11A.

For the finer discretization BEM calculations, the compute time is comparable to

Table 6.1: Compute time required to calculate the entries of the ligand desolvation matrix for barnase in the wild-type barnase–barstar complex. For both the curved BEM and FDM, the calculation was repeated at three discretization levels. For the curved BEM, the panel density reported is for all surfaces in the bound state geometry.

Method		
FDM	Grid Points Per Angstrom	Time (s)
	2.3	41,868
	4.6	637,930
	6.3	1,774,146
Curved BEM	Panels Per Angstrom ²	Time (s)
	9.4	755,343
	12.6	1,347,300
	19.0	2,024,024

that required for the finer FDM discretizations. Relating these discretization levels to the convergence plot suggests that for these multiple-solve problems, the BEM may offer superior accuracy for similar computational cost.

6.5 Conclusions

In conclusion, we have presented an implementation of the boundary-element method for linearized Poisson–Boltzmann continuum electrostatics that is capable of achieving high accuracy and solving the same topologies of dielectric boundaries, point charges, and salt regions that volume-based methods are capable of solving. Several techniques were employed to overcome several of the well-known practical limitations of the BEM. These included a general Green’s-theorem integral formulation for multiple embedded regions, curved panel discretization with robust integration methods, and preconditioned Krylov subspace methods combined with matrix compression using the FFTSVD algorithm.

Comparing the performance of the curved BEM against a reference finite-difference solver identified types of calculations for which improved accuracy may be important. For example, when computing absolute electrostatic solvation free energies or the

electrostatic component of rigid binding energies, the curved BEM method offers superior convergence properties. Even at the highest discretizations possible within 4 GB of computer memory, finite-difference methods did not appear to be converged, as the solutions continued to change significantly with increased expenditure of computing resources. However, when comparing differential rigid binding energies between mutant and wild-type protein complexes, even coarse finite-difference simulations sufficed to capture relative effects. This is not surprising considering that the local structural perturbations allow for cancellation of error. Relative rigid binding calculations with local geometry perturbations are prevalent in ranking the results of molecular design efforts [173], and finite-difference methods are an attractive tool for this class of computation. However, when non-rigid effects were introduced into the binding model, and the bound and unbound states were allowed to relax independently, finite-difference methods lost their convergence advantage. Therefore, as more sophisticated non-rigid models of binding are employed in ranking results of molecular design calculations, higher accuracy LPBE solvers such as the presented curved BEM may become necessary to make reliable predictions.

In the current implementation, the computational resources required to obtain solutions converged to tenths of kcal/mol on protein geometries are somewhat higher than what would be commonly available on a desktop workstation at this time. In order to compute a well converged protein solvation or binding energy in a few hours, a workstation with four processors and 4 GB of memory are currently required. Because the problem geometry is already represented essentially exactly, it is likely that the primary source of error in the method arises from the use of piecewise-constant representations of the surface variables. Higher-order basis functions may allow a significant reduction in the number of unknowns, and thus the required memory. However, two complications that may limit higher-order methods are that the numerical integrations are more time consuming, and that the compressibility of the

discretized operator may decrease. It is not yet clear where the optimal tradeoff lies between basis function complexity and these complications, and improvements in this area should be capable of reducing the time and memory usage of the curved BEM implementation to more accessible levels.

Chapter 7

General Conclusions

In conclusion, several computational techniques were developed and/or applied to solve problems in molecular design and the evaluation of energetic properties in molecular systems. A small-molecule design strategy, based on inverse methods and combinatorial search, was successfully developed and validated on several test systems. This work utilized an inverse phrasing of the small-molecule design problem to reduce the search space of potential ligands down to solving the functional group positioning problem on a set of discrete molecular scaffolds placed throughout the binding site. Grid-based energy functions and a pairwise solvation approximation played an important role in rapidly and accurately approximate binding energies. To account for the inaccuracies involved in the combinatorial search scoring function, more sophisticated binding energy models that were non-pairwise and slow to compute were applied in a hierarchical fashion to identify top-ranking compounds in a full binding energy model.

Validation of the method in engineered binding sites within the core of T4 lysozyme identified strengths and limitations of the presented inverse design strategy. Given the scoring and parameter sets currently used, the inverse strategy appears to do a better job optimizing the shape components of the binding free energy rather than the electrostatic components. As was observed in the T4 lysozyme system, polar groups were often buried for their favorable packing interactions and did not make hydrogen bonding interactions with the site. From the work presented here and other computational experiments, it is possible that the van der Waals component of the current energy function may not be suitably parameterized for the burial of polar

groups in hydrophobic pockets, or the electrostatic desolvation penalties of the polar groups may be underestimated. In order to avoid this problem in the design of HIV-1 protease inhibitors, a hydrogen bond filter was added to the design protocol to completely eliminate functional groups that bury unsatisfied polar groups. However, this solution to the problem is based on phenomenology and not physics, and is not likely to be a robust solution to the problem. Improving the binding energy function in a physics-based approach would be a fruitful avenue of future research, as the inverse design method is capable of using any scoring function that is pairwise additive.

The inverse design strategy was able to partially overcome this energy function issue because the complete nature of the combinatorial search approach allowed the shape and electrostatic contributions to the binding energy to be decoupled for a large portion of ligand space. Compounds that scored the most favorably in electrostatics were more similar to compounds known to bind in both the T4 lysozyme and *E. coli* chorismate mutase systems. The idea of selecting compounds from the search space that ranked the best in electrostatics and moderately well in packing was a recurring theme throughout all of the inverse design work. Validation in the HIV-1 protease system demonstrated another advantage of the inverse approach, which is that it allowed a fair comparison to be made between the fitness of two combinatorial libraries for targeting the protease active site because of a consistent sampling procedure and complete search.

Although the theoretical validation presented in Chapter 2 was encouraging, it was insufficient to prove that the inverse design method could propose novel compounds that would bind tightly experimentally. Therefore, the inverse design method was applied in a biologically relevant problem for which collaborators were willing to chemically synthesize and experimentally test computationally designed molecules. The goal of the collaboration was to test the substrate envelope hypothesis strategy as an inhibitor design principle to avoid drug resistance using HIV-1 protease as a

model system. Using the known structures of HIV-1 protease–substrate complexes, it was postulated that inhibitors that remained inside the consensus substrate volume would be less likely to induce drug resistance because mutations that prevent inhibitor binding would also affect substrate processing. The inverse small-molecule design technique was applied using the substrate envelope instead of an optimal ligand shape, a molecular scaffold similar to those present in several protease inhibitors, and a naive functional group library to diversify the scaffold at three positions.

Twenty designed compounds were proposed for synthesis, and fifteen of these were actually synthesized and experimentally tested for binding in the wild-type HIV-1 protease. All fifteen had observable binding to the protease, with the top four compounds exhibiting inhibition constants (K_i) between 30–50 nM. Clinically approved inhibitors range in binding from 0.002–0.1 nM. It is difficult to determine whether this represents a significant accomplishment for computational techniques. The designed inhibitors relied on a well-known molecular scaffold, and several groups predicted to be very favorable for binding, such as the isoxazole ring, ended up being poor for binding experimentally. However, considering other factors, such as the functional group library used in this first round of design possibly missing important classes of functional groups, and the fact that the crystal structures of designed compounds agree well with prediction, it is the belief of the author that the inverse design method is capable of enriching compound libraries for tight binding in practice.

When the four top binding designed compounds were tested for inhibition against three drug resistant mutant proteases, the compounds exhibited relatively flat resistance profiles, losing no more than 6–13 fold activity compared to wild type. These numbers are similar to the clinical inhibitor amprenavir (11 fold), and much less than the clinical inhibitor lopinavir (1,220 fold). Although the designed compounds do well in these three particular mutants, it could be argued that the reason why they have broad specificity is that they share a similar scaffold to amprenavir. For future

work, it would be very important to test the designed compounds in an amprenavir-specific resistance mutant, containing the I50V mutation, which has been reported to selectively reduce amprenavir binding by more than 100 fold [302]. If the designed compounds still perform well in this mutant, it would suggest that the functional groups chosen by the substrate envelope calculation, and not the scaffold chosen from the literature, were responsible for broad specificity.

Although it has not been formally demonstrated, there is a notion in the HIV research community that protease inhibitors with lower affinity to the wild type, such as those designed here, may inherently have broader specificity. In order to avoid this issue, it will be necessary to design tighter binding inhibitors that stay within the substrate envelope. Tighter binding compounds would also be extremely useful for cell-based assays and *in vitro* passaging experiments to see if it is possible to evolve resistance against envelope inhibitors. One possible way to achieve this goal through computational design would be to apply the inverse method using a substrate envelope placed inside a receptor structure derived from a complex with an inhibitor that uses the same scaffold as the one used for design. This was shown in Chapter 3 to improve the ability of computational methods to correctly rank compound affinity. In addition, it would be useful to enrich the functional group library with compounds known to work well in the HIV-1 protease subsites. These avenues are currently being explored, and preliminary results are encouraging.

In Chapter 4, charge optimization techniques and inverse protein design methods were applied to improve the affinity of a substrate peptide for an inactivated (D25N) mutant of HIV-1 protease. The motivation behind this project was to obtain cleaner thermodynamic data that could be used to gain insight into the way that the substrates interact with the protease, which could in turn be used for inhibitor design. Of three peptides synthesized and experimentally tested for binding with calorimetry, the two that were expected to bind tighter did by 10–12 fold and 2–3 fold respec-

tively. The single threonine-to-valine mutation that led to the dramatic improvement in binding free energy avoided burying an unsatisfied polar group. Although peptide binding was measured to be entropically driven, it is unclear whether these results can be extended to the active protease, where the aspartyl dyad is present. In addition, we hoped that the crystal structure of the 10–12-fold tighter binding peptide would reveal fundamental differences in the way that tighter binding peptides interact with the protease that could be exploited for inhibitor design. Unfortunately, the crystal structure of the tight binding peptide did not appear to be statistically different from that of the wild type RT–RH substrate. Further work on this project might include measuring catalytic rates for the designed peptides to determine if they are substrates, or the synthesis and testing of additional peptides suggested by protein design to identify even tighter binders.

The curved-panel boundary-element method solver presented in Chapters 5 and 6 describes an effort to improve accuracy when solving multi-region molecular continuum electrostatics problems with the linearized Poisson–Boltzmann equation (LPBE). Many practical challenges needed to be overcome to develop such a method, including the development of a matrix compression algorithm called FFTSVD, presented in Chapter 5. FFTSVD can sparsify the dense interaction matrices that arise in boundary-element formulations when solving partial differential equations such as the LPBE. The method combined the most sophisticated features from existing techniques into an algorithm that performed better, used less memory, and was capable of improving accuracy with less computational expense as compared to existing compression methods.

Using FFTSVD and other techniques, it was possible to develop an accurate curved boundary-element solver for the LPBE that was capable of solving multi-layer biomolecular continuum electrostatics problems to high accuracy. The method compared favorably to existing finite-difference approaches in terms of accuracy, demon-

strating that for many classes of electrostatic problems, errors in solution accuracy may dominate the predictions made from the model. However, one limitation of the current implementation was that its computational cost was somewhat high. Future work on this project should be dedicated to improving the accuracy per unit time of the solver. Several possible ways to achieve this goal include using higher-order basis functions on the curved elements, using alternative integral formulations that are better conditioned, improving the speed of computing near-singular or singular integrals over curved panels, or developing more uniform meshing schemes that avoid high panel densities in local regions.

Bibliography

- [1] C. Levinthal. Are there pathways for protein folding? *J. Chim. Phys.*, 65:44–45, 1968.
- [2] K. A. Dill and H. S. Chan. From Levinthal to pathways to funnels. *Nat. Struct. Biol.*, 4:10–19, 1997.
- [3] D. S. Goodsell, G. M. Morris, and A. J. Olson. Automated docking of flexible ligands: Applications of AutoDock. *J. Mol. Recognit.*, 9:1–5, 1996.
- [4] T. J. A. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aid. Mol. Des.*, 15:411–428, 2001.
- [5] B. K. Shoichet, S. L. McGovern, B. Q. Wei, and J. J. Irwin. Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.*, 6:439–446, 2002.
- [6] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, and P. S. Shenkin. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.*, 47:1739–1749, 2004.
- [7] W. P. Walters, M. T. Stahl, and M. A. Murcko. Virtual screening — An overview. *Drug Discov. Today*, 3:160–178, 1998.
- [8] R. Abagyan and M. Totrov. High-throughput docking for lead generation. *Curr. Opin. Chem. Biol.*, 5:375–382, 2001.
- [9] K. E. Drexler. Molecular engineering: An approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci. U.S.A.*, 78:5275–5278, 1981.
- [10] C. Pabo. Molecular technology: Designing proteins and peptides. *Nature*, 301:200–200, 1983.
- [11] J. Desmet, M. Demaeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356:539–542, 1992.

- [12] B. I. Dahiyat, C. A. Sarisky, and S. L. Mayo. De novo protein design: Towards fully automated sequence selection. *J. Mol. Biol.*, 273:789–796, 1997.
- [13] A. R. Leach and A. P. Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins*, 33:227–239, 1998.
- [14] D. B. Gordon, G. K. Hom, S. L. Mayo, and N. A. Pierce. Exact rotamer optimization for protein design. *J. Comp. Chem.*, 24:232–243, 2003.
- [15] G. Dantas, B. Kuhlman, D. Callender, M. Wong, and D. Baker. A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.*, 332:449–460, 2003.
- [16] P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, and P. S. Kim. High-resolution protein design with backbone freedom. *Science*, 282:1462–1467, 1998.
- [17] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302:1364–1368, 2003.
- [18] B. Kuhlman, J. W. O’Neill, D. E. Kim, K. Y. J. Zhang, and D. Baker. Conversion of monomeric protein L to an obligate dimer by computational protein design. *Proc. Natl. Acad. Sci. U.S.A.*, 98:10687–10691, 2001.
- [19] J. M. Shifman and S. L. Mayo. Modulating calmodulin binding specificity through computational protein design. *J. Mol. Biol.*, 323:417–423, 2002.
- [20] J. J. Havranek and P. B. Harbury. Automated design of specificity in molecular recognition. *Nat. Struct. Biol.*, 10:45–52, 2003.
- [21] L. L. Looger, M. A. Dwyer, J. J. Smith, and H. W. Hellinga. Computational design of receptor and sensor proteins with novel functions. *Nature*, 423:185–190, 2003.
- [22] D. Sitkoff, K. A. Sharp, and B. Honig. Accurate calculation of hydration free-energies using macroscopic solvent models. *J. Phys. Chem.*, 98:1978–1988, 1994.
- [23] L. P. Lee and B. Tidor. Optimization of electrostatic binding free energy. *J. Chem. Phys.*, 106:8681–8690, 1997.
- [24] E. Kangas and B. Tidor. Optimizing electrostatic affinity in ligand-receptor binding: Theory, computation, and ligand properties. *J. Chem. Phys.*, 109:7522–7545, 1998.
- [25] M. K. Gilson and B. Honig. Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis. *Proteins*, 4:7–18, 1988.

- [26] M. K. Gilson, K. A. Sharp, and B. H. Honig. Calculating the electrostatic potential of molecules in solution: Method and error assessment. *J. Comp. Chem.*, 9:327–335, 1988.
- [27] K. A. Sharp and B. Honig. Electrostatic interactions in macromolecules: Theory and applications. *Ann. Rev. Biophys. Biophys. Chem.*, 19:301–332, 1990.
- [28] A. Nicholls and B. Honig. A rapid finite-difference algorithm, utilizing successive over-relaxation to solve the Poisson–Boltzmann equation. *J. Comp. Chem.*, 12:435–445, 1991.
- [29] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268:1144–1149, 1995.
- [30] E. Kangas and B. Tidor. Charge optimization leads to favorable electrostatic binding free energy. *Phys. Rev. E*, 59:5958–5961, 1999.
- [31] L. P. Lee and B. Tidor. Optimization of binding electrostatics: Charge complementarity in the barnase–barstar protein complex. *Protein Sci.*, 10:362–377, 2001.
- [32] G. Stockman. Object recognition and localization via pose clustering. *Comput. Vision Graph.*, 40:361–387, 1987.
- [33] R. Nussinov and H. J. Wolfson. Efficient detection of 3-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc. Natl. Acad. Sci. U.S.A.*, 88:10495–10499, 1991.
- [34] H. J. Wolfson and I. Rigoutsos. Geometric hashing: An overview. *IEEE Comp. Sci. Eng.*, 4:10–21, 1997.
- [35] I. Pastan, M. Gottesman, C. R. Kahn, J. Flier, and P. Eder. Multiple-drug resistance in human cancer. *New Eng. J. Med.*, 316:1388–1393, 1987.
- [36] B. A. Larder and S. D. Kemp. Multiple mutations in HIV-1 reverse-transcriptase confer high-level resistance to zidovudine (AZT). *Science*, 246:1155–1158, 1989.
- [37] M. L. Cohen. Epidemiology of drug resistance: Implications for a postantimicrobial era. *Science*, 257:1050–1055, 1992.
- [38] A. Telenti, P. Imboden, F. Marchesi, D. Lowrie, S. Cole, M. J. Colston, L. Matter, K. Schopfer, and T. Bodmer. Detection of rifampicin resistance mutations in mycobacterium tuberculosis. *Lancet*, 341:647–650, 1993.
- [39] M. Tisdale, S. D. Kemp, N. R. Parry, and B. A. Larder. Rapid in vitro selection of human-immunodeficiency-virus type-1 resistant to 3'-thiacytidine inhibitors due to a mutation in the YMDD region of reverse-transcriptase. *Proc. Natl. Acad. Sci. U.S.A.*, 90:5653–5656, 1993.

- [40] J. H. Condra, W. A. Schleif, O. M. Blahy, L. J. Gabryelski, D. J. Graham, J. C. Quintero, A. Rhodes, H. L. Robbins, E. Roth, M. Shivaprakash, D. Titus, T. Yang, H. Teppler, K. E. Squires, P. J. Deutsch, and E. A. Emini. In-vivo emergence of HIV-1 variants resistant to multiple protease inhibitors. *Nature*, 374:569–571, 1995.
- [41] H. S. Gold and R. C. Moellering. Antimicrobial-drug resistance. *New Eng. J. Med.*, 335:1445–1453, 1996.
- [42] C. G. Whitney, M. M. Farley, J. Hadler, L. H. Harrison, C. Lexau, A. Reinhold, L. Lefkowitz, P. R. Cieslak, M. Cetron, E. R. Zell, J. H. Jorgensen, and A. Schuchat. Increasing prevalence of multidrug-resistant *Streptococcus pneumoniae* in the United States. *New Eng. J. Med.*, 343:1917–1924, 2000.
- [43] M. Prabu-Jeyabalan, E. Nalivaika, and C. A. Schiffer. How does a symmetric dimer recognize an asymmetric substrate? A substrate complex of HIV-1 protease. *J. Mol. Biol.*, 301:1207–1220, 2000.
- [44] M. Prabu-Jeyabalan, E. Nalivaika, and C. A. Schiffer. Substrate shape determines specificity of recognition for HIV-1 protease: Analysis of crystal structures of six substrate complexes. *Structure*, 10:369–381, 2002.
- [45] I. Luque, M. J. Todd, J. Gomez, N. Semo, and E. Freire. Molecular basis of resistance to HIV-1 protease inhibition: A plausible hypothesis. *Biochemistry*, 37:5791–5797, 1998.
- [46] A. Velazquez-Campoy, M. J. Todd, and E. Freire. HIV-1 protease inhibitors: Enthalpic versus entropic optimization of the binding affinity. *Biochemistry*, 39:2201–2207, 2000.
- [47] K. Nabors and J. White. FastCap: A multipole accelerated 3-D capacitance extraction program. *IEEE T. Comput. Aid. D.*, 10:1447–1459, 1991.
- [48] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *J. Chem. Phys.*, 73:325–348, 1987.
- [49] J. R. Phillips and J. K. White. A precorrected-FFT method for electrostatic analysis of complicated 3-D structures. *IEEE T. Comput. Aid. D.*, 16:1059–1072, 1997.
- [50] S. Kapur and D. E. Long. IES³: A fast integral equation solver for efficient 3-dimensional extraction. In *IEEE/ACM ICCAD*, pages 448–55, 1997.
- [51] W. L. Jorgensen. The many roles of computation in drug discovery. *Science*, 303:1813–1818, 2004.
- [52] G. Schneider and U. Fechner. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discov.*, 4:649–663, 2005.

- [53] H. J. Bohm. The computer-program LUDI: A new method for the de novo design of enzyme inhibitors. *J. Comput. Aid. Mol. Des.*, 6:61–78, 1992.
- [54] V. Gillet, A. P. Johnson, P. Mata, S. Sike, and P. Williams. SPROUT: A program for structure generation. *J. Comput. Aid. Mol. Des.*, 7:127–153, 1993.
- [55] S. H. Rotstein and M. A. Murcko. GroupBuild: A fragment-based method for de novo drug design. *J. Med. Chem.*, 36:1700–1710, 1993.
- [56] R. S. Bohacek and C. McMartin. Multiple highly diverse structures complementary to enzyme binding sites: Results of extensive application of a de novo design method incorporating combinatorial growth. *J. Am. Chem. Soc.*, 116:5560–5571, 1994.
- [57] D. E. Clark, D. Frenkel, S. A. Levy, J. Li, C. W. Murray, B. Robson, B. Waszkowycz, and D. R. Westhead. PRO-LIGAND: An approach to de novo molecular design. 1. application to the design of organic molecules. *J. Comput. Aid. Mol. Des.*, 9:13–32, 1995.
- [58] R. S. DeWitte and E. I. Shakhnovich. SMOG: De novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.*, 118:11733–11744, 1996.
- [59] D. A. Pearlman and M. A. Murcko. CONCERTS: Dynamic connection of fragments as an approach to de novo ligand design. *J. Med. Chem.*, 39:1651–1663, 1996.
- [60] N. P. Todorov and P. M. Dean. Evaluation of a method for controlling molecular scaffold diversity in de novo ligand design. *J. Comput. Aid. Mol. Des.*, 11:175–192, 1997.
- [61] D. Douguet, E. Thoreau, and G. Grassy. A genetic algorithm for the automated generation of small organic molecules: Drug design using an evolutionary algorithm. *J. Comput. Aid. Mol. Des.*, 14:449–466, 2000.
- [62] A. Miranker and M. Karplus. Functionality maps of binding sites: A multiple copy simultaneous search method. *Proteins*, 11:29–34, 1991.
- [63] M. B. Eisen, D. C. Wiley, M. Karplus, and R. E. Hubbard. HOOK: A program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding-site. *Proteins*, 19:199–221, 1994.
- [64] D. C. Roe and I. D. Kuntz. BUILDER v.2: Improving the chemistry of a de novo design strategy. *J. Comput. Aid. Mol. Des.*, 9:269–282, 1995.
- [65] A. Miranker and M. Karplus. An automated method for dynamic ligand design. *Proteins*, 23:472–490, 1995.

- [66] Y. Sun, T. J. A. Ewing, A. G. Skillman, and I. D. Kuntz. CombiDOCK: Structure-based combinatorial docking and library design. *J. Comput. Aid. Mol. Des.*, 12:597–604, 1998.
- [67] B. I. Dahiyat and S. L. Mayo. Protein design automation. *Protein Sci.*, 5:895–903, 1996.
- [68] A. G. Street and S. L. Mayo. Computational protein design. *Struct. Fold. Des.*, 7:R105–R109, 1999.
- [69] J. G. Saven. Combinatorial protein design. *Curr. Opin. Struct. Biol.*, 12:453–458, 2002.
- [70] N. A. Pierce, J. A. Spriet, J. Desmet, and S. L. Mayo. Conformational splitting: A more powerful criterion for dead-end elimination. *J. Comp. Chem.*, 21:999–1009, 2000.
- [71] A. R. Leach. Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.*, 235:345–356, 1994.
- [72] L. Schaffer and G. M. Verkhivker. Predicting structural effects in HIV-1 protease mutant complexes with flexible ligand docking and protein side-chain optimization. *Proteins*, 33:295–310, 1998.
- [73] P. Kallblad and P. M. Dean. Efficient conformational sampling of local side-chain flexibility. *J. Mol. Biol.*, 326:1651–1665, 2003.
- [74] E. C. Meng, B. K. Shoichet, and I. D. Kuntz. Automated docking with grid-based energy evaluation. *J. Comp. Chem.*, 13:505–524, 1992.
- [75] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comp. Chem.*, 19:1639–1662, 1998.
- [76] I. Muegge and Y. C. Martin. A general and fast scoring function for protein–ligand interactions: A simplified potential approach. *J. Med. Chem.*, 42:791–804, 1999.
- [77] H. Gohlke, M. Hendlich, and G. Klebe. Knowledge-based scoring function to predict protein–ligand interactions. *J. Mol. Biol.*, 295:337–356, 2000.
- [78] M. Stahl and M. Rarey. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.*, 44:1035–1042, 2001.
- [79] I. Halperin, B. Y. Ma, H. Wolfson, and R. Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, 47:409–443, 2002.
- [80] R. X. Wang, Y. P. Lu, and S. M. Wang. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.*, 46:2287–2303, 2003.

- [81] P. Ferrara, H. Gohlke, D. J. Price, G. Klebe, and C. L. Brooks III. Assessing scoring functions for protein–ligand interactions. *J. Med. Chem.*, 47:3032–3047, 2004.
- [82] B. K. Shoichet, A. R. Leach, and I. D. Kuntz. Ligand solvation in molecular docking. *Proteins*, 34:4–16, 1999.
- [83] X. Q. Zou, Y. X. Sun, and I. D. Kuntz. Inclusion of solvation in ligand binding free energy calculations using the generalized-Born model. *J. Am. Chem. Soc.*, 121:8033–8043, 1999.
- [84] D. S. Goodsell and A. J. Olson. Automated docking of substrates to proteins by simulated annealing. *Proteins*, 8:195–202, 1990.
- [85] B. A. Luty, Z. R. Wasserman, P. F. W. Stouten, C. N. Hodge, M. Zacharias, and J. A. McCammon. A molecular mechanics grid method for evaluation of ligand–receptor interactions. *J. Comp. Chem.*, 16:454–464, 1995.
- [86] G. S. Wu, D. H. Robertson, C. L. Brooks III, and M. Vieth. Detailed analysis of grid-based molecular docking: A case study of CDOCKER — A CHARMM-based MD docking algorithm. *J. Comp. Chem.*, 24:1549–1562, 2003.
- [87] A. Nicholls, K. A. Sharp, and B. Honig. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins*, 11:281–296, 1991.
- [88] N. Arora and D. Bashford. Solvation energy density occlusion approximation for evaluation of desolvation penalties in biomolecular interactions. *Proteins*, 43:12–27, 2001.
- [89] A. Akhiezer. *Theoretical and experimental studies of small molecule ligands*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, U.S.A., 2002.
- [90] A. E. Eriksson, W. A. Baase, J. A. Wozniak, and B. W. Matthews. A cavity-containing mutant of T4 lysozyme is stabilized by buried benzene. *Nature*, 355:371–373, 1992.
- [91] A. Morton, W. A. Baase, and B. W. Matthews. Energetic origins of specificity of ligand-binding in an interior nonpolar cavity of T4 lysozyme. *Biochemistry*, 34:8564–8575, 1995.
- [92] B. Q. Q. Wei, W. A. Baase, L. H. Weaver, B. W. Matthews, and B. K. Shoichet. A model binding site for testing scoring functions in molecular docking. *J. Mol. Biol.*, 322:339–355, 2002.
- [93] A. I. Su, D. M. Lorber, G. S. Weston, W. A. Baase, B. W. Matthews, and B. K. Shoichet. Docking molecules by families to increase the diversity of hits in database screens: Computational strategy and experimental evaluation. *Proteins*, 42:279–293, 2001.

- [94] A. P. Graves, R. Brenk, and B. K. Shoichet. Decoys for docking. *J. Med. Chem.*, 48:3714–3728, 2005.
- [95] G. W. Bemis and M. A. Murcko. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.*, 39:2887–2893, 1996.
- [96] G. W. Bemis and M. A. Murcko. Properties of known drugs. 2. Side chains. *J. Med. Chem.*, 42:5095–5099, 1999.
- [97] The Comprehensive Medicinal Chemistry database is available from MDL Information Systems, Inc. San Ramon, CA.
- [98] A. Husain, C. C. Galopin, S. Zhang, G. Pohnert, and B. Ganem. S(-)-dinitrobiphenic acid: A selective inhibitor of Escherichia coli chorismate mutase based on prephenate mimicry. *J. Am. Chem. Soc.*, 121:2647–2648, 1999.
- [99] L. Zhang and T. M. Rana. Solid-phase synthesis of alpha(2-(Benzylthio)1,4-dihydro-6-methyl-4-p-tolylpyrimidine-5-carboxamido) acids: A new strategy to create diversity in heterocyclic scaffolds. *J. Comb. Chem.*, 6:457–459, 2004.
- [100] D. B. Gordon and S. L. Mayo. Branch-and-Terminate: a combinatorial optimization algorithm for protein design. *Structure*, 7:1089–1098, 1999.
- [101] E. Kangas and B. Tidor. Electrostatic complementarity at ligand binding sites: Application to chorismate mutase. *J. Phys. Chem. B*, 105:880–888, 2001.
- [102] D. F. Green and B. Tidor. Escherichia coli glutaminyl-tRNA synthetase is electrostatically optimization for binding of its cognate substrates. *J. Mol. Biol.*, 342:435–452, 2004.
- [103] A. Y. Lee, P. A. Karplus, B. Ganem, and J. Clardy. Atomic structure of the buried catalytic pocket of Escherichia coli chorismate mutase. *J. Am. Chem. Soc.*, 117:3627–3628, 1995.
- [104] R. J. Iff, L. F. Ball, P. Lowe, and E. Haslam. Shikimate pathway. 5. chorismic acid and chorismate mutase. *J. Chem. Soc. Perkin T. 1.*, pages 1776–1783, 1976.
- [105] G. H. Braus. Aromatic amino-acid biosynthesis in the yeast *Saccharomyces cerevisiae*: A model system for the regulation of a eukaryotic biosynthetic pathway. *Microbiol. Rev.*, 55:349–370, 1991.
- [106] C. Poulsen and R. Verpoorte. Roles of chorismate mutase, isochorismate synthase and anthranilate synthase in plants. *Phytochemistry*, 30:377–386, 1991.
- [107] P. A. Bartlett and C. R. Johnson. An inhibitor of chorismate mutase resembling the transition-state conformation. *J. Am. Chem. Soc.*, 107:7792–7793, 1985.

- [108] P. A. Bartlett, Y. Nakagawa, C. R. Johnson, S. H. Reich, and A. Luis. Chorismate mutase inhibitors: Synthesis and evaluation of some potential transition-state analogs. *J. Org. Chem.*, 53:3195–3210, 1988.
- [109] T. Clarke, J. D. Stewart, and B. Ganem. Transition-state analog inhibitors of chorismate mutase. *Tetrahedron*, 46:731–748, 1990.
- [110] H. B. Wood, H. P. Buser, and B. Ganem. Phosphonate analogs of chorismic acid: Synthesis and evaluation as mechanism-based inactivators of chorismate mutase. *J. Org. Chem.*, 57:178–184, 1992.
- [111] M. E. Hediger. Design, synthesis, and evaluation of aza inhibitors of chorismate mutase. *Bioorg. Med. Chem.*, 12:4995–5010, 2004.
- [112] P. Wipf and A. Cunningham. A solid-phase protocol of the Biginelli dihydropyrimidine synthesis suitable for combinatorial chemistry. *Tetrahedron Lett.*, 36:7819–7822, 1995.
- [113] C. O. Kappe. 100 years of the Biginelli dihydropyrimidine synthesis. *Tetrahedron*, 49:6937–6963, 1993.
- [114] D. L. N. G. Surleraux, A. Tahri, W. G. Verschuere, G. M. E. Pille, H. A. de Kock, T. H. M. 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. T. P. Wigerinck. Discovery and selection of TMC114, a next generation HIV-1 protease inhibitor. *J. Med. Chem.*, 48:1813–1822, 2005.
- [115] F. A. Momany and R. Rone. Validation of the general-purpose QUANTA 3.2/CHARMm force-field. *J. Comp. Chem.*, 13:888–900, 1992.
- [116] A. T. Brünger and M. Karplus. Polar hydrogen positions in proteins: Empirical energy placement and neutron-diffraction comparison. *Proteins*, 4:148–156, 1988.
- [117] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, 4:187–217, 1983.
- [118] R. Dennington II, T. Keith, J. Millam, K. Eppinnett, W. L. Hovell, and R. Gilliland. *GaussView, Version 3.09*. Semichem, Inc., Shawnee Mission, KS, 2003.
- [119] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, Ö. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, P. Salvador, J. J. Dannenberg, D. K. Malick, A. D. Rabuck,

- K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komáromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle, and J. A. Pople. *Gaussian 98* (Gaussian, Inc., Pittsburgh, PA, 1998).
- [120] C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model. *J. Phys. Chem.*, 97:10269–10280, 1993.
- [121] K. A. Sharp and B. Honig. Calculating total electrostatic energies with the nonlinear Poisson–Boltzmann equation. *J. Phys. Chem.*, 94:7684–7692, 1990.
- [122] W. Rocchia, S. Sridharan, A. Nicholls, E. Alexov, A. Chiabrera, and B. Honig. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects. *J. Comp. Chem.*, 23:128–137, 2002.
- [123] The GNU triangulated surface library is available at <http://gts.sourceforge.net>.
- [124] M. M. Gottesman and I. Pastan. Biochemistry of multidrug resistance mediated by the multidrug transporter. *Ann. Rev. Biochem.*, 62:385–427, 1993.
- [125] H. Nikaido. Prevention of drug access to bacterial targets: Permeability barriers and active efflux. *Science*, 264:382–388, 1994.
- [126] C. Tantillo, J. P. Ding, A. Jacobomolina, R. G. Nanni, P. L. Boyer, S. H. Hughes, R. Pauwels, K. Andries, P. A. J. Janssen, and E. Arnold. Locations of anti-AIDS drug-binding sites and resistance mutations in the 3-dimensional structure of HIV-1 reverse-transcriptase. Implications for mechanisms of drug-inhibition and resistance. *J. Mol. Biol.*, 243:369–387, 1994.
- [127] A. Hochhaus, S. Kreil, A. S. Corbin, P. La Rosee, M. C. Muller, T. Lahaye, B. Hanfstein, C. Schoch, N. Cross, U. Berger, H. Gschaidmeier, B. J. Druker, and R. Hehlmann. Molecular and chromosomal mechanisms of resistance to imatinib (ST1571) therapy. *Leukemia*, 16:2190–2196, 2002.
- [128] L. V. Gubareva, R. Bethell, G. J. Hart, K. G. Murti, C. R. Penn, and R. G. Webster. Characterization of mutants of influenza A virus selected with the neuraminidase inhibitor 4-guanidino-Neu5Ac2en. *J. Virol.*, 70:1818–1827, 1996.
- [129] M. I. Allen, M. Deslauriers, C. W. Andrews, G. A. Tipples, K. A. Walters, D. L. J. Tyrrell, N. Brown, and L. D. Condreay. Identification and characterization of mutations in hepatitis B virus resistant to lamivudine. *Hepatology*, 27:1670–1677, 1998.

- [130] E. E. Kim, C. T. Baker, M. D. Dwyer, M. A. Murcko, B. G. Rao, R. D. Tung, and M. A. Navia. Crystal-structure of HIV-1 protease in complex with VX-478, a potent and orally bioavailable inhibitor of the enzyme. *J. Am. Chem. Soc.*, 117:1181–1182, 1995.
- [131] S. R. Turner, J. W. Strohbach, R. A. Tommasi, P. A. Aristoff, P. D. Johnson, H. I. Skulnick, L. A. Dolak, E. P. Seest, P. K. Tomich, M. J. Bohanan, M. M. Horng, J. C. Lynn, K. T. Chong, R. R. Hinshaw, K. D. Watenpaugh, M. N. Janakiraman, and S. Thaisrivongs. Tipranavir (PNU-140690): A potent, orally bioavailable nonpeptidic HIV protease inhibitor of the 5,6-dihydro-4-hydroxy-2-pyrone sulfonamide class. *J. Med. Chem.*, 41:3467–3476, 1998.
- [132] N. P. Shah, C. Tran, F. Y. Lee, P. Chen, D. Norris, and C. L. Sawyers. Overriding imatinib resistance with a novel ABL kinase inhibitor. *Science*, 305:399–401, 2004.
- [133] B. A. Larder, K. Hertogs, S. Bloor, C. van den Eynde, W. DeCian, Y. Y. Wang, W. W. Freimuth, and G. Tarpley. Tipranavir inhibits broadly protease inhibitor-resistant HIV-1 clinical samples. *AIDS*, 14:1943–1948, 2000.
- [134] J. A. Partaledis, K. Yamaguchi, M. Tisdale, E. E. Blair, C. Falcione, B. Maschera, R. E. Myers, S. Pazhanisamy, O. Futer, A. B. Cullinan, C. M. Stuver, R. A. Byrn, and D. J. Livingston. In-vitro selection and characterization of human-immunodeficiency-virus type-1 (HIV-1) isolates with reduced sensitivity to hydroxyethylamino sulfonamide inhibitors of HIV-1 aspartyl protease. *J. Virol.*, 69:5228–5235, 1995.
- [135] M. Maguire, D. Shortino, A. Klein, W. Harris, V. Manohitharajah, M. Tisdale, R. Elston, J. Yeo, S. Randall, F. Xu, H. Parker, J. May, and W. Snowden. Emergence of resistance to protease inhibitor amprenavir in human immunodeficiency virus type 1-infected patients: Selection of four alternative viral protease genotypes and influence of viral susceptibility to coadministered reverse transcriptase nucleoside inhibitors. *Antimicrob. Agents Chemother.*, 46:731–738, 2002.
- [136] N. M. King, M. Prabu-Jeyabalan, E. A. Nalivaika, P. Wigerinck, M. P. de Bethune, and C. A. Schiffer. Structural and thermodynamic basis for the binding of TMC114, a next-generation human immunodeficiency virus type 1 protease inhibitor. *J. Virol.*, 78:12012–12021, 2004.
- [137] N. M. King, M. Prabu-Jeyabalan, E. A. Nalivaika, and C. A. Schiffer. Combating susceptibility to drug resistance: Lessons from HIV-1 protease. *Chem. Biol.*, 11:1333–1338, 2004.
- [138] S. Tuske, S. G. Sarafianos, A. D. Clark, J. P. Ding, L. K. Naeger, K. L. White, M. D. Miller, C. S. Gibbs, P. L. Boyer, P. Clark, G. Wang, B. L. Gaffney, R. A.

- Jones, D. M. Jerina, S. H. Hughes, and E. Arnold. Structures of HIV-1 RT-DNA complexes before and after incorporation of the anti-AIDS drug tenofovir. *Nat. Struct. Mol. Biol.*, 11:469–474, 2004.
- [139] P. Kirkpatrick. Antiviral drugs: Inside the envelope. *Nat. Rev. Drug Discov.*, 3:100–100, 2004.
- [140] J. M. Coffin. HIV population-dynamics in-vivo: Implications for genetic-variation, pathogenesis, and therapy. *Science*, 267:483–489, 1995.
- [141] T. D. Wu, C. A. Schiffer, M. J. Gonzales, J. Taylor, R. Kantor, S. W. Chou, D. Israelski, A. R. Zolopa, W. J. Fessel, and R. W. Shafer. Mutation patterns and structural correlates in human immunodeficiency virus type 1 protease following different protease inhibitor treatments. *J. Virol.*, 77:4836–4847, 2003.
- [142] N. G. Hoffman, C. A. Schiffer, and R. Swanstrom. Covariation of amino acid positions in HIV-1 protease. *Virology*, 314:536–548, 2003.
- [143] B. Schmidt, K. Korn, B. Moschik, C. Paatz, K. Uberla, and H. Walter. Low level of cross-resistance to amprenavir (141W94) in samples from patients pretreated with other protease inhibitors. *Antimicrob. Agents Chemother.*, 44:3213–3216, 2000.
- [144] M. Vaillancourt, D. Irlbeck, T. Smith, R. W. Coombs, and R. Swanstrom. The HIV type 1 protease inhibitor saquinavir can select for multiple mutations that confer increasing resistance. *AIDS Res. Hum. Retroviruses*, 15:355–363, 1999.
- [145] T. Watkins, W. Resch, D. Irlbeck, and R. Swanstrom. Selection of high-level resistance to human immunodeficiency virus type 1 protease inhibitors. *Antimicrob. Agents Chemother.*, 47:759–769, 2003.
- [146] L. J. Hyland, T. A. Tomaszek, and T. D. Meek. Human immunodeficiency virus-1 protease. 2. Use of pH rate studies and solvent kinetic isotope effects to elucidate details of chemical mechanism. *Biochemistry*, 30:8454–8463, 1991.
- [147] T. Yamazaki, L. K. Nicholson, D. A. Torchia, P. Wingfield, S. J. Stahl, J. D. Kaufman, C. J. Eyermann, C. N. Hodge, P. Y. S. Lam, Y. Ru, P. K. Jadhav, C. H. Chang, and P. C. Weber. NMR and X-ray evidence that the HIV protease catalytic aspartyl groups are protonated in the complex formed by the protease and a nonpeptide cyclic urea-based inhibitor. *J. Am. Chem. Soc.*, 116:10791–10792, 1994.
- [148] J. Trylska, J. Antosiewicz, M. Geller, C. N. Hodge, R. M. Klabe, M. S. Head, and M. K. Gilson. Thermodynamic linkage between the binding of protons and inhibitors to HIV-1 protease. *Protein Sci.*, 8:180–195, 1999.
- [149] J. J. Irwin and B. K. Shoichet. ZINC — A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Mod.*, 45:177–182, 2005.

- [150] L. David, R. Luo, and M. K. Gilson. Ligand–receptor docking with the Mining Minima optimizer. *J. Comput. Aid. Mol. Des.*, 15:157–171, 2001.
- [151] V. Kairys and M. K. Gilson. Enhanced docking with the mining minima optimizer: Acceleration and side-chain flexibility. *J. Comp. Chem.*, 23:1656–1670, 2002.
- [152] T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard, and J. L. Banks. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.*, 47:1750–1759, 2004.
- [153] H. L. Sham, C. Zhao, K. D. Stewart, D. A. Betebenner, S. Q. Lin, C. H. Park, X. P. Kong, W. Rosenbrook, T. Herrin, D. Madigan, S. Vasavanonda, N. Lyons, A. Molla, A. Saldivar, K. C. Marsh, E. McDonald, N. E. Wideburg, J. F. Denissen, T. Robins, D. J. Kempf, J. J. Plattner, and D. W. Norbeck. A novel, picomolar inhibitor of human immunodeficiency virus type 1 proteases. *J. Med. Chem.*, 39:392–397, 1996.
- [154] S. D. Young, L. S. Payne, W. J. Thompson, N. Gaffin, T. A. Lyle, S. F. Britcher, S. L. Graham, T. H. Schultz, A. A. Deana, P. L. Darke, J. Zugay, W. A. Schleif, J. C. Quintero, E. A. Emini, P. S. Anderson, and J. R. Huff. HIV-1 protease inhibitors based on hydroxyethylene dipeptide isosteres: An investigation into the role of the P1' side chain on structure activity. *J. Med. Chem.*, 35:1702–1709, 1992.
- [155] B. Kuhn, P. Gerber, T. Schulz-Gasch, and M. Stahl. Validation and use of the MM-PBSA approach for drug discovery. *J. Med. Chem.*, 48:4040–4048, 2005.
- [156] C. W. Murray, C. A. Baxter, and A. D. Frenkel. The sensitivity of the results of molecular docking to induced fit effects: Application to thrombin, thermolysin and neuraminidase. *J. Comput. Aid. Mol. Des.*, 13:547–562, 1999.
- [157] A. C. R. Martin. <http://www.bioinf.org.uk/software/profit/>.
- [158] C. Debouck. The HIV-1 protease as a therapeutic target for AIDS. *AIDS Res. Hum. Retroviruses*, 8:153–164, 1992.
- [159] A. Molla, M. Korneyeva, Q. Gao, S. Vasavanonda, P. J. Schipper, H. M. Mo, M. Markowitz, T. Chernyavskiy, P. Niu, N. Lyons, A. Hsu, G. R. Granneman, D. D. Ho, C. A. B. Boucher, J. M. Leonard, D. W. Norbeck, and D. J. Kempf. Ordered accumulation of mutations in HIV protease confers resistance to ritonavir. *Nat. Med.*, 2:760–766, 1996.
- [160] C. Flexner. HIV protease inhibitors. *New Eng. J. Med.*, 338:1281–1292, 1998.
- [161] G. Croteau, L. Doyon, D. Thibeault, G. McKercher, L. Pilote, and D. Lamarre. Impaired fitness of human immunodeficiency virus type 1 variants with high-level resistance to protease inhibitors. *J. Virol.*, 71:1089–1096, 1997.

- [162] J. Martinez-Picado, L. V. Savara, L. Sutton, and R. T. D'Aquila. Replicative fitness of protease inhibitor-resistant mutants of human immunodeficiency virus type 1. *J. Virol.*, 73:3744–3752, 1999.
- [163] I. T. Weber, J. Wu, J. Adomat, R. W. Harrison, A. R. Kimmel, E. M. Wondrak, and J. M. Louis. Crystallographic analysis of human immunodeficiency virus 1 protease with an analog of the conserved CA-p2 substrate — Interactions with frequently occurring glutamic acid residue at P2' position of substrates. *Eur. J. Biochem.*, 249:523–530, 1997.
- [164] B. Mahalingam, J. M. Louis, J. Hung, R. W. Harrison, and I. T. Weber. Structural implications of drug-resistant mutants of HIV-1 protease: High-resolution crystal structures of the mutant protease/substrate analogue complexes. *Proteins*, 43:455–464, 2001.
- [165] Y. F. Tie, P. I. Boross, Y. F. Wang, L. Gaddis, F. L. Liu, X. F. Chen, J. Tozser, R. W. Harrison, and I. T. Weber. Molecular basis for substrate recognition and drug resistance from 1.1 to 1.6 Angstrom resolution crystal structures of HIV-1 protease mutants with substrate analogs. *FEBS J.*, 272:5265–5277, 2005.
- [166] M. Prabu-Jeyabalan, E. A. Nalivaika, N. M. King, and C. A. Schiffer. Viability of a drug-resistant human immunodeficiency virus type 1 protease variant: Structural insights for better antiviral therapy. *J. Virol.*, 77:1306–1315, 2003.
- [167] M. Prabu-Jeyabalan, E. A. Nalivaika, N. M. King, and C. A. Schiffer. Structural basis for coevolution of a human immunodeficiency virus type 1 nucleocapsid-p1 cleavage site with a V82A drug-resistant mutation in viral protease. *J. Virol.*, 78:12446–12454, 2004.
- [168] J. S. Bardi, I. Luque, and E. Freire. Structure-based thermodynamic analysis of HIV-1 protease inhibitors. *Biochemistry*, 36:6588–6596, 1997.
- [169] L. P. Lee and B. Tidor. Barstar is electrostatically optimized for tight binding to barnase. *Nat. Struct. Biol.*, 8:73–76, 2001.
- [170] T. Sulea and E. O. Purisima. Optimizing ligand charges for maximum binding affinity. A solvated interaction energy approach. *J. Phys. Chem. B*, 105:889–899, 2001.
- [171] E. Kangas and B. Tidor. Electrostatic specificity in molecular ligand design. *J. Chem. Phys.*, 112:9120–9131, 2000.
- [172] P. A. Sims, C. F. Wong, and J. A. McCammon. Charge optimization of the interface between protein kinases and their ligands. *J. Comp. Chem.*, 25:1416–1429, 2004.
- [173] D. F. Green and B. Tidor. Design of improved protein inhibitors of HIV-1 cell entry: Optimization of electrostatic interactions at the binding interface. *Proteins*, 60:644–657, 2005.

- [174] M. K. Gilson. Sensitivity analysis and charge-optimization for flexible ligands: Applicability to lead optimization. *J. Chem. Theory Comput.*, 2:259–270, 2006.
- [175] A. D. MacKerell, 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 III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102:3586–3616, 1998.
- [176] E. Neria, S. Fischer, and M. Karplus. Simulation of activation free energies in molecular systems. *J. Chem. Phys.*, 105:1902–1921, 1996.
- [177] W. E. Reiher III. *Theoretical studies of hydrogen bonding*. PhD thesis, Harvard University, Cambridge, MA, U.S.A., 1985.
- [178] R. J. Vanderbei. LOQO: An interior point code for quadratic programming. *Optim. Method. Softw.*, 11-2:451–484, 1999.
- [179] R. J. Vanderbei. LOQO User’s Manual - Version 3.10. *Optim. Method. Softw.*, 11-2:485–514, 1999.
- [180] R. L. Dunbrack and M. Karplus. Conformational-analysis of the backbone-dependent rotamer preferences of protein side-chains. *Nat. Struct. Biol.*, 1:334–340, 1994.
- [181] J. M. Scholtz and R. L. Baldwin. The mechanism of alpha-helix formation by peptides. *Ann. Rev. Biophys. Biomol. Struct.*, 21:95–118, 1992.
- [182] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. A new force-field for molecular mechanical simulation of nucleic-acids and proteins. *J. Am. Chem. Soc.*, 106:765–784, 1984.
- [183] J. O. Hui, A. G. Tomasselli, I. M. Reardon, J. M. Lull, D. P. Brunner, C. S. C. Tomich, and R. L. Heinrikson. Large-scale purification and refolding of HIV-1 protease from Escherichia coli inclusion bodies. *J. Protein Chem.*, 12:323–327, 1993.
- [184] W. Minor. XDISPLAYF program. Purdue University: West Lafayette, Indiana, 1993.
- [185] Z. Otwinowski. Oscillation data reduction program. In *CCP4 Study Weekend: Data collection and processing, 29–30 Jan 1993*, England, 1993. SERC Daresbury Laboratory.
- [186] S. Bailey. The CCP4 suite: Programs for protein crystallography. *Acta Crystallogr. D*, 50:760–763, 1994.

- [187] J. Navaza. AMoRe: An automated package for molecular replacement. *Acta Crystallogr. A*, 50:157–163, 1994.
- [188] R. J. Morris, A. Perrakis, and V. S. Lamzin. ARP/wARP's model-building algorithms. I. The main chain. *Acta Crystallogr. D*, 58:968–975, 2002.
- [189] T. A. Jones, M. Bergdoll, and M. Kjeldgaard. O: A macromolecular modeling environment. In C. Bugg and S. Ealick, editors, *Crystallographic and Modeling Methods in Molecular Design*, pages 189–195. Springer-Verlag, Berlin, 1990.
- [190] G. N. Murshudov, A. A. Vagin, and E. J. Dodson. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D*, 53:240–255, 1997.
- [191] J. Kuriyan and W. I. Weis. Rigid protein motion as a model for crystallographic temperature factors. *Proc. Natl. Acad. Sci. U.S.A.*, 88:2773–2777, 1991.
- [192] V. Schomaker and K. N. Trueblood. On the rigid-body motion of molecules in crystals. *Acta Crystallogr. B*, 24:63–76, 1968.
- [193] I. J. Tickle and D. S. Moss. Modelling rigid-body thermal motions in macromolecular crystal structure refinement. In *IUCr99 Computing School*. IUCr, London, 1999.
- [194] R. A. Laskowski, M. W. Macarthur, D. S. Moss, and J. M. Thornton. PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, 26:283–291, 1993.
- [195] J. Voldman, M. L. Gray, and M. A. Schmidt. Microfabrication in biology and medicine. *Ann. Rev. Biomed. Eng.*, 1:401–425, 1999.
- [196] D. S. Gray, T. L. Tan, J. Voldman, and C. S. Chen. Dielectrophoretic registration of living cells to a microelectrode array. *Biosens. Bioelectron.*, 19:771–80, 2004.
- [197] G.-B. Lee, S.-H. Chen, G.-R. Huang, W.-C. Sung, and Y.-H. Lin. Microfabricated plastic chips by hot embossing methods and their applications for DNA separation and detection. *Sens. Actuators, B*, 75:142–148, 2001.
- [198] T. P. Burg and S. R. Manalis. *Appl. Phys. Lett.*, 83:2698–2700, 2003.
- [199] T. Korsmeyer. Design tools for BioMEMS. In *Design Automation Conference*, pages 622–627, 2004.
- [200] J. White. CAD challenges in BioMEMS design. In *IEEE Design Automation Conference (DAC)*, pages 629–632, 2004.
- [201] S. D. Senturia, R. M. Harris, B. P. Johnson, S. Kim, K. Nabors, M. A. Shulman, and J. K. White. A computer-aided design system for microelectromechanical systems (MEMCAD). *J. Microelectromech. S.*, 1:3–13, 1992.

- [202] C. A. Savran, S. M. Knudsen, A. D. Ellington, and S. R. Manalis. Micromechanical detection of proteins using aptamer-based receptor molecules. *J. Anal. Chem.*, 76:3194–3198, 2004.
- [203] R. A. Potyrailo, R. C. Conrad, A. D. Ellington, and G. M. Hieftje. Adapting selected nucleic acid ligands (aptamers) to biosensors. *J. Anal. Chem.*, 70:3419–3425, 1998.
- [204] I. Klapper, R. Hagstrom, R. Fine, K. Sharp, and B. Honig. Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: Effects of ionic strength and amino-acid modification. *Proteins*, 1:47–59, 1986.
- [205] A. Jean-Charles, A. Nicholls, K. Sharp, B. Honig, A. Tempczyk, T. F. Hendrickson, and W. C. Still. Electrostatic contributions to solvation energies: Comparison of free energy perturbation and continuum calculations. *J. Am. Chem. Soc.*, 113:1454–1455, 1991.
- [206] M. Holst, N. Baker, and F. Wang. Adaptive multilevel finite element solution of the Poisson–Boltzmann equation. I. Algorithms and examples. *J. Comp. Chem.*, 21:1319–1342, 2000.
- [207] B. J. Yoon and A. M. Lenhoff. A boundary element method for molecular electrostatics with electrolyte effects. *J. Comp. Chem.*, 11:1080–1086, 1990.
- [208] K. Nabors, F. T. Korsmeyer, F. T. Leighton, and J. K. White. Preconditioned, adaptive, multipole-accelerated iterative methods for three-dimensional first-kind integral equations of potential theory. *SIAM J. Sci. Comput.*, 15:713–735, 1994.
- [209] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *J. Comp. Phys.*, 73:325–348, 1987.
- [210] W. Hackbusch. A sparse matrix arithmetic based on H-matrices. I. Introduction to H-matrices. *Computing*, 62:89–108, 1999.
- [211] W. Hackbusch and B. N. Khoromskij. A sparse H-matrix arithmetic. II. Application to multi-dimensional problems. *Computing*, 64:21–47, 2000.
- [212] S. Borm, L. Grasedyck, and W. Hackbusch. Introduction to hierarchical matrices with applications. *Eng. Anal. Bound. Elem.*, 27:405–22, 2003.
- [213] W. Shi, J. Liu, N. Kakani, and T. Yu. A fast hierarchical algorithm for 3-D capacitance extraction. In *Design Automation Conference*, 1998.
- [214] J. Tausch and J. K. White. A multiscale method for fast capacitance extraction. In *Design Automation Conference*, pages 537–542, 1999.
- [215] E. T. Ong, K. M. Lim, K. H. Lee, and H. P. Lee. A fast algorithm for three-dimensional potential fields calculation: fast fourier transform on multipoles. *J. Comp. Phys.*, 192:244–61, 2003.

- [216] E. T. Ong, H. P. Lee, and K. M. Lim. A parallel fast Fourier transform on multipoles (FFTM) algorithm for electrostatics analysis of three-dimensional structures. *IEEE T. Comput. Aid. D.*, 23:1063–1072, 2004.
- [217] G. Biros, L. Ying, and D. Zorin. A fast solver for the Stokes equations with distributed forces in complex geometries. *J. Comp. Phys.*, 193:317–348, 2004.
- [218] L. Ying, G. Biros, and D. Zorin. A kernel-independent adaptive fast multipole algorithm in two and three dimensions. *J. Comp. Phys.*, 196:591–626, 2004.
- [219] S. Kapur and D. E. Long. IES³: Efficient electrostatic and electromagnetic simulation. *IEEE Comput. Sci. Eng.*, 5:60–7, 1998.
- [220] L. Greengard, J. Huang, V. Rokhlin, and S. Wandzura. Accelerating fast multipole methods for the Helmholtz equation at low frequencies. *IEEE Comput. Sci. Eng.*, 5:32–38, 1998.
- [221] D. Gope and V. Jandhyala. PILOT: A fast algorithm for enhanced 3D parasitic extraction efficiency. In *IEEE Electrical Performance of Electronic Packaging*, 2003.
- [222] L. Greengard and V. Rokhlin. A new version of the fast multipole method for the Laplace equation in three dimensions. *Acta Num.*, pages 229–269, 1997.
- [223] Z. Zhu, B. Song, and J. White. Algorithms in FastImp: A fast and wideband impedance extraction program for complicated 3D geometries. *IEEE/ACM Design Automation Conference*, 2003.
- [224] N. R. Aluru and J. White. A fast integral equation technique for analysis of microflow sensors based on drag force calculations. In *Modeling and Simulation of Microsystems*, 1998.
- [225] W. Ye, X. Wang, and J. White. A fast Stokes solver for generalized flow problems. In *Modeling and Simulation of Microsystems*, 2000.
- [226] X. Wang. *FastStokes: A fast 3-D fluid simulation program for micro-electromechanical systems*. PhD thesis, Massachusetts Institute of Technique, 2002.
- [227] S. S. Kuo, M. D. Altman, J. P. Bardhan, B. Tidor, and J. K. White. Fast methods for simulation of biomolecule electrostatics. *International Conference on Computer Aided Design (ICCAD)*, 2002.
- [228] Y. Saad and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comp.*, 7:856–869, 1986.
- [229] F. M. Richards. Areas, volumes, packing, and protein structure. *Ann. Rev. Biophys. Bioeng.*, 6:151–176, 1977.

- [230] J. Kanapka, J. Phillips, and J. White. Fast methods for extraction and sparsification of substrate coupling. In *Design Automation Conference*, pages 738–743, 2000.
- [231] J. Fliege and U. Maier. The distribution of points on the sphere and corresponding cubature formulae. *IMA J. Num. Anal.*, 19:317–334, 1999.
- [232] M. Frigo and S. G. Johnson. FFTW: An adaptive software architecture for the FFT. In *Proceedings of the 1998 IEEE International Conference on Acoustics Speech and Signal Processing*, volume 3, pages 1381–1384. IEEE, 1998.
- [233] B. P. Mosier, J. I. Malho, and J. G. Santiago. Photobleached-fluorescence imaging of microflows. *Exp. Fluid.*, 33:545–554, 2002.
- [234] S. I. Cho, S.-H. Lee, D. S. Chung, and Y.-K. Kim. Bias-free pneumatic sample injection in microchip electrophoresis. *J. Chromat. A*, 1063:253–256, 2005.
- [235] Z. Zhu. Efficient techniques for wideband impedance extraction of complex 3-dimensional geometries. Master’s thesis, Massachusetts Institute of Technology, 2002.
- [236] A. E. Ruehli and P. A. Brennan. Efficient capacitance calculations for three-dimensional multiconductor systems. *IEEE Transactions on Microwave Theory and Techniques*, 21:76–82, 1973.
- [237] E. T. Ong, K. H. Lee, and K. M. Lim. Singular elements for electro-mechanical coupling analysis of micro-devices. *J. Micromech. Microeng.*, 13:482–490, 2003.
- [238] E. T. Ong and K. M. Lim. Three-dimensional singular boundary elements for corner and edge singularities in potential problems. *Engineering Analysis with Boundary Elements*, 29:175–189, 2005.
- [239] B. Roux and T. Simonson. Implicit solvent models. *Biophys. Chem.*, 78:1–20, 1999.
- [240] T. Simonson. Macromolecular electrostatics: Continuum models and their growing pains. *Curr. Opin. Struct. Biol.*, 11:243–252, 2001.
- [241] N. A. Baker. Poisson–Boltzmann methods for biomolecular electrostatics. *Meth. Enzymol.*, 383:94–118, 2004.
- [242] I. Stakgold. *Green’s Functions and Boundary Value Problems*. John Wiley & Sons, 2nd edition, 1998.
- [243] M. Holst and F. Saied. Multigrid solution of the Poisson–Boltzmann equation. *J. Comp. Chem.*, 14:105–113, 1993.

- [244] J. D. Madura, J. M. Briggs, R. C. Wade, M. E. Davis, B. A. Luty, A. Ilin, J. Antosiewicz, M. K. Gilson, B. Bagheri, L. R. Scott, and J. A. McCammon. Electrostatics and diffusion of molecules in solution: Simulations with the University of Houston Brownian dynamics program. *Comput. Phys. Comm.*, 91:57–95, 1995.
- [245] W. Rocchia, E. Alexov, and B. Honig. Extending the applicability of the nonlinear Poisson–Boltzmann equation: Multiple dielectric constants and multivalent ions. *J. Phys. Chem. B*, 105:6507–6514, 2001.
- [246] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.*, 98:10037–10041, 2001.
- [247] T. J. You and S. C. Harvey. Finite-element approach to the electrostatics of macromolecules with arbitrary geometries. *J. Comp. Chem.*, 14:484–501, 1993.
- [248] C. M. Cortis and R. A. Friesner. Numerical solution of the Poisson–Boltzmann equation using tetrahedral finite-element meshes. *J. Comp. Chem.*, 18:1591–1608, 1997.
- [249] N. Baker, M. Holst, and F. Wang. Adaptive multilevel finite element solution of the Poisson–Boltzmann equation II. Refinement at solvent-accessible surfaces in biomolecular systems. *J. Comp. Chem.*, 21:1343–1352, 2000.
- [250] R. J. Zauhar and R. S. Morgan. A new method for computing the macromolecular electric-potential. *J. Mol. Biol.*, 186:815–820, 1985.
- [251] R. J. Zauhar and R. S. Morgan. The rigorous computation of the molecular electric-potential. *J. Comp. Chem.*, 9:171–187, 1988.
- [252] R. J. Zauhar and R. S. Morgan. Computing the electric potential of biomolecules: Application of a new method of molecular surface triangulation. *J. Comp. Chem.*, 11:603–622, 1990.
- [253] A. H. Juffer, E. F. F. Botta, B. A. M. Vankeulen, A. Vanderploeg, and H. J. C. Berendsen. The electric potential of a macromolecule in a solvent: A fundamental approach. *J. Comp. Phys.*, 97:144–171, 1991.
- [254] H. X. Zhou. Boundary-element solution of macromolecular electrostatics: Interaction energy between 2 proteins. *Biophys. J.*, 65:955–963, 1993.
- [255] R. Bharadwaj, A. Windemuth, S. Sridharan, B. Honig, and A. Nicholls. The fast multipole boundary-element method for molecular electrostatics: An optimal approach for large systems. *J. Comp. Chem.*, 16:898–913, 1995.
- [256] E. O. Purisima and S. H. Nilar. A simple yet accurate boundary-element method for continuum dielectric calculations. *J. Comp. Chem.*, 16:681–689, 1995.

- [257] R. J. Zauhar and A. Varnek. A fast and space-efficient boundary element method for computing electrostatic and hydration effects in large molecules. *J. Comp. Chem.*, 17:864–877, 1996.
- [258] Y. N. Vorobjev and H. A. Scheraga. A fast adaptive multigrid boundary element method for macromolecular electrostatic computations in a solvent. *J. Comp. Chem.*, 18:569–583, 1997.
- [259] A. H. Boschitsch, M. O. Fenley, and H. X. Zhou. Fast boundary element method for the linear Poisson–Boltzmann equation. *J. Phys. Chem. B*, 106:2741–2754, 2002.
- [260] A. J. Bordner and G. A. Huber. Boundary element solution of the linear Poisson–Boltzmann equation and a multipole method for the rapid calculation of forces on macromolecules in solution. *J. Comp. Chem.*, 24:353–367, 2003.
- [261] D. M. Chipman. Solution of the linearized Poisson–Boltzmann equation. *J. Chem. Phys.*, 120:5566–5575, 2004.
- [262] K. E. Atkinson. *The Numerical Solution of Integral Equations of the Second Kind*. Cambridge University Press, 1997.
- [263] S. Hofinger and T. Simonson. Dielectric relaxation in proteins: A continuum electrostatics model incorporating dielectric heterogeneity of the protein and time-dependent charges. *J. Comp. Chem.*, 22:290–305, 2001.
- [264] B. Z. Lu, D. Q. Zhang, and J. A. McCammon. Computation of electrostatic forces between solvated molecules determined by the Poisson–Boltzmann equation using a boundary element method. *J. Chem. Phys.*, 122, 2005.
- [265] A. H. Boschitsch and M. O. Fenley. Hybrid boundary element and finite difference method for solving the nonlinear Poisson–Boltzmann equation. *J. Comp. Chem.*, 25:935–955, 2004.
- [266] B. Lee and F. M. Richards. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, 55:379–400, 1971.
- [267] M. L. Connolly. Analytical molecular surface calculation. *J. Appl. Cryst.*, 16:548–558, 1983.
- [268] M. L. Connolly. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221:709–713, 1983.
- [269] J. Liang and S. Subramaniam. Computation of molecular electrostatics with boundary element methods. *Biophys. J.*, 73:1830–1841, 1997.
- [270] J. P. Bardhan, M. D. Altman, B. Tidor, and J. K. White. Efficient integration techniques for curved panel discretizations of molecule-solvent interfaces. Unpublished.

- [271] A. H. Boschitsch, M. O. Fenley, and W. K. Olson. A fast adaptive multipole algorithm for calculating screened Coulomb (Yukawa) interactions. *J. Comp. Phys.*, 151:212–241, 1999.
- [272] H. Cheng, L. Greengard, and V. Rokhlin. A fast adaptive multipole algorithm in three dimensions. *J. Comp. Phys.*, 155:468–498, 1999.
- [273] L. X. Ying, G. Biros, and D. Zorin. A kernel-independent adaptive fast multipole algorithm in two and three dimensions. *J. Comp. Phys.*, 196:591–626, 2004.
- [274] M. D. Altman, J. P. Bardhan, B. Tidor, and J. K. White. FFTSVD: A fast multiscale boundary-element method solver suitable for bio-MEMS and biomolecule simulation. *IEEE T. Comput. Aid. D.*, 25:274–284, 2006.
- [275] J. L. Hess and A. M. O. Smith. Calculation of non-lifting potential flow about arbitrary three-dimensional bodies. *J. Ship Res.*, 8:22–44, 1962.
- [276] J. N. Newman. Distribution of sources and normal dipoles over a quadrilateral panel. *J. Eng. Math.*, 20:113–126, 1986.
- [277] J.-L. Guermond. Numerical quadratures for layer potentials over curved domains in R^3 . *SIAM J. Numer. Anal.*, 29:1347–1369, 1992.
- [278] A. M. Buckle, G. Schreiber, and A. R. Fersht. Protein–protein recognition: Crystal structural-analysis of a barnase barstar complex at 2.0-Angstrom resolution. *Biochemistry*, 33:8878–8889, 1994.
- [279] Z. S. Hendsch and B. Tidor. Electrostatic interactions in the GCN4 leucine zipper: Substantial contributions arise from intramolecular interactions enhanced on binding. *Protein Sci.*, 8:1381–1392, 1999.
- [280] S. Spector, M. H. Wang, S. A. Carp, J. Robblee, Z. S. Hendsch, R. Fairman, B. Tidor, and D. P. Raleigh. Rational modification of protein stability by the mutation of charged surface residues. *Biochemistry*, 39:872–879, 2000.
- [281] J. D. Jackson. *Classical Electrodynamics*. Wiley, 3rd edition, 1998.
- [282] R. Kress. *Linear Integral Equations*. Springer–Verlag, 2nd edition, 1999.
- [283] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 1983.
- [284] R. J. Zauhar. SMART: A solvent-accessible triangulated surface generator for molecular graphics and boundary-element applications. *J. Comput. Aid. Mol. Des.*, 9:149–159, 1995.
- [285] A. Stroud. *Approximate Calculation of Multiple Integrals*. Prentice Hall, 1971.

- [286] X. Wang, J. N. Newman, and J. K. White. Robust algorithms for boundary-element integrals on curved surfaces. In *Technical Proceedings of the 2000 International Conference on Modeling and Simulation of Microsystems*, pages 473–476. Nano Science and Technology Institute, 2000.
- [287] D. J. Willis, J. Peraire, and J. K. White. A quadratic basis function, quadratic geometry, high order panel method. In *44th AIAA Aerospace Sciences Meeting, AIAA-2006-1253*, 2006.
- [288] J. Schöberl. NETGEN — An advancing front 2D/3D-mesh generator based on abstract rules. *Comput. Visual. Sci.*, 1:42–52, 1997.
- [289] J. Shen and J. Wendoloski. Electrostatic binding energy calculation using the finite difference solution to the linearized Poisson–Boltzmann equation: Assessment of its accuracy. *J. Comp. Chem.*, 17:350–357, 1996.
- [290] G. Schreiber and A. R. Fersht. Interaction of barnase with its polypeptide inhibitor barstar studied by protein engineering. *Biochemistry*, 32:5145–5150, 1993.
- [291] G. Schreiber and A. R. Fersht. Energetics of protein–protein interactions: Analysis of the barnase–barstar interface by single mutations and double mutant cycles. *J. Mol. Biol.*, 248:478–486, 1995.
- [292] F. B. Sheinerman and B. Honig. On the role of electrostatic interactions in the design of protein–protein interfaces. *J. Mol. Biol.*, 318:161–177, 2002.
- [293] L. T. Chong, S. E. Dempster, Z. S. Hendsch, L. P. Lee, and B. Tidor. Computation of electrostatic complements to proteins: A case of charge stabilized binding. *Protein Sci.*, 7:206–210, 1998.
- [294] F. Dong, M. Vijayakumar, and H. X. Zhou. Comparison of calculation and experiment implicates significant electrostatic contributions to the binding stability of barnase and barstar. *Biophys. J.*, 85:49–60, 2003.
- [295] C. Frisch, G. Schreiber, C. M. Johnson, and A. R. Fersht. Thermodynamics of the interaction of barnase and barstar: Changes in free energy versus changes in enthalpy on mutation. *J. Mol. Biol.*, 267:696–706, 1997.
- [296] G. Schreiber, C. Frisch, and A. R. Fersht. The role of Glu73 of barnase in catalysis and the binding of barstar. *J. Mol. Biol.*, 270:111–122, 1997.
- [297] D. G. Covell and A. Wallqvist. Analysis of protein–protein interactions and the effects of amino acid mutations on their energetics. The importance of water molecules in the binding epitope. *J. Mol. Biol.*, 269:281–297, 1997.
- [298] T. Kortemme and D. Baker. A simple physical model for binding energy hot spots in protein–protein complexes. *Proc. Natl. Acad. Sci. U.S.A.*, 99:14116–14121, 2002.

- [299] T. Wang, S. Tomic, R. R. Gabdouliline, and R. C. Wade. How optimal are the binding energetics of barnase and barstar? *Biophys. J.*, 87:1618–1630, 2004.
- [300] I. Massova and P. A. Kollman. Computational alanine scanning to probe protein–protein interactions: A novel approach to evaluate binding free energies. *J. Am. Chem. Soc.*, 121:8133–8143, 1999.
- [301] H. Gohlke and D. A. Case. Converging free energy estimates: MM-PB(GB)SA studies on the protein–protein complex Ras–Raf. *J. Comp. Chem.*, 25:238–250, 2004.
- [302] H. Ohtaka, A. Velazquez-Campoy, D. Xie, and E. Freire. Overcoming drug resistance in HIV-1 chemotherapy: The binding thermodynamics of amprenavir and TMC-126 to wild-type and drug-resistant mutants of the HIV-1 protease. *Protein Sci.*, 11:1908–1916, 2002.