View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by DSpace@MIT

Optimization of Electrostatic Binding Free Energy: Applications to the Analysis and Design of Ligand Binding in Protein Complexes

by

David Francis Green

BACHELOR OF SCIENCE IN CHEMISTRY
SIMON FRASER UNIVERSITY, 1997

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

September 2002

© Massachusetts Institute of Technology 2002.
All rights reserved.

Author . . . .
Department of Chemistry
July 30, 2002

Certified by . . . .
Bruce Tidor
Associate Professor of Bioengineering and Computer Science
Thesis Supervisor

Accepted by . . . .
Robert W. Field
Chairman, Department Committee on Graduate Students

This thesis has been examined by a committee of the
Department of Chemistry as follows:

JoAnne Stubbe . . . .
Thesis Committee Chair

Bruce Tidor . . . .
Thesis Supervisor

Robert W. Field . . . .

Optimization of Electrostatic Binding Free Energy: Applications to the Analysis and Design of Ligand Binding in Protein Complexes

by

David Francis Green

Submitted to the Department of Chemistry on July 30, 2002,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Biological Chemistry

Abstract

Electrostatic interactions play an important role in determining the energetics of association in biomolecular complexes. Previous work has shown that, within a continuum electrostatic model, for any given complex there exists a ligand charge distribution which optimizes the electrostatic binding free energy — the electrostatic complement of the target receptor. This electrostatic affinity optimization procedure was applied to several systems both in order to understand the role of electrostatic interactions in natural systems and as a tool in the design of ligands with improved affinity. Comparison of the natural and optimal charges of several ligands of glutaminyl-tRNA synthetase from *E. coli*, an enzyme with a strong natural requirement for specificity, shows remarkable similarity in many areas, suggesting that the optimization of electrostatic interactions played a role in the evolution of this system. The optimization procedure was also applied to the design of improvements to two inhibitors of HIV-1 viral–cell membrane fusion. Two tryptophan residues that are part of a D-peptide inhibitor were identified as contributing most significantly to binding, and a novel computational screening procedure based on the optimization methodology was developed to screen a library of tryptophan derivatives at both positions. Additionally, the optimization methodology was used to predict four mutations to standard amino acids at three positions on 5-Helix, a protein inhibitor of membrane fusion. All mutations were computed to improve the affinity of the inhibitor, with a five hundred-fold improvement calculated for one triple mutant. In the complex of β-lactamase inhibitor protein with TEM1 β-lactamase, a novel type of electrostatic interaction was identified, with surface exposed charged groups on the periphery of the binding interface projecting significant energetic effects through as much as 10 Å of solvent. Finally, a large number of *ab initio* methods for determining partial atomic charges on small molecules were evaluated in terms of their ability to reproduce experimental values in continuum electrostatic calculations, with several preferred methods identified.

Thesis Supervisor: Bruce Tidor
Title: Associate Professor of Bioengineering and Computer Science

Acknowledgements

First and foremost I would like to thank my thesis advisor Bruce Tidor for all his advice and guidance. Bruce has been the best advisor I could have hoped for, allowing me to choose the direction of my research while always providing helpful suggestions and insight.

My undergraduate research advisors, B. Mario Pinto and Roland K. Pomeroy, started me out on my scientific career, providing my first training in how to be a researcher and a scientist. Mario gave me the chance to do research from my first year, and helped me discover my strong interest in biophysical chemistry. Pom never let me forget that there is much more to science than the hottest new thing, and that all applied science builds on a foundation of fundamental research. I can't overstate how much they helped me become the scientist I am today.

I would also like to thank the current and past members of the Tidor group for making it a great environment in which to work. In particular, I would like to thank Erik Kangas and Lee-Peng Lee, whose work on charge optimization lay the foundation for all the work contained herein, and who provided helpful guidance at the beginning of my graduate career. Zachary Hendsch, Justin Caravella, Karl Hanf, Shari Spector, Philip Kim, Woody Sherman, Michael Altman, and Brian Joughin also all deserve special acknowledgement for numerous helpful discussions. Brian contributed to the work on β-lactamase inhibitor protein, and much of Appendix A is his work.

Special thanks are due to Peter S. Kim, Debra M. Eckert, and Michael J. Root for helpful discussions and for providing insight into the HIV-1 gp41 system, as well as for sharing experimental data at various stages. Michael Root also deserves special recognition for experimentally validating some of the results predicted in the 5-Helix system.

Very heartfelt thanks are due to my partner Faye Yu for support throughout the years, and to my parents and brothers for support and encouragement of my love of science from the very beginning.

Contents

1 **General Introduction** **13**

1.1 Electrostatics and biomolecular association 13

1.2 Computational studies of biomolecules 16

1.3 Optimization of electrostatic interactions 18

2 **Analysis of Electrostatic Interactions** **23**

2.1 The importance of electrostatic interactions 23

2.2 Solvent–solute interactions 24

2.3 The continuum electrostatic model 25

2.4 Electrostatics in affinity and specificity 27

2.4.1 Electrostatic contributions to binding 27

2.4.2 Optimization of electrostatic interactions 29

3 **Electrostatic Optimization of Enzyme Ligands** **33**

3.1 Introduction 34

3.2 Methods 38

3.3 Results 40

3.3.1 Optimization of GlnRS ligands 40

3.3.2 Robustness of results 47

3.3.3 Specificity of glutamine over glutamic acid 51

3.4 Discussion 53

3.5 Conclusions 56

4 Design Methods for Peptide Inhibitors 59

4.1 Introduction 60

4.1.1 Inhibition of HIV-1 cell entry 60

4.1.2 Inhibitors targeting the N-terminal coiled coil of gp41 64

4.2 Database screening strategy 66

4.3 Methods 69

4.4 Results 72

4.4.1 Electrostatics of C34 and D-peptide binding 72

4.4.2 Optimization of D-peptide binding 74

4.4.3 Design of enhanced electrostatic interactions 78

4.4.4 Ligand scanning of D-peptide tryptophans 82

4.4.5 Stabilization of the D-peptide bound conformation 88

4.5 Discussion 90

4.6 Conclusions 96

5 Designing Improved Protein Inhibitors 99

5.1 Introduction 100

5.2 Methods 102

5.3 Results 106

5.3.1 5-Helix–C-peptide electrostatic binding free energy 106

5.3.2 Electrostatic contributions to 5-Helix binding 106

5.3.3 Optimization of 5-Helix binding 111

5.3.4 Binding energetics of 5-Helix mutants 117

5.3.5 Electrostatic contributions to 5-Helix stability 119

5.4 Discussion 121

5.5 Conclusions 126

6 "Action-at-a-Distance" Interactions 129

6.1 Introduction 130

6.2 Methods 132

6.3 Results 134

6.3.1 Electrostatic contributions to BLIP–TEM1 binding 134
6.3.2 Variation of results with internal dielectric 140
6.3.3 Optimization of BLIP binding 142
6.4 Discussion 149
6.5 Conclusions 154

7 *Ab initio* Charge Determination 157
7.1 Introduction 158
7.2 Methods 160
7.3 Results 161
7.3.1 Molecules representative of protein groups 161
7.3.2 Extended set of small organic molecules 166
7.4 Discussion 176
7.5 Conclusions 183

8 General Conclusions 187

A Residual Potentials and Affinity 191
A.1 Introduction 192
A.2 Results and Discussion 194
A.3 Conclusion 202
A.4 Methods 202

B ICE User's Manual 205
B.1 Introduction 205
B.1.1 Molecular binding free energy 205
B.1.2 Electrostatic binding free energies 206
B.1.3 Surface area dependent binding free energies 207
B.1.4 Calculation of relative free energies in solution 208
B.1.5 Component analysis of electrostatic interactions 210
B.1.6 Electrostatic complementarity: The residual potential 211
B.1.7 Optimization of electrostatic binding free energy 212

B.2 Theory 213
B.2.1 The continuum electrostatic model 213
B.2.2 Electrostatic contributions to binding 214
B.2.3 Optimization of electrostatic interactions 217
B.2.4 Type-I (Affinity) Optimum 218
B.2.5 Type-II (Specificity) Optimum 219
B.2.6 Type-III (Best Hapten) Optimum 219
B.3 Optimization in real-world problems 220
B.3.1 Fixed charges in optimization 220
B.3.2 Simultaneous optimization of multiple ligands 221
B.3.3 Optimization over multiple conformations 224
B.3.4 Poor behavior of optimization matrices 228
B.3.5 Application of constraints 229
B.4 Overview of Program 229
B.4.1 Continuum electrostatic calculations 231
B.4.2 Component analysis 231
B.4.3 Optimization 232
B.4.4 Optimization of protein ligands 234
B.4.5 Ligand scanning 234
B.5 Manual Pages 236
B.5.1 binpot 236
B.5.2 comp_anal.prl 238
B.5.3 delphi.prl 240
B.5.4 getmatrix 249
B.5.5 ligand_scan.prl 253
B.5.6 matrix_elements.prl 254
B.5.7 optimize 256
B.5.8 protein_scan.prl 262
B.5.9 rank_comp.prl 264
B.6 Sample Configuration Files 265

B.6.1 comp_anal.prl (component.cfg) 266
B.6.2 matrix_elements.prl (matrix.cfg) 266
B.6.3 optimize (optimize.cfg) 271
B.6.4 ligand_scan.prl (ligandscan.cfg) 276

Bibliography **279**

Curriculum Vitae **299**

Chapter 1

General Introduction

1.1 Electrostatics and biomolecular association

Molecular association, and in particular the association of proteins with various other molecules, plays a central role in biology: enzymes must bind their substrates; signaling receptors must bind their target signal molecules; regulatory proteins must bind their appropriate targets. In addition, the vast majority of drugs act by binding to, and thus affecting the activity of, one or more protein targets within the body. In all of these cases, there is a need for an appropriate balance of affinity and specificity in the binding reaction. If two molecules in a complex are required to dissociate as a requirement for function, as, for example, in enzyme release of a product or in signals transduced by transient interactions, the affinity of the complex can not be too high. On the other hand, in the design of many drug molecules, and for natural enzyme inhibitors such as bovine pancreatic trypsin inhibitor (BPTI) and barstar, the goal may to be to attain maximal affinity. Similar variations exist in the requirement for specificity in binding. In some cases, such as in many of the associations involved in protein synthesis and DNA replication, as well as in the design of drugs meant to act on a single target, an extremely high degree of specificity may be essential. In other situations, though, a much weaker degree of specificity may be beneficial — this is the case for enzymes which act on multiple substrates, and in the design of drugs that will be active against a similar target in a range of pathogens.

It is generally accepted that the force which drives most biologically relevant molecular association events is non-electrostatic. In particular, the so-called "hydrophobic effect", related to the substantial energetic cost of disrupting the structure of bulk water upon solvation of most molecules, substantially favors the bound state [24, 26, 135]. This is due to the approximate dependence of the hydrophobic effect on molecular surface area (the greater the surface area of a molecule, the greater the disruption of the water structure) [24, 25]; barring any large scale conformational changes, the surface area of a complex will always be significantly smaller than the sum of the surface areas of the two free ligands.

While the hydrophobic effect may contribute the majority of the stabilization of a complex, other energetic contributions can be equally important. A high degree of shape complementarity has been recognized as being very important for high affinity ligands — a steric clash can severely reduce the binding affinity of a complex, and, since most solutes make near optimal van der Waals interactions with solvent in the unbound state, a lack of optimal contacts in the bound state can also reduce the binding affinity. Entropic changes, both due to loss of translational and rotational degree of freedom and due to changes in the populations of internal degrees of freedom, also can play an important role [53, 151], leading, for example, to a preponderance of rigid molecules among small molecule drugs. Finally, electrostatic interactions must be considered, as the binding interfaces of most associating molecules are at least somewhat polar [27].

Electrostatic interactions play an important, and interesting, role in modulating the free energy of complex formation. Due to the favorable interactions with solvent that any polar group makes in the unbound state, the net energetic contribution of electrostatic interactions which are made in the complex is not necessarily favorable [69]. In fact, for many of the complexes studied to date, the reverse seems to generally true — the overall electrostatic contribution to the binding free energy is unfavorable [69, 94]. While initially counter-intuitive, since so many electrostatic interactions are observed in biomolecular complexes, this is not particularly surprising, as the primary role of electrostatic interactions may be to impart specificity to the association reaction. While the seemingly well-designed electrostatic interactions seen in complexes may not contribute favorably to the binding energetics, if the polar groups involved were not oriented appropriately in the bound state, making interactions to compensate the loss of interactions made with solvent in the unbound state, the contribution would be substantially more unfavorable. Thus, by requiring reasonable electrostatic interactions to be made in the bound state in order to balance the unfavorable desolvation penalty, a single orientation of binding can be enforced, as can specificity against molecules with a different distribution of polarity.

Although the electrostatic interactions in existing complexes may tend to be unfavorable, this does not necessarily imply that the appropriate electrostatic interactions can not contribute to the favorable binding free energy of high affinity ligands. To the contrary, it may be that the highest affinity ligands are those for which the electrostatics are the most favorable — reducing a generally unfavorable contribution to binding, or even making a favorable contribution, will lead to tighter binding — and some initial studies suggest this is the case [93]. Furthermore, optimizing the electrostatic interactions that a ligand makes on binding may provide a useful means by which to design novel high affinity ligands. It is these questions which are addressed in the work described here. Using computational methods for the study of protein complexes, the optimization of electrostatic interactions is investigated, both in the analysis of natural complexes and in ligand design, with promising results for further applications of these techniques.

1.2 Computational studies of biomolecules

Over the past twenty-five years, advances in theory, coupled with the explosion of computational power, has stimulated the rapid development of methods to analyze biological systems *in silico*. Computational studies of biological systems can play an important role complementary to experimental studies. One particularly useful application of theoretical studies has been in separating out the individual contributions of the various parts of a complex, and the various parts of the energy, in ways inaccessible by experiment. This separation often allows for a more intuitive understanding of the energetics, while maintaining a rigorous framework for quantitative analysis.

Biological systems are very large and very complex, and thus the study of biomolecules at the most fundamental level — quantum mechanics — is infeasible. However, a great deal of success has been had by applying theories and methods developed for macroscopic physical systems to the microscopic systems of biological macromolecules. Two highly successful examples of this are molecular mechanics and continuum electrostatics.

Molecular mechanics. Molecular mechanics methods treat molecules as a collection of atoms described by a mass and a partial charge [11, 29, 73, 100, 152, 159]. Bonds are described energetically by springs of an appropriate strength, and similar terms are used to describe the interactions of atoms connected by two or three bonds (bond angles and dihedral angles, respectively). Interactions between non-bonded atoms are described by Lennard–Jones or similar potentials for van der Waals interactions, and by a Coulomb's Law-type potential for electrostatic interactions. Energies of states can be evaluated individually, or molecular dynamics can be propagated using Newton's laws of motion. Calculations may be done *in vacuo*, or solvent molecules may be explicitly included in the description of the system.

Molecular mechanics methods have been applied to numerous problems in biology [82, 157]. The dynamics of macromolecules [32, 44, 81, 103, 131], and the relation of these dynamics to function [12, 13, 99], have been studied in detail. The effects of mutations on the energetics and dynamics of biological macromolecules have also been extensively studied by these means [30, 52, 120, 145, 150]. In the area of molecular recognition, molecular mechanics methods have been applied to unraveling the energetic contributions to binding thermodynamics and kinetics [45, 89, 115, 121, 126], and to understanding the differences in binding of chemically related ligands to a common receptor [104, 123, 156]. In addition, molecular mechanics force fields have been used in *de novo* design applications, both of ligands for protein targets [36, 38, 74, 95, 122], and of novel proteins themselves [31, 35, 59, 60, 72, 85, 91].

Continuum electrostatics. A second method derived from macroscopic physics which has been successfully applied to biological systems is the continuum electrostatic model [56, 148, 158]. In this approach, molecules are generally described as a set of point charges located at atomic centers embedded in a region of low dielectric constant described by the molecular surface [56, 57, 108]. Solvent is described implicitly as a region of high dielectric constant, possibly with some description of mobile ions. In a commonly used approach, the electrostatic potential produced by such a system can be obtained by solution of the linearized Poisson–Boltzmann equation. The electrostatic free energy of any system of charges can then be obtained by taking the sum over all charges of one half the product of electrostatic potential and the partial charge. For the evaluation of electrostatic free energies, continuum electrostatics provides a significant benefit over molecular mechanics. In order to account for solvation effects in an explicit solvent model, the configurational space of the solvent must be adequately sampled, which can be a highly computationally expensive process. However, in an implicit model, solvent rearrangement and configurational sampling is included in the continuum description of the solvent, and thus continuum methods are much more computationally efficient.

Continuum electrostatic calculations have been applied nearly as diversely as have molecular mechanics applications [70, 137]. Numerous studies have investigated the nature of the electrostatic field in and around biological molecules [56, 83, 158]. Continuum solvation methods have also been applied to the prediction of the $\mathrm{pK_a}$s of both small molecules and protein side chains [2, 18, 118, 153, 165]. However, by far the greatest number of applications of continuum electrostatics has been in determining the details of the energetic contributions of electrostatic interactions to protein stability [54, 65, 68, 164] and to protein–ligand association [3, 51, 69, 106, 107, 166].

1.3 Optimization of electrostatic interactions

Within the linearized Poisson–Boltzmann model, all charges act independently, and thus the contributions to the electrostatic free energy from various parts of the system are separable. As a result, the electrostatic contribution of any group of atoms to the free energy of a system can be decoupled from the rest of the system. This enables a complete breakdown of important quantities such as binding affinity and protein stability into contributions from various groups, and thus allows an analysis of the system to be used to pinpoint key contributors to both the stabilization and the destabilization of proteins and protein complexes. Such an approach has been used in the study of binding in several protein–protein [66, 69, 113] and protein–DNA [17, 64] complexes, as well as in the study of electrostatic contributions to protein stability [65, 68, 141]. In addition, the ability to separate individual atomic contributions to the electrostatic binding free energy, and the linear response of the electrostatic potential with respect to variation of the magnitude of the partial atomic charges, allows the energy to be written as a product of the charges on the ligand and receptor, with matrices dependent only on the binding geometry describing the desolvation of the ligand and of the receptor as well as the ligand–receptor interaction. This formulation gives rise to an electrostatic optimization procedure, in which the charge distribution on a ligand or receptor which optimizes the electrostatic contribution to either the affinity or the specificity of binding can be computed [77–79, 92]. The theoretical bases of these results are outlined in Chapter 2, as these methods form the foundation on which the bulk of the work detailed here rests.

Some initial applications of the electrostatic optimization theory have been presented previously, but have been somewhat limited in scope. Electrostatic optimization in the barnase–barstar enzyme–inhibitor complex showed that, while significant gains in binding affinity could be gained by electrostatic optimization, wild-type barstar is close to optimal for binding barnase, particularly in the context of natural amino acids [23, 93, 94]. Analysis of ligand binding in chorismate mutase from *Bacillus subtilis* revealed close to optimal charges in regions making close interactions in the bound state, and furthermore suggested a role played by an electrostatic preference for the transition state in promoting catalysis [76, 80]. The purpose of the work described here is to demonstrate several further applications of the electrostatic optimization methodology, both in furthering our understanding of the role of electrostatic interactions in the formation of natural complexes, and in the design of ligands with enhanced binding affinity.

Electrostatic optimization and enzyme–substrate binding. Chapter 3 focuses on applying the optimization procedure to enzyme substrates, and using the comparison of natural and optimal charges to gain insight into how enzymes have evolved to use electrostatic interactions in determining affinity and specificity in binding to their cognate substrates. The enzyme chosen is glutaminyl-tRNA synthetase (GlnRS) which plays a key role in protein synthesis, and thus must be highly specific in order to minimize errors leading to dysfunction. GlnRS is particularly interesting for the study of electrostatic optimization as it has two polar cognate substrates and must discriminate against polar, charged, and hydrophobic alternatives.

Electrostatic optimization and ligand design. Chapters 4 and 5 deal with applications of electrostatic optimization to the design of inhibitors of HIV-1 viral–

cell membrane fusion. In Chapter 4 we consider a D-peptide inhibitor targeting a trimeric coiled coil of the viral glycoprotein gp41, known to be a useful target for inhibiting a large-scale conformational change required for membrane fusion. We also describe a hierarchical scheme for efficiently evaluating binding free energies for a large database of potential ligand modifications to an arbitrary level of detail and accuracy. In Chapter 5 we focus on a protein inhibitor of membrane fusion which targets a separate region of gp41, but the same conformational change. The charge optimization procedure is used to identify several positions whose mutation to another natural amino acid is predicted to improve binding.

"Action-at-a-distance" electrostatic interactions. Chapter 6 outlines an alternative approach to designing mutations which enhance the binding affinity of a protein ligand to its target receptor. Interactions involving charged groups on the periphery of a binding interface can make interactions through a region of solvent. While these interactions may be significantly screened by solvent, this is balanced by a smaller desolvation cost on binding relative to interfacial residues. In addition, the moderate range of the interaction makes the design of these interactions less sensitive to imperfections in structural models. The energetics of these interactions in the complex of β-lactamase inhibitor protein with TEM1 β-lactamase are explored and are compared with those of interactions involving more buried groups. The use of residual potentials (a measure of electrostatic complementarity) for interpreting these results is presented in Appendix A.

Methods of *de novo* charge development. Chapter 7 diverges a little in content. Continuum solvation methods generally require a set of partial atomic charges for all molecules, but for many small molecule ligands, these parameters are not available. We have evaluated the performance of numerous *ab initio* charge determination methods in continuum solvation calculations. Charges were computed for a large set of small organic molecules, and solvation free energies were calculated using these charges in a continuum model. The computed solvation energies were compared to experimental values to evaluate which methods perform particularly well, and which perform more poorly.

Software for electrostatic optimization. Appendix B contains the manual for the ICE (Integrated Continuum Electrostatics) suite of computer software. This software package, which extends the work of Erik Kangas and Zachary Hendsch, is a complete set of tools for the analysis and optimization of electrostatic contributions to binding in biomolecular complexes.

Chapter 2

Analysis of Electrostatic Interactions: Methods and Theory

2.1 The importance of electrostatic interactions

Of the many interactions made between associating molecules, electrostatic interactions are particularly interesting for several reasons. It is generally accepted that the driving force for most macromolecular association events is the hydrophobic effect, the entropic benefit of releasing solvent from the binding surfaces of each molecule [24, 26, 135]. However, this effect is non-specific, with any burial of the same surface area contributing equally [24, 25]. Van der Waals interactions are also relatively non-specific, with only substantial steric clashes resulting in large unfavorable energies, and favorable interactions being quite small in magnitude. On the other hand, electrostatic interactions are highly specific; electrostatic interaction energies can range from highly favorable to highly unfavorable depending on the identity and geometry of the interacting groups. Furthermore, electrostatic interactions act over a significantly longer-range — the energy of interaction between two charged groups falls off linearly with distance, and the interaction of two dipoles decreases with the cube of the distance — than do van der Waals interactions, which decrease with the sixth power of the distance between the interacting groups. In addition, solvation effects can make the energetics of electrostatic interactions non-intuitive; groups making favorable interactions in the bound state of a complex may make even more favorable interactions with solvent in the unbound state, making the net contribution to binding unfavorable [69]. While it is relatively clear that the most favorable van der Waals interactions are made by making the maximal contact between groups without steric interference, and that the hydrophobic effect favors the burial (and conversely disfavors the solvent exposure) of non-polar groups, in order to understand electrostatic interactions it is necessary to consider in detail both the bound and unbound states.

2.2 Solvent–solute interactions

Solvent plays a key role in determining the behavior and energetics of biological systems. As essentially all of biology takes place in an aqueous, moderate ionic strength environment, it is the behavior of biological molecules in this aqueous milieu that imparts their function. Thus, in order to understand how biological molecules interact with each other, it is also necessary to understand how they interact with water. Solvent–solute interactions can loosely be classified into three types. Firstly, solvent molecules can associate with the solute with a reasonably high binding affinity. In large molecular systems, such as proteins and protein complexes, polar groups on the solute can coordinate water molecules with up to four stable hydrogen bonds, potentially making the bound state significantly more favorable than that of bulk solvent, despite the entropic penalty for reducing the mobility of the water molecule. Secondly, as water molecules are both highly polar and highly mobile (in the liquid state), water can react strongly to the electrostatic field of a solute. In the primary solvation shell of a solute, these interactions may involve transient solute–solvent hydrogen bonds, but the water molecules are not tightly bound and exchange freely and rapidly with one another. The electrostatic field can extend a significant distance into solvent, with the water molecules many layers removed from the solute reorienting (in an average sense) to interact with the solute's electrostatic field. Thirdly, in bulk water every water molecule interacts strongly, if transiently, with all its neighbors in a tetrahedral geometry, and the introduction of a non-polar solute into water disrupts this network, leading to a reorganization of the solvent, with a significant entropic cost. This hydrophobic effect drives the association of non-polar groups in water, with many important ramifications. All these interactions are present in biological systems, and all must be appropriately considered in order to understand the energetics of molecular association in biological environments.

2.3 The continuum electrostatic model

In the computation of biomolecular energetics, to treat solvent explicitly, by placing the system of interest within a large region of individually considered solvent molecules, is very costly. In such a system, to compute the hydrophobic and bulk electrostatic interactions a solute makes with the solvent requires that the conformational space of the solvent is adequately sampled, which, with six degrees of freedom available to every solvent molecule, is a very computationally intensive process. An alternative to the explicit modeling of solvent is to employ a continuum model, considering the effects of solvent as a bulk entity, rather than as a microscopically distinct ensemble of molecules. For hydrophobic interactions, this treatment most often leads to a surface area dependent energy term [24, 25, 140]; the greater the surface area of the solute, the greater the required reorganization of water, and thus the greater the entropic penalty. For electrostatic interactions, a dielectric continuum model is frequently used [56, 148, 158]. The dielectric response mimics the average reorientation of water molecules in an electrostatic field, including both the favorable enthalpic interaction term and the entropic cost of orienting of the solvent. The effect of mobile ions can also be treated, using a bulk treatment such as the Debye–Hückel model.

In the continuum electrostatic approach, molecules are frequently described as a set of point charges located at atomic centers embedded in a region of low dielectric described by the molecular surface, with the solvent treated as a region of higher dielectric with, possibly, some concentration of mobile ions [55, 57, 108]. The electrostatic potential produced by such a system can be obtained by solution of the Poisson–Boltzmann equation:

$$\vec{\nabla} \cdot [\epsilon(\vec{\mathrm{r}})\vec{\nabla}\phi(\vec{\mathrm{r}})] - \epsilon(\vec{\mathrm{r}})\kappa^2(\vec{\mathrm{r}})\mathrm{sinh}[\phi(\vec{\mathrm{r}})] = -4\pi\rho(\vec{\mathrm{r}}) \tag{2.1}$$

where $\kappa^2 = \frac{8\pi z^2 I}{ekT}$ describes the effect of mobile ions using a Debye–Hückel model. From the electrostatic potential, the electrostatic free energy of any system of charges is given by $\mathrm{G} = \frac{1}{2}\sum_i \phi_i q_i$, with the sum taken over all charges.

When the electrostatic potential in solvent is relatively small, as is the case for many systems of biological interest, the Poisson–Boltzmann equation can be linearized, replacing the hyperbolic sine dependence of the salt term with the first term in the series expansion ($\sinh(x) = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots$) and yielding:

$$\vec{\nabla} \cdot [\epsilon(\vec{\mathrm{r}})\vec{\nabla}\phi(\vec{\mathrm{r}})] - \epsilon(\vec{\mathrm{r}})\kappa^2(\vec{\mathrm{r}})\phi(\vec{\mathrm{r}}) = -4\pi\rho(\vec{\mathrm{r}}) \tag{2.2}$$

Within this linearized Poisson–Boltzmann model, all charges act independently, and thus the contributions to the electrostatic free energy from various parts of the system are separable. When the contribution from any subset of the system can be considered independently, with the total energy being a simple sum of the various parts, the binding free energy can easily be partitioned into the contributions from each molecule, each functional group, or even each atom. This ability to partition the energy simply and rigorously makes the linearized Poisson–Boltzmann equation a powerful tool for the analysis of electrostatic interactions in macromolecular systems.

2.4 Electrostatics in affinity and specificity

2.4.1 Electrostatic contributions to binding

Previous work by Hendsch and Tidor made significant in-roads to understanding the electrostatic contribution to the binding energetics in several biological systems [67, 69], as well as related work on electrostatic contributions to protein stability [65, 66, 68]. Separating the contributions of various chemical groups (for example, side chain, backbone amino and backbone carbonyl groups for proteins, and base, ribose and phosphate groups for nucleic acids) provides a logical separation of contributions to the energy into three terms for every group. These are the desolvation energy of the individual group, the solvent screened interactions between the group and all groups on the binding partner in the bound state (intermolecular interactions), and the difference in solvent screening of the interactions between the group and other groups on the same molecule in the bound and unbound states (intramolecular interactions). These are termed the desolvation, the direct interactions, and the indirect interactions, respectively, and we can reconstitute the full electrostatic binding energy by:

$$\Delta \mathrm{G}^{es} = \sum_{i} \Delta \mathrm{G}_{i}^{solv.} + \sum_{i} \sum_{j} \Delta \mathrm{G}_{ij}^{dir.} + \frac{1}{2} \sum_{i} \sum_{j} \Delta \mathrm{G}_{ij}^{indir.} \tag{2.3}$$

with the indirect terms halved to avoid double counting.

In addition to the individual group solvation energies and the pair-wise interaction energies, two measures of the overall contribution of a group to the binding free energy can readily be defined. The first, denoted the mutation energy, corresponds the difference in binding free energy of the natural system and that of a hypothetical system in which the group in question (and only that group) is replaced by a hydrophobic isostere. That is, the mutation energy is the energy gained by “turning on” the charges on the group of interest in the context of the natural charges at the atoms of the rest of the system. When the charges on a group are eliminated, all the interactions made by that group are lost along with the desolvation of the group, and thus the mutation energy is defined as:

$$\Delta \mathrm{G}_i^{mut.} = \Delta \mathrm{G}_i^{solv.} + \sum_j \Delta \mathrm{G}_{ij}^{dir.} + \sum_j \Delta \mathrm{G}_{ij}^{indir.} \tag{2.4}$$

While the mutation energy is particularly useful in that it corresponds exactly to a physical transformation (if not one that can be experimentally implemented), it suffers one drawback — the sum of the mutation energies of every group does not equal the binding free energy, because all interactions are counted twice. As it is useful for understanding a system to be able to partition the energy between groups, the contribution energy is defined as:

$$\Delta \mathrm{G}_i^{contrib.} = \Delta \mathrm{G}_i^{solv.} + \frac{1}{2} \sum_j \Delta \mathrm{G}_{ij}^{dir.} + \frac{1}{2} \sum_j \Delta \mathrm{G}_{ij}^{indir.} \tag{2.5}$$

such that the sum of all contribution energies is the total electrostatic binding free energy. While useful for partitioning the energy between groups in a meaningful way, the contribution energy does not correspond to any physical transformation. Thus neither the contribution nor the mutation energy is a perfect measure, but both are complementary, and used together can give significant insight to how various groups contribute to the overall energetics of binding.

The overall electrostatic contribution to binding free energy of a ligand (l) binding to a receptor (r) can be written as:

$$\Delta \mathrm{G}^{es} = \Delta \mathrm{G}_l^{hyd.} + \Delta \mathrm{G}_{r,l}^{int.} + \Delta \mathrm{G}_r^{hyd.} \tag{2.6}$$

where $\Delta \mathrm{G}_{r,l}^{int.}$ is the solvent screened interaction free energy between the receptor and ligand in the bound state given by:

$$\Delta \mathrm{G}_{r,l}^{int.} = \sum_{i \in r} \sum_{j \in l} \Delta \mathrm{G}_{ij}^{dir.} \tag{2.7}$$

$\Delta \mathrm{G}_{l}^{hyd.}$ is the change is the ligand hydration free energy on binding given by:

$$\Delta \mathrm{G}_{l}^{hyd.} = \sum_{i \in l} \Delta \mathrm{G}_{i}^{solv.} + \frac{1}{2} \sum_{i \in l} \sum_{j \in l} \Delta \mathrm{G}_{ij}^{indir.} \tag{2.8}$$

and $\Delta \mathrm{G}_{r}^{hyd}$ is the equivalent term for the receptor.

2.4.2 Optimization of electrostatic interactions

Breaking down the electrostatic binding free energy further, and considering every atom in the system as its own group, leads to a particularly interesting result. When each group is an atom, the solvation free energy of each group can be written as $\Delta \mathrm{G}_{i}^{solv.} = \frac{1}{2}(\phi_{ii}^{bound} - \phi_{ii}^{unbound})q_i$, where ϕ_{ii} is the potential produced by charge i at position i. However, due to the linear response of the linearized Poisson–Boltzmann model, the potential produced by any charge at position i can be related to the potential produced by a single unit charge at the same position (Φ_i) by $\phi_i = q_i \Phi_i$. This leads to an expression for the atomic solvation energy in terms of the partial atomic charge and the bound and unbound potentials of a unit charge at the atom center:

$$\Delta \mathrm{G}_{i}^{solv.} = \frac{1}{2} q_i (\Phi_{ii}^{bound} - \Phi_{ii}^{unbound}) q_i \tag{2.9}$$

Similarly, with single atom groups the pairwise indirect interactions can be written in terms of the potential generated by charge i at position j as $\Delta \mathrm{G}_{ij}^{indir.} = (\phi_{ij}^{bound} - \phi_{ij}^{unbound})q_j$, into which the substitution of $\phi_{ij} = q_i \Phi_{ij}$ gives:

$$\Delta \mathrm{G}_{ij}^{indir.} = q_i (\Phi_{ij}^{bound} - \Phi_{ij}^{unbound}) q_j \tag{2.10}$$

Using the same procedure for the direct interactions yields:

$$\Delta \mathrm{G}_{ij}^{dir.} = q_i(\Phi_{ij}^{bound})q_j \tag{2.11}$$

with only the bound state potentials contributing. For both the direct and indirect interactions, $\Delta \mathrm{G}_{ij} = \Delta \mathrm{G}_{ji}$, by the reciprocity implicit in the continuum model.

Substituting Equation 2.11 into Equation 2.7 gives:

$$\Delta \mathrm{G}_{r,l}^{int.} = \sum_{i \in r} \sum_{j \in l} q_i(\Phi_{ij}^{bound})q_j \tag{2.12}$$

This can be written in matrix form as $\vec{Q}_r^{\dagger}\mathbf{C}\vec{Q}_l$, where $\vec{Q}_r$ is a vector of the charges on the receptor, $\vec{Q}_l$ is a vector of the charges on the ligand, and the elements of the matrix $\mathbf{C}$ are given by Φ_{ij}^{bound}. In a similar fashion, substituting Equations 2.9 and 2.10 into Equation 2.8, gives:

$$\Delta \mathrm{G}_{l}^{hyd.} = \frac{1}{2}\sum_{i \in l} q_i(\Phi_{ii}^{bound} - \Phi_{ii}^{unbound})q_i + \frac{1}{2}\sum_{i \in l}\sum_{j \in l} q_i(\Phi_{ij}^{bound} - \Phi_{ij}^{unbound})q_j \tag{2.13}$$

This too can be written in matrix form as $\vec{Q}_l^{\dagger}\mathbf{L}\vec{Q}_l$, where the diagonal elements of the matrix $\mathbf{L}$ are given by $\frac{1}{2}(\Phi_{ii}^{bound} - \Phi_{ii}^{unbound})$, and the off-diagonal elements are given by $\frac{1}{2}(\Phi_{ij}^{bound} - \Phi_{ij}^{unbound})$. Naturally, the change in receptor hydration free energy on binding, $\Delta G_r^{hyd.}$, can be written in the same fashion as $\vec{Q}_r^{\dagger}\mathbf{R}\vec{Q}_r$, with the receptor desolvation matrix, $\mathbf{R}$, analogous to the ligand desolvation matrix, $\mathbf{L}$. Combining these terms gives an expression for the overall electrostatic binding free energy in matrix form:

$$\Delta \mathrm{G}^{es} = \vec{Q}_l^{\dagger}\mathbf{L}\vec{Q}_l + \vec{Q}_r^{\dagger}\mathbf{C}\vec{Q}_l + \vec{Q}_r^{\dagger}\mathbf{R}\vec{Q}_r \tag{2.14}$$

For a given receptor and a fixed ligand geometry, $\vec{Q}_r$, $\mathbf{L}$, $\mathbf{R}$, and $\mathbf{C}$ are all constant and thus the electrostatic binding free energy depends only on the ligand charges. $\Delta \mathrm{G}^{es}$ is quadratic in $\vec{Q}_l$ (see Figure 2-1), thus forming a paraboloid in ligand charge

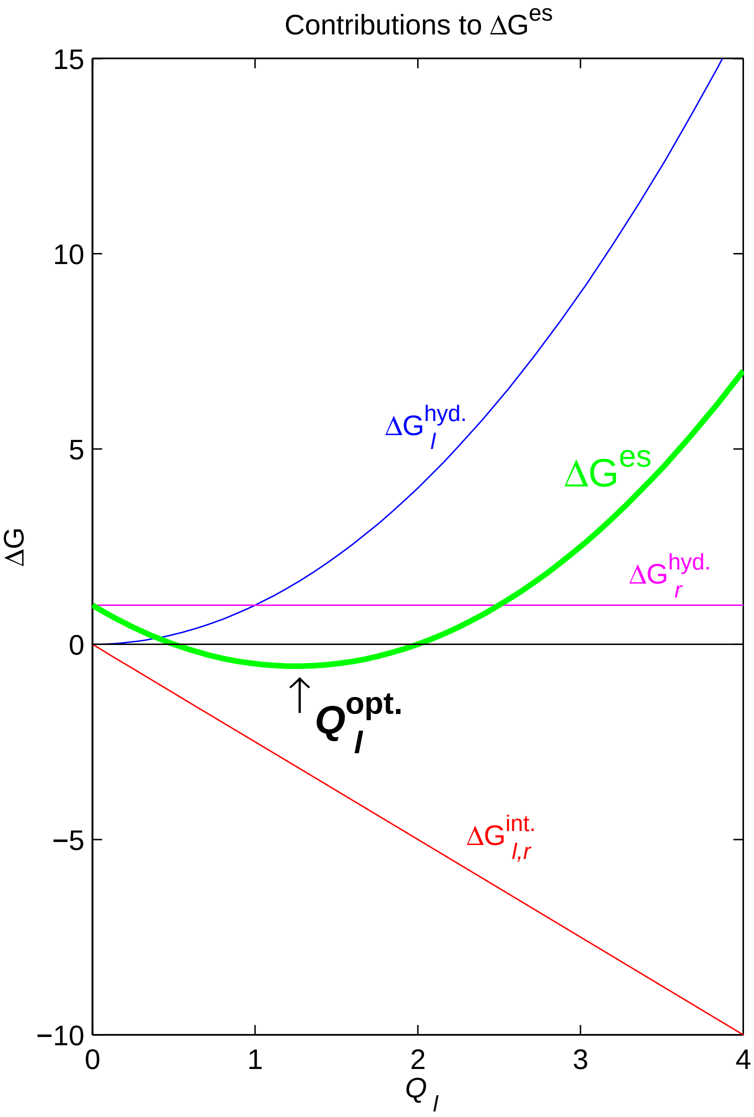


Figure 2-1: **The electrostatic binding free energy varies quadratically with ligand charge.** The desolvation free energy of the ligand varies with the square of the charges on the ligand, while the free energy of interaction with the receptor varies linearly with the ligand charges. As the receptor desolvation free energy is independent of the ligand charges, the net electrostatic binding free energy is a quadratic function of the ligand charge distribution. As a result, there is a single minimum on the free energy surface, corresponding to the optimal ligand charge distribution.

space and allowing $\Delta\mathrm{G}^{es}$ to be easily minimized with respect to $\vec{Q}_l$[1]:

$$\frac{\partial\Delta\mathrm{G}^{es}}{\partial\vec{Q}_l} = 2\vec{Q}_l^{\dagger}\mathbf{L} + \vec{Q}_r^{\dagger}\mathbf{C} = 0 \tag{2.15}$$

This gives a ligand charge distribution which optimizes the electrostatic contribution to the binding free energy, the optimal charge vector being given by:

$$\vec{Q}_l^{\,\mathrm{opt}} = -\frac{1}{2}\mathbf{L}^{\mathbf{-1}}\mathbf{C}^{\dagger}\vec{Q}_r \tag{2.16}$$

The theory behind these results, and related expressions for optimization of specificity in binding, has been derived in detail by Lee and Tidor [92] and by Kangas and Tidor [77–79]. Optimal charge distributions which meet certain specifications (such as a fixed total charge or proportionalities of certain partial atomic charges) can easily be obtained by minimizing the binding free energy with respect to the ligand charges, subject to the applied constraints, using a variety of standard methods.

---

[1] It has been shown that the **L** and **R** matrices are positive definite [77], essentially because it can never be electrostatically favorable to desolvate a molecule. As a result of the positive definite nature of the matrices, the global minimum of $\Delta\mathrm{G}^{es}$ with respect to variation of $\vec{Q}_l$ is the single stationary point on the free energy surface.

Chapter 3

Electrostatic Optimization of Enzyme Ligands: A Study of Glutaminyl-tRNA Synthetase

Abstract

Molecular mechanisms have evolved to impart appropriate affinity and specificity to protein interactions. Here we have analyzed the interactions of an aminoacyl-tRNA synthetase for which strong evolutionary pressure is believed to enforce strong specificity of substrate binding and catalysis. Electrostatic interactions have been hypothesized to be particularly efficient at enhancing binding specificity, and the effects of charged and polar groups were the focus of this study. The binding of glutaminyl-tRNA synthetase from *Eschericia coli* to several ligands, including the natural substrates, was analyzed. An affinity optimization procedure based on continuum electrostatics was used to evaluate the relative complementarity of the enzyme to its ligands. The natural and optimal ligand charges show remarkable agreement, most significantly in regions somewhat removed from the sites of chemical reaction. In particular, regions of the ligands observed to make several electrostatic interactions with the enzyme in the bound state have optimal charge distributions with identical positive–negative patterning, as well as similar magnitudes, as those of the natural ligands. The enzyme's cognate substrates are, in the regions where specific binding is presumed to be an important goal of the enzyme, very close to optimal, and thus the results suggest that the optimization of electrostatic interactions has played an important role in guiding the evolution of this enzyme.

3.1 Introduction

The affinity of a ligand for an enzyme to which it binds is dependent on a number of factors. It has been shown that shape complementarity of the ligand to the binding site plays an important role, and that the driving force in many cases is the entropic benefit of the release of structured water from around the ligand and within the binding site [24, 26, 135]. In addition, the unfavorable entropic cost of binding a flexible molecule in a single conformation has been recognized [53, 151]. The role of electrostatic interactions, such as hydrogen bonds and salt bridges, however, is somewhat less well understood. It is clear from the analysis of the structures of complexes of small molecules and their receptors that electrostatic interactions are made in the bound state [27]. However, since any polar group will make favorable interactions with water in the unbound state, the energetic role of electrostatic interactions is not obvious. In principle, electrostatics can play an important role in binding specificity, both in terms of binding a cognate substrate over a decoy, as well as in determining a unique orientation of binding; uncompensated polar and charged groups buried at a binding interface incur a large energetic cost due to desolvation. Likewise, even if compensating interactions are present, the large desolvation penalty incurred implies that the net electrostatic effect on binding need not be favorable, and the details are likely to be system dependent [69, 94]. The penalty for even partially undercompensated polar groups appears significant, as electrostatic interactions appear very close to optimal for the tight-binding barnase–barstar complex [93]. In contrast to the large amount of work that has been done to investigate effects on protein stability and binding affinity, much less is known about the determinants of specificity.

Aminoacyl-tRNA synthetases (aaRS) play a vital role in cells, catalyzing the aminoacylation of transfer RNA (tRNA) as a preliminary step in protein synthesis. The reaction takes place in two steps: the amino acid is first activated by reaction with ATP to form an aminoacyl-adenylate (Equation 3.1); a free hydroxyl on the 3′-terminal adenosine of the tRNA then displaces the adenyl moiety to form the aminoacyl-tRNA complex (Equation 3.2).

$$\begin{aligned} \text{aa} + \text{ATP} &\rightleftharpoons \text{aa-AMP} + \text{PP}_i && (3.1) \\ \text{aa-AMP} + \text{tRNA} &\rightleftharpoons \text{aa-tRNA} + \text{AMP} && (3.2) \end{aligned}$$

In order for the genetic code to be faithfully translated from messenger RNA to polypeptide, aaRSs must be highly specific, both for the correct tRNA and for the correct amino acid.

Glutaminyl-tRNA synthetase (GlnRS) must be able to effectively discriminate between its cognate amino acid substrate and other amino acids of similar size and shape. In particular, the enzyme must be highly selective for glutamine and against glutamic acid, which in its protonated state differs from glutamine solely by the replacement of the amide NH_2 by a hydroxyl. In addition, asparagine differs in structure from glutamine by a single methylene in the aliphatic portion of the side chain, and both leucine and methionine are of a similar size to glutamine but are hydrophobic. While GlnRS is one of only three aaRSs (along with GluRS and ArgRS) which require tRNA for the activation reaction (Eq. 3.1), the basis of this requirement is still unclear, as the enzyme will bind both the substrates and the product of this reaction in the absence of tRNA [48]. Atomic resolution structures of GlnRS bound to a variety of ligands, including ATP and an analogue of the glutaminyl-adenylate intermediate, have been determined [4, 5, 117, 124] (see Figure 3-1 for an overview of the structure).

Due to the sequential nature of the biochemical reaction, there are two opportunities for the enzyme to enforce specificity based on affinity alone (Figure 3-2). Differential binding affinities of free amino acids will result in differential rates of production of intermediate, and, as long as the intermediate has a lifetime which is long compared to the dissociation rate, differential affinities for the intermediate may

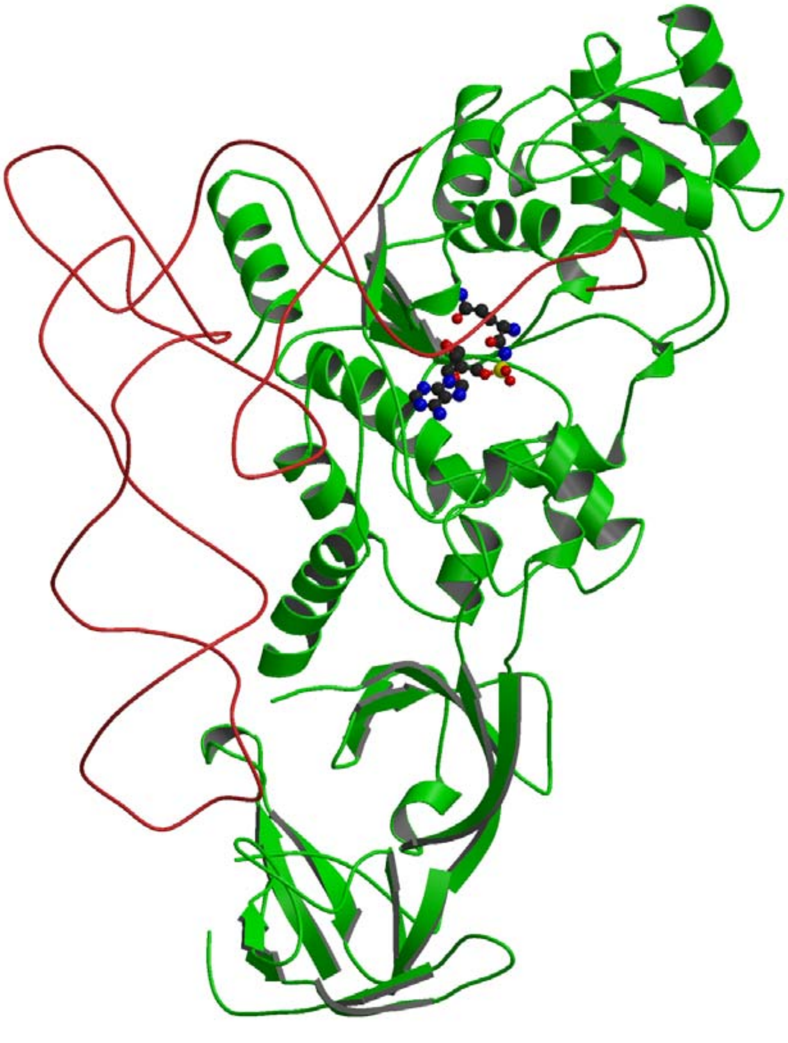


Figure 3-1: **Structure of GlnRS in complex with tRNA$^{\mathrm{Gln}}$ and an analogue of Gln-AMP.** GlnRS is displayed in green ribbon, tRNA$^{\mathrm{Gln}}$ in red tube, and QSI (a Gln-AMP analogue) in atom colored ball-and-stick representation. This figure was prepared with MOLSCRIPT [87] and RASTER3D [105].

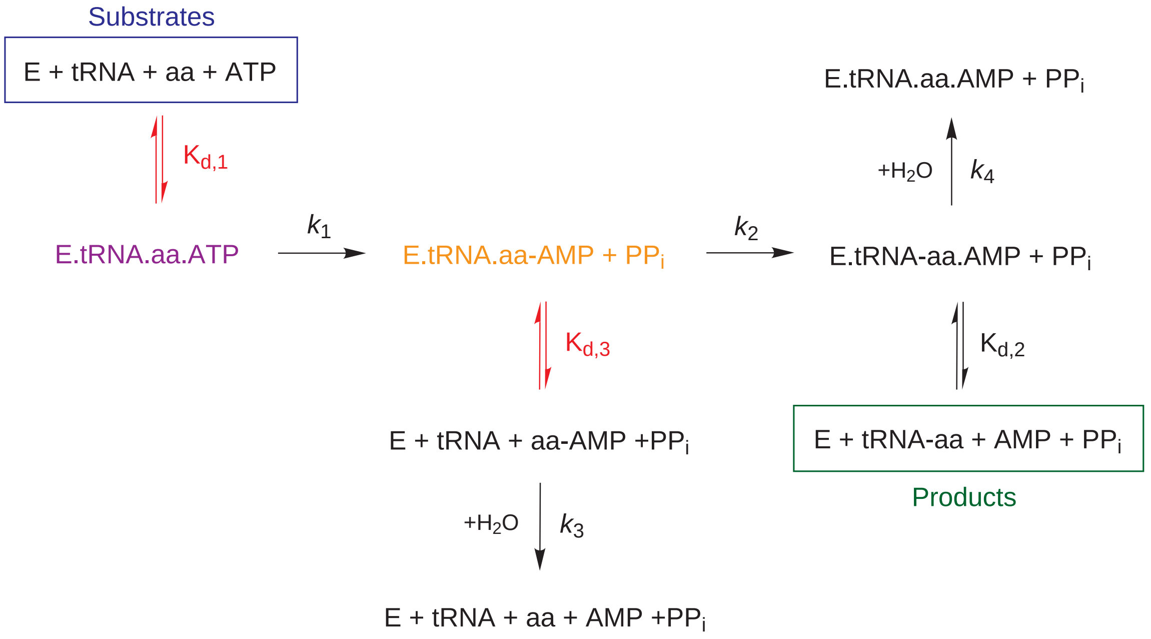


Figure 3-2: **Opportunities for enforcement of specificity in aminoacyl-tRNA synthetases.** Two primary means of discrimination by affinity alone are available in the transformation of amino acids and tRNA into charged aminoacyl-tRNAs. Differential binding affinities of the free amino acids in the first binding step ($\mathrm{K}_{\mathrm{d},1}$) clearly have an effect, but so may differential binding affinities of the aminoacyl-adenylate intermediate($\mathrm{K}_{\mathrm{d},3}$) . A post-charging editing step (k_4) is known to be active in some systems, but has not been observed in GlnRS.

provide a secondary means of providing specificity[1] . Even if the intermediate has a relatively short lifetime, any differences in the dissociation rate which perturb the

---

[1] Assuming that binding equilibria are fast relative to the chemical steps, that chemical steps are irreversible, and that enzyme and substrate concentrations are constant, the steady-state rate of product formation is:

$$\frac{d[\text{tRNA-aa}]}{dt} = \frac{k_1 k_2 C D[\text{aa}][\text{ATP}]}{k_2 D + k_3 E} \tag{3.3}$$

where $C = K_{d,1}[\text{E}][\text{tRNA}]$, $D = \frac{K_{d,3}[\text{E}][\text{tRNA}]}{1+K_{d,3}[\text{E}][\text{tRNA}]}$, and $E = \frac{1}{1+K_{d,3}[\text{E}][\text{tRNA}]}$, and the constants correspond to the steps detailed in Figure 3-2. Thus the discrimination between substrates i and j (given by the ratio of the rates of product formation) is:

$$\Delta^{ij} = \frac{k_1^i k_2^i K_{d,1}^i K_{d,3}^i}{k_1^j k_2^j K_{d,1}^j K_{d,3}^j} \cdot \frac{(k_2^j K_{d,3}^j[\text{E}][\text{tRNA}] + k_3^j)}{(k_2^i K_{d,3}^i[\text{E}][\text{tRNA}] + k_3^i)} \tag{3.4}$$

When the free aa-AMP is never hydrolyzed ($k_3 \to 0$), this reduces to an expression dependent only lifetimes of non-cognate intermediates to the same scale as the dissociation rate may have an effect on specificity. Thus, in looking at the specificity of aminoacyl-tRNA synthetases, it is important to consider not only the binding of the substrates, but also the affinity of the aminoacyl-adenylate intermediate. An editing mechanism, in which the enzyme cleaves the amino acid moiety from non-cognate aminoacyl-tRNAs, adds an additional level of specificity in some systems [47], but this activity has not been observed in GlnRS.

As GlnRS must be highly selective, and as both its substrates and some of those against which it must discriminate are polar, it may be that GlnRS utilizes electrostatic discrimination to bind its cognate substrates specifically. This, coupled with the biological importance of this class of enzymes, makes this system a particularly interesting model in which to analyze the energetic role of electrostatics and other interactions in order to increase our understanding of how natural systems perform with high specificity.

3.2 Methods

Preparation of structures. The structure used for analysis of ATP bound to GlnRS is a 2.5 Å resolution structure of a ternary complex of GlnRS, tRNA^{Gln}, and ATP (Protein Data Bank (PDB) [125] ID 1gtr) [117]. The structure of 5′-*O*-[*N*-(L-glutaminyl) sulfamoyl] adenosine (QSI), an analogue of glutaminyl-adenylate, bound to GlnRS is a 2.4 Å resolution structure of a ternary complex of GlnRS, tRNA^{Gln}, and

---

on the relative affinities of the free amino acid and the rates of the first chemical step:

$$\Delta^{ij} = \frac{k_1^i K_{d,1}^i}{k_1^j K_{d,1}^j} \tag{3.5}$$

However, when all free aa-AMP is hydrolyzed (k_3 is large), the discrimination depends on the relative rates of all chemical steps and the affinities of both the free amino acids and the aminoacyl-adenylates:

$$\Delta^{ij} = \frac{k_1^i k_2^i K_{d,1}^i K_{d,3}^i k_3^j}{k_1^j k_2^j K_{d,1}^j K_{d,3}^j k_3^i} \tag{3.6}$$

QSI (PDB ID 1qtq) [124]. The tRNA was neglected in all calculations. Hydrogen-atom positions were determined using the HBUILD facility [14] of the CHARMM computer program [11]. The PARAM19 parameter set[11] was used for the protein atoms, with the addition of aromatic hydrogens on Phe, Tyr, Trp and His for consistency with the parameters used in the continuum electrostatic calculations. Parameters for the adenine base and the ribose of QSI, as well as for ATP, were taken from an experimental polar-hydrogen parameter set [163].

Partial atomic charges. Partial charges for the sulfamoyl and phosphodiester groups were obtained by restrained fitting to quantum mechanically derived electrostatic potentials for model compounds. The fitting was done using the RESP computer program [6, 28] and was based on electrostatic potentials obtained at the HF/6-311G** level of theory using GAUSSIAN94 [49]. Minimum energy geometries for the model compounds were determined at the HF/6-31G** level using the JAGUAR computer program [130]. The calculated charges (Figure 3-5) were slightly modified so as to be consistent with the partial charges used from other parameter sets (Figure 3-4).

Continuum electrostatic calculations. Electrostatic analysis was carried out with a continuum model. The protein and ligand were treated as regions of low dielectric with partial point charges placed at atomic centers embedded in high dielectric solvent with a bulk ionic strength of 0.145 M. In all calculations a dielectric constant of 80 was used for the solvent and, unless noted otherwise, a dielectric constant of 4 was used for the interior of all molecules. The solvent boundary was determined using a 1.4 Å radius probe, and an ion exclusion (Stern) layer [9] of 2.0 Å was applied. Numerical solutions of the linearized Poisson–Boltzmann equation were computed using a finite difference method, as implemented in a locally modified version of the DELPHI computer program [55, 57, 134, 136]. Protein partial atomic charges and radii were taken from the PARSE parameter set [140] with a few minor changes. Charges on the bridging ring carbons of tryptophan were assigned to $0e$, charges for proline and for disulfide bridged cysteine residues were taken from the PARAM19 parameter set [11], and the charges from glutamate and lysine side chains were used for charged C and N termini respectively. Partial charges for the adenine base and the ribose of QSI, and for ATP, were taken from an experimental polar-hydrogen parameter set [163]. The calculations were done on a $65 \times 65 \times 65$ unit grid using a four-step focusing procedure in which the terms involving only the region near the active site were calculated on a grid on which the longest dimension of the complex occupied 368% of one edge, terms involving regions further from the active site were calculated on a 184% fill grid, and terms involving the regions of the complex most distant from the active site were calculated on a 92% fill grid. Boundary potentials for each level were obtained from the previous focusing level, with those for the 92% fill obtained from a 23% fill calculation using Debye–Hückel boundary potentials. In all cases the finest resolution grid corresponded to a grid spacing of 0.37 Å. All calculations were averaged over ten translations of the structure on the grid in order to minimize artifacts from the the placement of the point charges and molecular boundaries onto the finite difference grid.

Calculations of the charge distributions which optimize the electrostatic binding free energy were performed as previous described [23, 77–80, 92–94] using locally written software. All matrices were well-behaved, with no negative or near-zero eigenvalues. Calculations of optimal charge distributions with constraints on the total charge were done using the LOQO computer program [133, 154, 155].

3.3 Results

3.3.1 Optimization of GlnRS ligands

The competitive inhibitor 5′-*O*-[*N*-(L-glutaminyl) sulfamoyl] adenosine (QSI) binds GlnRS moderately well (1.32 μM inhibition constant [124]). QSI is an analogue of the reaction intermediate L-glutaminyl adenylate in which the reactive phosphate group

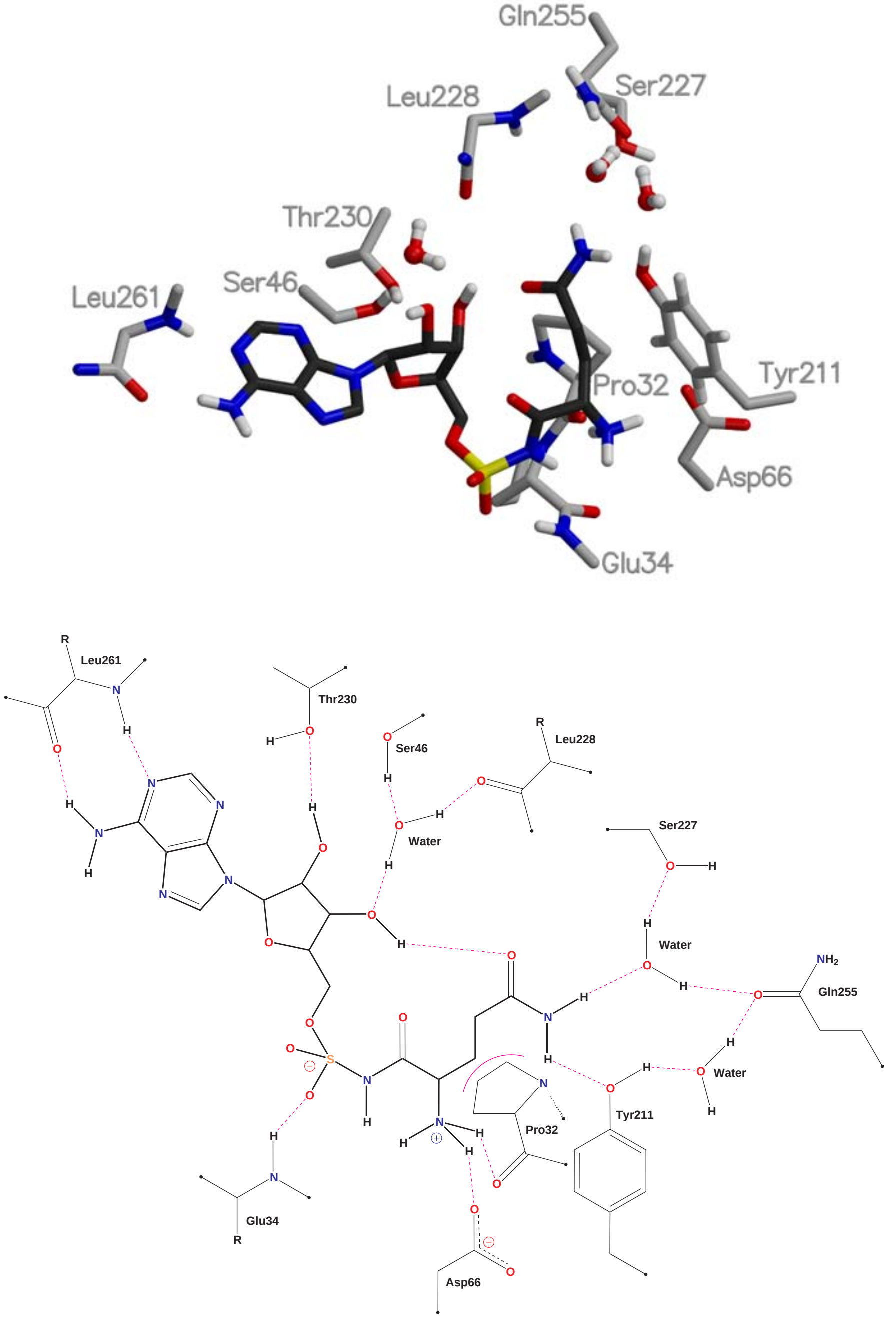


Figure 3-3: **Structure of QSI in the GlnRS active site.** The structure of QSI bound in the GlnRS active site is displayed, both as a detailed representation of atom positions from the crystal structure (QSI in dark grey and GlnRS residues in light grey) and as a schematic describing the key contacts. The structural figure was prepared with MOLSCRIPT [87] and RASTER3D [105].

is replaced by a sulfamoyl moiety. By means of comparison, the K_d for L-glutaminyl adenylate has been estimated at 20–30 nM under similar conditions[2], indicating that the analogue fails to mimic some important aspects of the intermediate effectively. The structure of QSI bound in the GlnRS active site is displayed in Figure 3-3. To examine the extent to which electrostatic complementarity is important in the binding of ligands, an analysis was performed on the structure of the QSI complex in which a hypothetical charge distribution for QSI was determined that is perfectly complementary to the enzyme. This so-called "optimized" QSI charge distribution would lead to the tightest binding to the enzyme [77, 80, 92, 93].

The partial charges on the atom centers of QSI that optimize the electrostatic binding free energy to GlnRS are displayed in Figure 3-4. Also shown are charges consistent with the corresponding parameterized partial atomic charges for nucleic acids and proteins. In many regions, the positive–negative patterning of the optimal charge distribution matches that of the actual ligand, and the magnitudes are also remarkably similar. This is most notable along the Watson–Crick base-pairing edge of the adenine base, at both the free amino group and the amide NH_2 group of the glutaminyl moiety, at the sulfamoyl group, and at the ribose hydroxyls. The greatest deviations between the optimal charges and those on QSI are around the five-membered ring of the adenine base, on the ring atoms of the ribose, along the carbon chain of glutamine, and at the backbone carbonyl of glutamine.

Because QSI substitutes the aminoacyl-adenylate intermediate's phosphodiester group by a sulfamoyl, it is useful to compare the optimal charges in this region to the

---

[2] The affinity of L-glutamine for GlnRS in the absence of tRNA has been determined to be 460 μM, and that of methyl adenylate has been determined to be 71 μM [8]. These two molecules have little overlap, but together include all components of the L-glutaminyl adenylate — the methyl group of methyl adenylate occupies the position of the glutamine carbonyl carbon of the intermediate, and one carboxylate oxygen of glutamine occupies the position of one adenylate phosphate oxygen. Due to the small overlap, it may be a reasonable first approximation that the free energy of binding of L-glutaminyl adenylate is simply the sum of the binding free energies of methyl adenylate and L-glutamine, making the dissociation constant for the intermediate the product of the dissociation constants of these two ligands, and giving an estimated K_d of 33 nM. In the presence of tRNA, the K_d of L-glutamine is 360 μM and that of methyl adenylate is 55 μM [8], giving an estimate for the K_d of L-glutaminyl adenylate of 20 nM under these conditions.

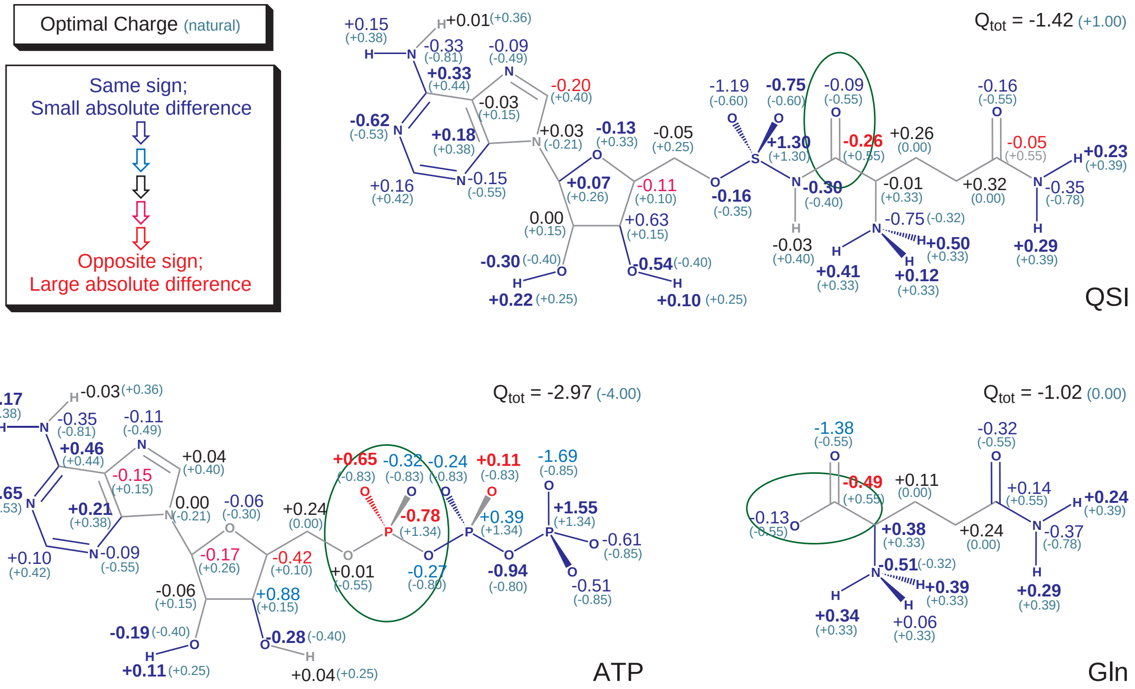


Figure 3-4: **Partial charges at atomic positions of QSI, ATP and glutamine that optimize binding to GlnRS.** Values in parentheses are partial atomic charges consistent with an experimental nucleic acid parameter set [163] and the PARSE protein parameter set [140]. Optimal charges are colored according to the degree to which they match the natural charges; similar charges in **blue** and substantially different charges in **red**.

actual charges of a phosphodiester. Optimal charges for the atoms of the sulfamoyl group of QSI along with partial atomic charges determined for model compounds containing sulfamoyl, phosphamoyl and phosphodiester linkages are shown in Figure 3-5. It is interesting to note that the optimal charges in this region more closely mimic those of the true intermediate than they do the charges of the intermediate analogue. The free oxygens of both the phosphamoyl and the phosphodiester are more highly charged than those of the sulfamoyl, as are the optimal charges in those positions. Also, the optimal charge at the sulfamoyl proton (which is absent in the phosphodiester) is near zero. These effects may contribute to the reduced affinity of QSI. The greater similarity between the optimum and the true intermediate than between the

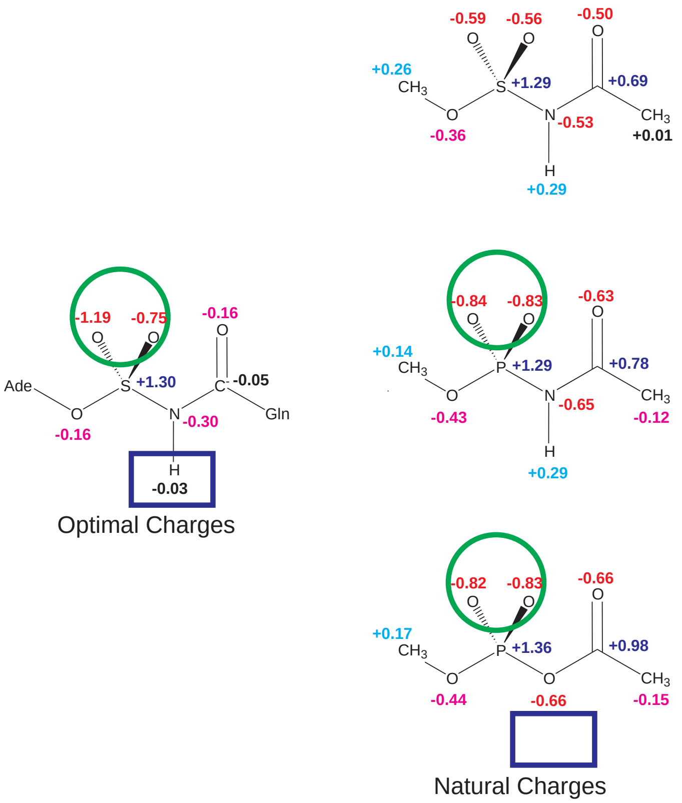


Figure 3-5: **Quantum mechanically derived partial atomic charges for sulfamoyl, phosphamoyl and phosphodiester model compounds.** Optimal charges at the sulfamoyl atoms of QSI are shown for equivalent positions. **Green circle:** The terminal oxygens have larger negative charges in the phosphorous containing compounds than in the sulfamoyl, consistent with the optimum. **Blue box:** The sulfamoyl proton has a near zero optimal charge, and is absent only in the phosphodiester.

optimum and the intermediate analogue is also apparent in the net charges on the ligands. The net charge of QSI, with a neutral sulfamoyl group, is $+1e$, while that of glutaminyl-adenylate is $0e$ due to the negative charge on the phosphate. The net charge of the optimal ligand is $-1.42e$, closer in total charge to glutaminyl-adenylate than to QSI. This is even more apparent focusing on the sulfamoyl group; the sum of the charges on these atoms in the optimal ligand is $-1.43e$, compared with $0e$ for the sulfamoyl and $-1e$ for the phosphodiester. Energetically, the greater similarity of the phosphorous containing compounds is very pronounced (see Table 3-1). The

	Natural Charges		Alternate Charges		
Ligand	ΔG_{es}	Q_{tot}	Charge Set	$\Delta\Delta G_{es}$	Q_{tot}
ATP	+12.59	-4	Optimum	−12.73	-2.97
Gln	+3.02	0	Optimum	−3.12	-1.02
			Glu	+13.83	-1
			Glu-H (*anti*)	+1.70	0
			Glu-H (*syn*)	+1.05	0
QSI	+22.56	+1	Optimum	−23.35	-1.42
			QPI	−12.71	0
			Gln-AMP	−14.18	0
			ESI	+13.66	0
			ESI-H (*anti*)	+1.55	+1
			ESI-H (*syn*)	+0.89	+1

Table 3-1: **Electrostatic free energies of binding of GlnRS ligands.** The electrostatic binding free energy of each ligand is shown along with the total ligand charge. Differences in binding free energy relative to the natural ligand, as well as total charges, are also shown for alternate charge distributions, including the optimum. QSI and QPI are the sulfamoyl and phosphamoyl analogues of glutaminyl adenylate, ESI is the glutamate analogue of QSI, and the -H designates a protonated glutamic acid, with the proton in either the *syn* or *anti* orientation.

electrostatic binding free energy of the phosphamoyl equivalent of QSI is computed to be 12.7 kcal·mol^{-1} more favorable than that of QSI, and glutaminyl adenylate is computed to have a 14.2 kcal·mol^{-1} more favorable binding free energy. Thus, while the QSI is more than 23 kcal·mol^{-1} suboptimal, the electrostatic binding free energy of glutaminyl adenylate is less than 10 kcal·mol^{-1} worse than the optimum.

ATP is one of two substrates in the first reaction catalyzed by GlnRS, and thus the optimal partial charges at the atomic centers of ATP were similarly computed and analyzed (see Figure 3-4). As for QSI, the optimal charges on the Watson–Crick base-pairing edge of the adenine base show remarkable agreement to the natural charges, and in fact are almost identical to those of the QSI optimum. The structures of both ATP (see Figure 3-6) and QSI (see Figure 3-3) bound to GlnRS are very similar in the overlapping region, with all the same contacts between the protein and both the base and ribose seen in both structures, and very similar conformations adopted by both

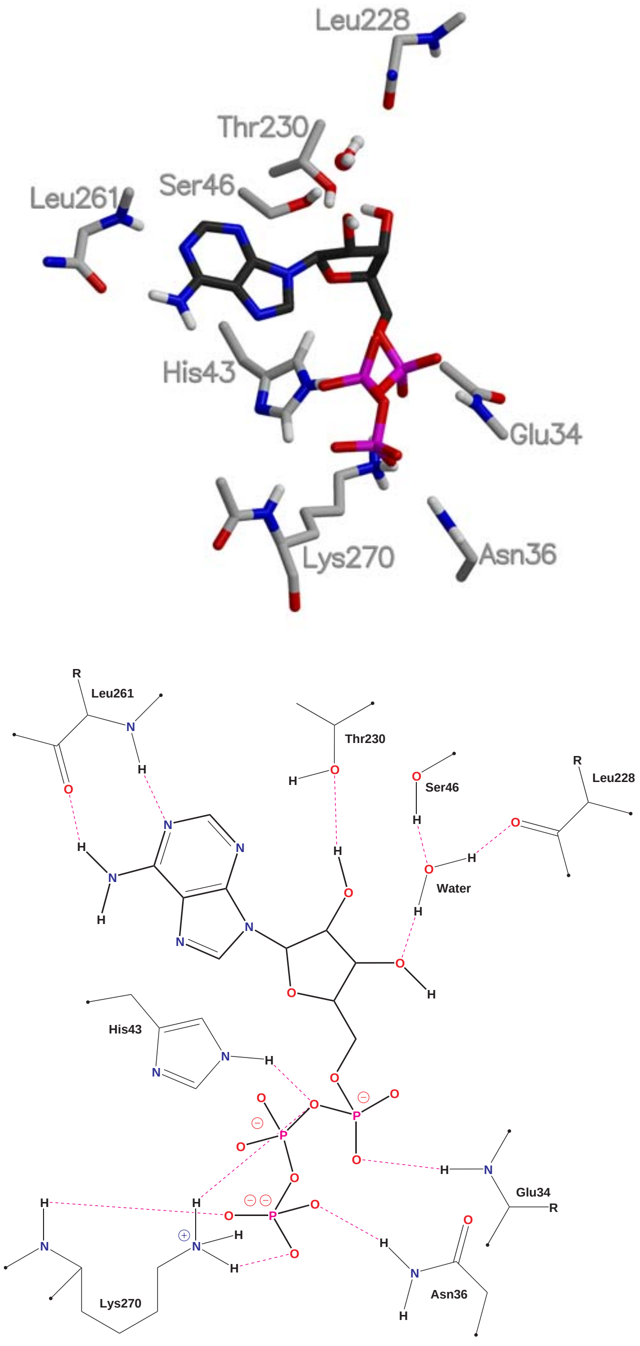


Figure 3-6: **Structure of ATP in the GlnRS active site.** The structure of ATP bound in the GlnRS active site is displayed, both as a detailed representation of atom positions from the crystal structure (ATP in dark grey and GlnRS residues in light grey) and as a schematic describing the key contacts. The structural figure was prepared with MOLSCRIPT [87] and RASTER3D [105].

the protein side chains and the adenosine. In addition, the optimal charges on the γ-phosphate are very similar to the natural charges at this position. A lesser degree of similarity is seen on the five-membered ring of the adenine base, on the ribose ring atoms, and on the β-phosphate. The α-phosphate, on the other hand, shows large differences between the optimal and natural charges, with the central phosphorous and one of the terminal oxygens reversing sign.

A model of glutamine (the second substrate of the first reaction) bound in the active site was generated using the position of the glutaminyl portion of QSI as a guide and the optimal charges were computed (see Figure 3-4). Again, the optimal and natural charges are very similar in many regions, particularly at the side-chain amide NH_2 and the backbone ammonium, as was seen in the QSI optimum. The greatest deviations in this case are seen at the backbone carboxylate, which in the optimum is much more negative than in the natural ligand.

3.3.2 Robustness of results

One question which arises about the optimal charge distribution is the effect of constraints on the total charge of the system. Real molecules are constrained to have integral net charges, while the total charges of the calculated optimal charge distributions are non-integral. In addition, while the optimal net charges obtained certainly fall within the range of the charges of naturally occurring ligands, they deviate from the total charges of the natural ligands by up to more than $2e$. To investigate the effect of such constraints, the optimal charges at the atom centers of the ligands were determined under the constraint that the total charge be that of the natural ligand. In the case of ATP the constraint was that the total charge be $-4e$ (a $1.03e$ difference from the free optimum), for glutamine the total charge was fixed at $+1e$ (a $1.02e$ difference), and for QSI the total charge was constrained both to $0e$ (the net charge of glutaminyl-adenylate, and a $1.42e$ difference) and $+1e$ (the net charge of QSI, and a $2.42e$ difference). The variations of the optimal charges obtained with constraints

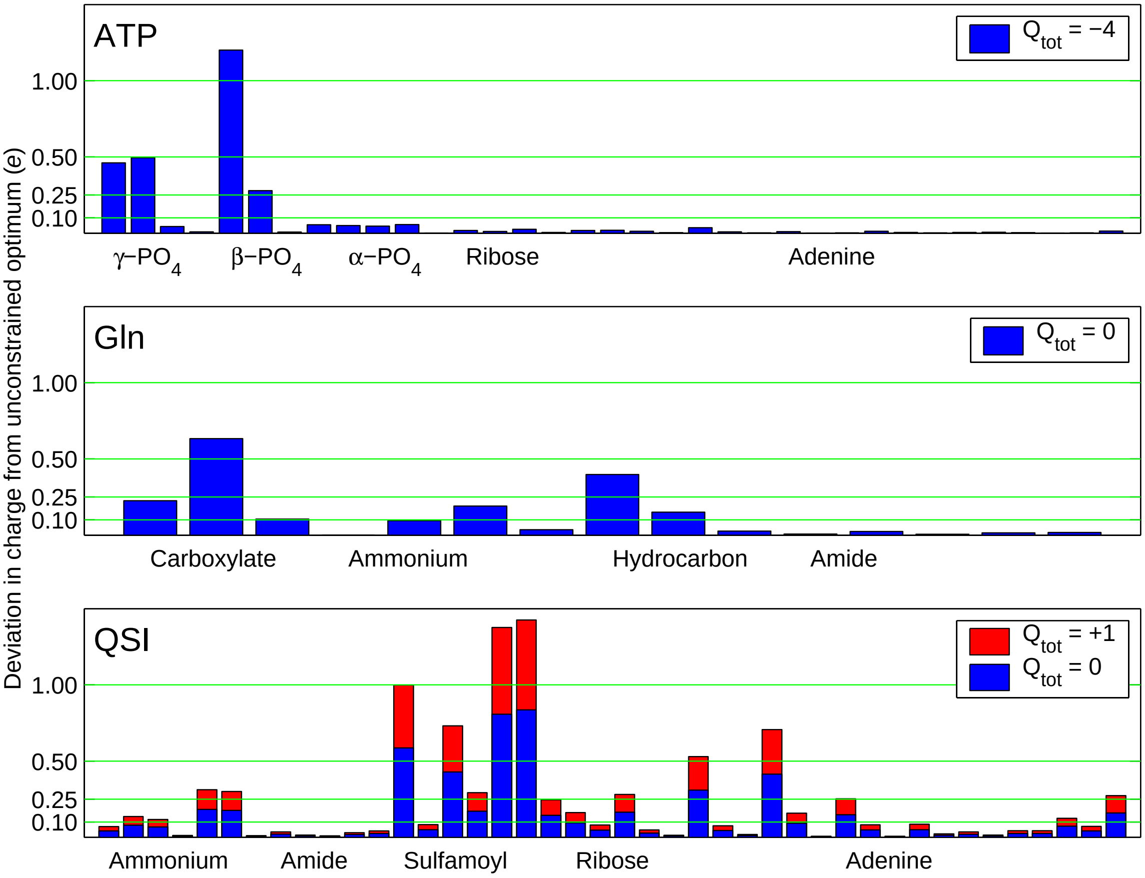


Figure 3-7: **Variation of optimal charges at atomic positions of GlnRS ligands with constraints on total ligand charge.** The deviation of the constrained optimal charges of each GlnRS ligand from the unconstrained optimal charges are displayed, grouped by the regions of each molecule. In all cases, the majority of the charges vary by less than $0.2e$ and many by less than $0.1e$. The greatest variations are localized to specific regions of each molecule.

from those of the unconstrained optima are displayed in Figure 3-7.

For the atom centers of ATP, almost all variation is localized to four atoms from the γ- and β-phosphates, with all other variations being less than 0.1 charge units. Those positions which vary the most are those which are the least buried on binding, as measured by the inverse of the diagonal element of the desolvation matrix, and thus may be expected to contribute the least to the binding free energy. Energetically, the imposition of this constraint costs very little, with the constrained optimum binding only 0.63 kcal·mol^{-1} worse than the free optimum.

For glutamine, the greatest variation is seen in the atoms of the carboxylate and in the hydrophobic portion of the side chain. Virtually identical charges are seen at the amide atoms in both the constrained and the free optimum. This constraint, a similar in magnitude deviation in total charge as in the case of ATP, is slightly more costly in an energetic sense, with the constrained optimum binding 1.63 kcal·mol^{-1} worse than the free optimum.

For the atom centers of QSI, the largest variations occur at the atom positions of the sulfamoyl which, as in the case of the ATP phosphates and the glutamine carboxylate, are the most solvent exposed in the complex. However, in this case there are more significant changes elsewhere, particularly at several ribose atom positions. Still, the optimal charges in some of the regions which showed the best agreement with natural charges (the adenine base, particularly on the Watson–Crick base-pairing edge, and the glutamine amide) show very little variation, even when the total charge is constrained to $+1e$, more than 2.0 charge units away from optimum. Comparing the variation of the charges in the neutral optimum and the positively charged optimum, the same positions are seen to vary the most relative to the free optimum (net charge $-1.02e$) — the deviation is simply larger for the positively charged optimum. The neutral optimum binds 3.22 kcal·mol^{-1} worse than the free optimum, with a slightly larger variation in total charge ($1.42e$) than in the case of ATP and glutamine, and the positively charged optimum binds 9.36 kcal·mol^{-1} worse for a variation in charge of $2.42e$ — the energetic cost of the constraint increases as the square of the total difference in charge from the unconstrained optimum as a result of the quadratic nature of the energy surface.

The results of continuum electrostatic calculations can be dependent on the value of the dielectric constant used for the interior of the protein and ligand. To determine the effect of this value on these results, the optimal charges at the atomic positions of QSI were re-calculated using a number of internal dielectric constants ranging from

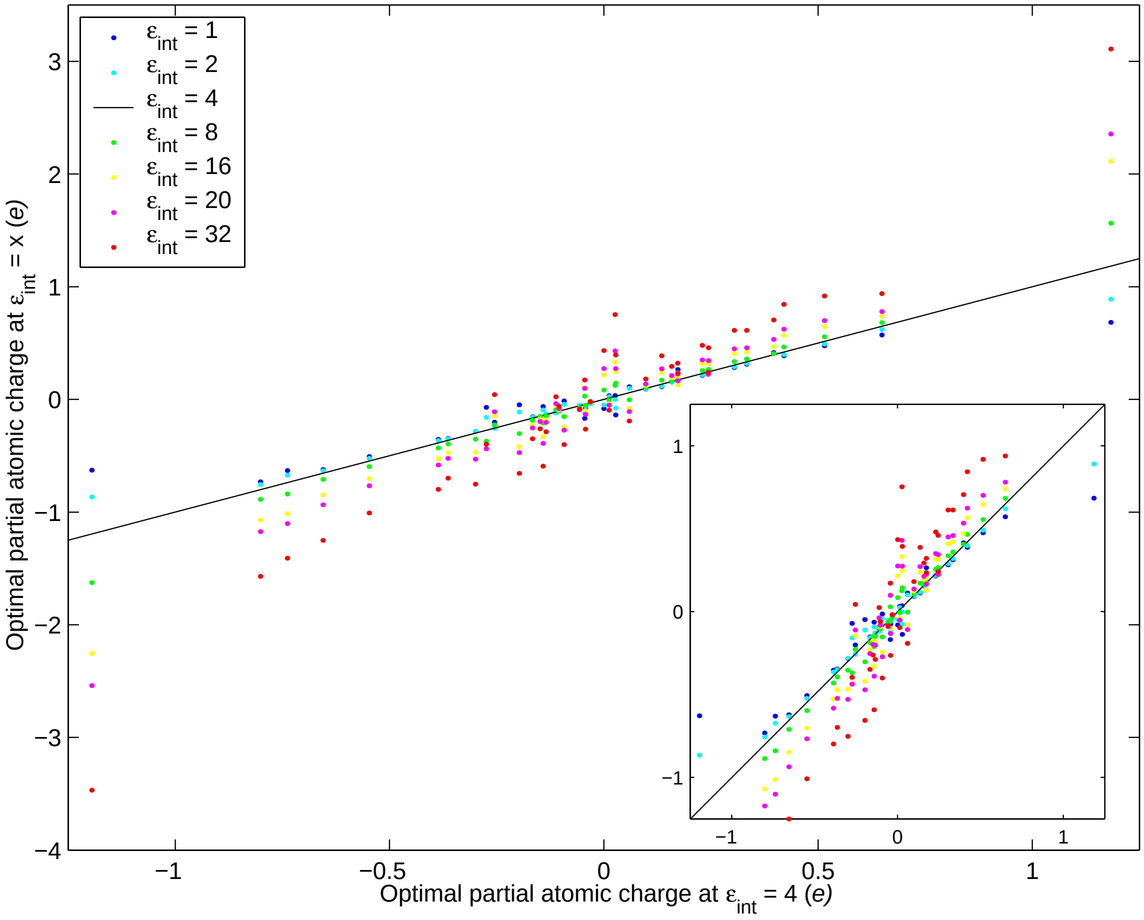


Figure 3-8: **Variation of optimal charges at atomic positions of QSI with internal dielectric constant.** The optimal charges calculated at QSI atom positions with a variety of internal dielectric constants are plotted relative to those computed with $\epsilon_{int} = 4.0$. The inset shows the same data in the range -1.2 to $+1.2e$ with both axes on the same scale. With the exception of the most negative and the most positive charges, the optimal charges vary little when ϵ_{int} is changed.

1.0 to 32.0. The variation of the resultant optimal charges with respect to those obtained using an internal dielectric constant of 4.0 is displayed in Figure 3-8.

For internal dielectric constants between 1.0 and 8.0, the optimal point charges vary only very slightly, with the majority of the variation localized to the two partial charges largest in magnitude. These correspond to the position of the sulfamoyl sulfur, which takes on a large positive charge in the optimum, and one of the univalent oxygens attached to the sulfur, which in the optimum is highly negative. These two atoms are less desolvated on binding than any others in the system, and thus variations in the charges of these atoms results in the smallest energetic cost. These optimal partial charges vary the most with the choice of internal dielectric, and do so in the expected manner. As the internal dielectric is increased, charges must be larger to effect a similar interaction, while at the same time, the desolvation penalty for larger charges is reduced as the difference between the external and internal dielectric constants becomes less pronounced. This effect can also be seen in the net optimal charge of the ligand, which increases with increasing internal dielectric constant.

3.3.3 Specificity of glutamine over glutamic acid

The primary role of GlnRS is to faithfully link glutamine with tRNA^{Gln} and to avoid reaction with similarly structured amino acids such as glutamic acid. The results of the charge optimization procedure show that the amino portion of the glutamine amide has natural charges very close to optimal; the best ligand for this enzyme has an NH_2 group, immediately demonstrating a preference for glutamine even over the protonated form of glutamic acid.

In addition to providing details of optimal charges, the affinity optimization procedure provides a fast method for evaluating the electrostatic contribution of any charge distribution on a given ligand geometry. As glutamine and glutamic acid are very close in structure, the electrostatic binding free energy of glutamic acid, both in its charged and neutral states, was estimated using the atomic positions of glutamine but the charges of glutamic acid (absent hydrogen atoms being assigned a charge of zero). Two protonated conformations were considered, with the proton either in the *syn* or *anti* position, on the oxygen equivalent to the amide nitrogen. The comparison was done both in the context of the free amino acid and in the context of the sulfamoyl analogue of the adenylate intermediate.

The results (Table 3-1) show that GlnRS strongly discriminates against the negatively charged state of glutamate, with glutamate computed to bind almost 14 kcal·mol^{-1} worse than glutamine both in the context of the free amino acid and in the context of the sulfamoyl inhibitor. The protonated form of glutamic acid is less strongly discriminated against, with only a roughly 1 kcal·mol^{-1} loss of binding free energy (which would result in less than a ten-fold difference in binding affinity) in going from glutamine to glutamic acid with the proton in either conformation. However, as the pK$_\mathrm{a}$ of glutamic acid is roughly 4, an additional approximately 4 kcal·mol^{-1} is required for binding the protonated form at pH 7, resulting in an overall computed difference in affinity of 5000-fold[3].

---

[3] The general expression describing the dissociation of an acid, $\mathrm{K_a} = \frac{[\mathrm{H^+}][\mathrm{A^-}]}{[\mathrm{HA}]}$, can be rearranged to give the concentration of the protonated form, [HA], in terms of the total concentration of acid, $[\mathrm{A^{tot}}]$, as $[\mathrm{HA}] = \frac{[\mathrm{H^+}][\mathrm{A^{tot}}]}{\mathrm{K_a}+[\mathrm{H^+}]}$. If only the protonated form can bind to a given receptor, R, the dissociation constant of the complex can then written as $\mathrm{K_d} = \frac{[\mathrm{HA}][\mathrm{R}]}{[\mathrm{HA.R}]}$. Substituting in the expression for [HA] gives:

$$\mathrm{K_d} = \frac{[\mathrm{H^+}]}{\mathrm{K_a} + [\mathrm{H^+}]} \cdot \frac{[\mathrm{A^{tot}}][\mathrm{R}]}{[\mathrm{HA.R}]} \tag{3.7}$$

The second term is the apparent complex dissociation constant for the acid, $\mathrm{K_d^{\mathit{app}}}$, describing the relative concentrations of free receptor, of free acid (in both protonated and unprotonated forms) and of the complex. Converting Equation 3.7 to give dissociation free energy, by $\Delta\mathrm{G} = -RT\ln\mathrm{K}$, yields:

$$\Delta\mathrm{G} = -RT\ln\frac{[\mathrm{H^+}]}{\mathrm{K_a} + [\mathrm{H^+}]} - RT\ln\mathrm{K_d^{\mathit{app}}} \tag{3.8}$$

which rearranges to:

$$\Delta\mathrm{G}^{app} = \Delta\mathrm{G} + RT\ln\frac{[\mathrm{H^+}]}{\mathrm{K_a} + [\mathrm{H^+}]} \tag{3.9}$$

with $\Delta\mathrm{G}^{app}$ being the apparent dissociation free energy of the acid–receptor complex. Under conditions where the proton concentration is significantly lower than the K$_\mathrm{a}$, as is the case here, with a pH of 7 and a pK$_a$ of 4, the term of $\frac{[\mathrm{H^+}]}{\mathrm{K_a}+[\mathrm{H^+}]}$ can be approximately by $\frac{[\mathrm{H^+}]}{\mathrm{K_a}}$, and thus Equation 3.9 can be written as:

$$\Delta\mathrm{G}^{app} = \Delta\mathrm{G} + \frac{RT}{\log_{10} e}(\mathrm{pK_a} - \mathrm{pH}) \tag{3.10}$$

where the substitutions of $\mathrm{pH} = -\log[\mathrm{H^+}]$ and $\mathrm{pK_a} = -\log\mathrm{K_a}$ have been made. The dissociation free energy is lowered in proportion to the difference in pK$_\mathrm{a}$ and pH by a factor of $\frac{RT}{\log_{10} e} = 1.36$ kcal·mol^{-1}. Thus a three unit difference in pK$_\mathrm{a}$ and pH reduces the dissociation free energy (or increases the binding free energy) by 4.1 kcal·mol^{-1}, yielding a 1000-fold reduction in affinity.

3.4 Discussion

The most striking aspect of this work is the remarkable similarity between the optimal charges and those of the actual ligands, particularly in areas where the ligands can be seen to make interactions with the enzyme in the structures of the complexes. It thus seems that, at least in some cases, nature does optimize electrostatic interactions; the binding site is constructed such that its optimal ligand is electrostatically similar to the desired ligand. The minor differences are not surprising as the electrostatic interactions possible in a binding site are limited by the somewhat small set of polar functionalities which exist in the twenty naturally occurring amino acids.

For each ligand, the greatest deviations between the optimal and natural charges are localized in one region of the molecule. For ATP this is the α-phosphate, for glutamine it is the carboxylate, and for QSI (or Gln-AMP) it is the backbone carbonyl of the glutamine moiety. These regions have two common features: they are largely exposed to solvent in the bound state, and they are the sites at which the chemistry of the reactions takes place. Both these features may contribute to the deviations observed.

First of all, the affinity optimization procedure involves balancing the unfavorable desolvation penalty of binding with favorable electrostatic interactions made in the bound state. When a portion of a molecule is largely exposed to solvent in the bound state, the desolvation of this region is small, and thus large partial atomic charges are less unfavorable in these areas than in areas which are more substantially buried in the bound state. Energetically, the cost of the deviation of a partial atomic charge from the optimum is inversely related to the degree of burial (as measured by the change in the desolvation potential of a charge located at that point upon binding), and thus atomic centers with small desolvation potentials can deviate significantly from optimal without paying a large energetic penalty. Therefore, there is less pressure for solvated regions to have a close match between the natural and optimal charges.

While the solvent exposure may partially explain the difference between the optimal and natural charges at these positions, other regions in the ligands have similar solvent exposure in the bound state, yet have optimal and natural charges which agree much more closely (the γ-phosphate of ATP and the sulfur center of QSI are particularly good examples of this). However, it may not be expected that the enzyme would have evolved to bind tightly to regions which are involved directly in the chemistry of the catalyzed reactions. Strong interactions between the enzyme and these areas may lead to reduced mobility, which could hinder chemistry and thus reduce the catalytic efficiency of the enzyme. In the bimolecular reactions catalyzed by GlnRS, one likely mode of catalysis is by the enzyme binding the substrates in an orientation favorable for reaction, while leaving the atoms directly involved in the reaction free to move. The close agreement between the optimal and natural charges at the ends of the ligands, somewhat removed from the site of chemistry, supports such a mechanism. For the first chemical step, the enzyme has evolved to bind tightly to the adenine base and to the γ-phosphate of ATP, positioning the α-phosphate in a good location for reaction with the carboxylate of glutamine, which is similarly positioned by interactions between the enzyme and both the ammonium and the side-chain amide. In the second chemical step, interactions between the adenine base, the glutamine amide, and the ammonium position the carbonyl of the glutaminyl-adenylate in just the right orientation for reaction with tRNA. In addition, the enzyme may have evolved to make favorable interactions with the transition state more so than with the substrates, and thus some of the deviations seen may be due to the differences in geometry and charge distribution between the transition states of each reaction and the bound substrates.

The results of the affinity optimization procedure are quite robust to variations in the calculations. Equivalent regions in different ligands, such as the adenine base in both ATP and QSI, and the glutamine side chain and ammonium in both glutamine and QSI, have similar optimal charges. Thus the results of the optimization in these areas are generally independent of the global shape of the molecule, and are localized from the portions of the molecules which vary during the chemical reactions. The choice of internal dielectric constant has very little effect on the optimal partial atomic charges on QSI outside the most solvent exposed region of the ligand, with the vast majority of the variation localized to two atoms. In addition, in no case did the qualitative distribution of positive and negative charges or the relative ordering of charge magnitudes change significantly under different conditions. Similar results are seen when reasonable constraints are applied to the system; the computed partial atomic charges over most regions of the molecules vary little even when the total charge is constrained to a value differing from the unconstrained optimum by more than $2e$.

It is important to note that large regions of a molecule can be very close to the optimum even when the electrostatic binding free energy is significantly unfavorable. For example, QSI is computed to have an electrostatic contribution to the binding free energy of $+22.6$ kcal·mol^{-1}, yet has close agreement between optimal and natural charges in many areas. This binding free energy can be greatly improved by only slightly varying a small region of the molecule — the phosphamoyl analogue of QSI (a single atom difference) has a electrostatic binding free energy 12.7 kcal·mol^{-1} more favorable than QSI, and the glutaminyl-adenylate has an electrostatic binding free energy 14.2 kcal·mol^{-1} more favorable. That even as closely matching charge distributions as are seen here between the natural and optimal charges can have such significantly different binding free energies simply shows the importance of electrostatic interactions and their optimization. The electrostatic binding free energy varies quadratically with the ligand charge distribution, and thus deviations in charge which move away from the optimal charges have an amplified energetic effect. As a result, the change of the sulfamoyl group to a phosphamoyl or a phosphodiester, which brings the ligand closer to optimal by bridging $1.0e$ of a difference in total charge of $2.4e$, makes up more than half of the difference in binding free energy between the ligand and the optimum, despite only acting on a few atoms. While an experimental value for the binding affinity of glutaminyl-adenylate is not available, the binding affinity of tyrosyl-adenylate to TyrRS is 13.2 pM and that of phenylalanyl-adenylate to PheRS is 4.4 nM. Simply combining the known affinities of glutamine (460 μM) and methyl-adenylate (71 μM) [8] gives an estimated binding affinity of glutaminyl-adenylate of 20–30 nM, and thus it seems reasonable to consider that the affinity of glutaminyl-adenylate to GlnRS is similar to that of the other aminoacyl-adenylates to their cognate aaRS, in the range of picomolar to low nanomolar. On the other hand, in an inhibition assay, QSI has a $\mathrm{K_i}$ of 1.32 μM, and thus seems to bind much more weakly. This is in good qualitative agreement with the computed results.

3.5 Conclusions

The partial atomic charges which optimize the electrostatic contribution to binding of several ligands to glutaminyl-tRNA synthetase from *E. coli* were determined, and were compared with the natural charges of these ligands. Remarkable agreement is seen between the optimal and natural charges in many regions, suggesting that the enzyme has evolved to optimize many of the electrostatic interactions it makes with its ligands. The optimization results also indicate that analogues of the glutaminyl-adenylate which preserve the phosphorous center (and which thus preserve the net negative charge of this region) are likely to be more effective inhibitors than those containing electrostatically neutral sulfur centers. The results are seen to be quite robust to changes in the details of the computation, making it clear that the observed behaviors are a result of the nature of the enzyme and the mode of binding, and are not artifacts of the theoretical procedure.

An interesting question which arises out of this work is that of the meaning of the regions where the optimal charges deviate significantly from those of the natural ligand. The most substantial differences were seen between the optimal and natural charges at the sites at which the chemistry of the activation and the aminoacyl transfer reactions occur, and thus it is very possible that in these regions the enzyme has not optimized binding to the substrate, rather evolving either to bind more preferentially to the transition state or to allow for the required mobility of atoms involved in the chemical reaction. Further work, such as performing a similar analysis on transition state charge distributions and geometries, may help answer some of these questions.

Chapter 4

Design Methods for Peptide Inhibitors: Optimization of HIV-1 Cell-Entry Inhibitors Targeting the N-Terminal Coiled Coil of gp41

Abstract

HIV infection of a cell requires that the viral membrane is able to fuse with that of the target cell. This membrane fusion event is mediated by the viral membrane glycoprotein gp41, which is thought to undergo a conformational change involving the docking of three helices (from the C-terminal portion of gp41) against a preformed trimeric coiled coil (from the N-terminal portion of gp41) , as a prerequisite for membrane fusion. Molecules that bind to the trimeric coiled coil have been shown to block the conformational change, making them effective inhibitors of the infection of cells by HIV. These include a short, cyclic D-peptide identified by mirror-image phage display which binds in a relatively hydrophobic pocket on the coiled coil. The structure of a complex of this D-peptide and a model of the coiled coil has been solved to atomic resolution.

Here calculational approaches were applied to analyze the crystal structure in search of defects in either packing or electrostatics that could be exploited in the design of enhanced affinity ligands, potentially utilizing amino acids beyond the standard twenty. Areas of small electrostatic non-complementarity involving two key

tryptophan side chains were identified. To search for modified ligands with enhanced affinity, a procedure based on the electrostatic optimization framework was developed, in which a large database of tryptophan derivatives were computationally screened for enhanced binding of the D-peptide to the target coiled coil. Using a hierarchical procedure in which increasingly accurate, but more costly, calculations are done on a ranked subset of molecules identified by a more cost-effective procedure, enabled a library of over 9000 D-peptide derivatives to be screened. While the computed improvement in binding free energy of the top ranked ligand was only 0.9 kcal·mol^{-1} better than the original D-peptide, the procedure was validated as being a useful tool in the design of improved inhibitors.

4.1 Introduction

4.1.1 Inhibition of HIV-1 cell entry

Human Immunodeficiency Virus (HIV) is a membrane enveloped virus, and therefore, to infect a cell, the viral membrane must fuse with that of the target cell. This membrane fusion event is facilitated by gp41 and gp120, two HIV membrane glycoproteins. gp41 and gp120 are synthesized as a single polypeptide (gp160), and then proteolytically cleaved into the functional subunits after folding into their native states. The C-terminal region of gp41 spans the viral membrane, while gp120 is bound to gp41 on the viral surface. Membrane fusion takes place through a series of steps. First gp120 binds to a receptor, CD4, on the target cell, and to one of several chemokine co-receptors. On binding, gp120 undergoes a conformational change and may dissociate from gp41, leaving gp41 in a transient intermediate state. A fusion peptide at the N-terminus of gp41 then inserts into the target membrane, with gp41 still in the transient conformation. A major conformational change is then thought to take place, involving the docking of a region of gp41 proximal to the viral membrane against a pre-formed trimeric coiled coil near the N-terminus. This docking, involving three chains of gp41, ultimately results in the formation of a fusogenic "trimer-of-hairpins" structure, with a six-helial bundle as the primary structural element, in which the viral and cellular membranes are in close proximity. At this point the cellular and

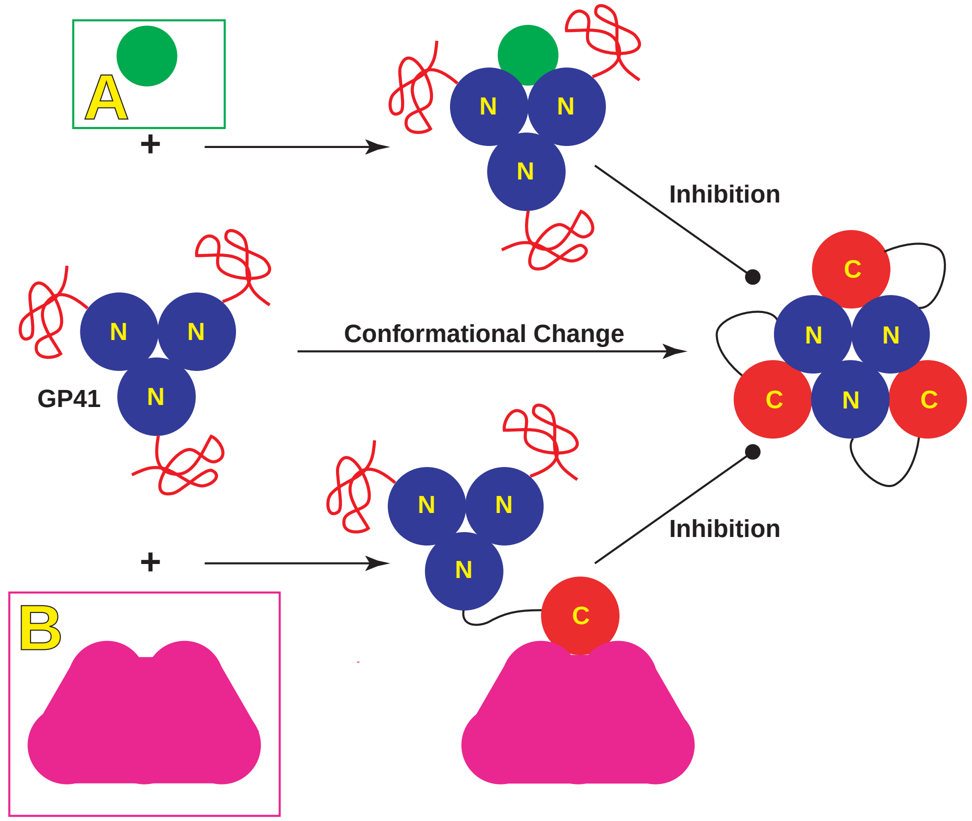


Figure 4-1: **Inhibition of the gp41 conformational change.** There are two general mechanisms by which the formation of the gp41 fusion active conformation can be inhibited. **A:** A ligand binding to the N-terminal coiled coil of gp41 in its transient non-fusogenic conformation would block the binding site occupied by the C-terminal helix in the fusogenic form. **B:** A ligand that binds to the C-terminal helix would similarly prevent the helix from docking against the N-terminal coiled coil.

viral membranes fuse, allowing the viral contents to enter the cell. This mechanism and the data leading to its elucidation has recently been reviewed by Eckert and Kim [42].

The conformational change of gp41 provides a excellent target for the development of inhibitors of HIV cell entry, since the helical regions of gp41 forming the six-helical bundle structure are highly evolutionarily conserved relative to other portions of the sequence (see below for details). Two possible modes of inhibition are readily proposed, both targeting the transient state of gp41 which exists after gp120 binds to CD4, but before the conformational change that forms the trimer-of-hairpins structure. First, a molecule that binds to the site on the N-terminal coiled coil against which the C-terminal peptides dock would block the binding of the C-terminal domain and thus the conformational change. Peptides from the C-terminal region are active inhibitors by this mechanism [20, 75, 101], and several other inhibitors targeting this site have been developed. These include a short D-peptide isolated by phage display [43] and several small organic molecules found through various screening methodologies [34, 46, 71, 167]. A second mode of inhibition targets the C-terminal region. A molecule which binds to this portion of gp41 would similarly prevent the requisite docking event for the conformational change. Acting through this mechanism, peptides from the N-terminal coiled-coil region of gp41 are also weak inhibitors in membrane fusion assays, trimerizing and then binding to the C-terminal region [98, 162]. Several protein constructs designed around the N-terminal coiled coil, including 5-Helix, which is the focus of Chapter 5, similarly bind to the C-terminal region, sequestering it away from the N-terminal coiled coil [41, 97, 127]. These constructs are all potent inhibitors of HIV Type-1 (HIV-1) viral–cell membrane fusion. A schematic summarizing the conformational change and the targets of inhibition is shown in Figure 4-1.

The sequence of the envelope glycoprotein gp160, the precursor of both gp41 and gp120, is quite variable across HIV-1 isolates. Over thirty isolates, the overall length of gp160 varies between 847 and 867 residues, and of these, only 370 of these are strictly conserved (43%). Furthermore, only 529 are at least moderately conserved (62%), and only 582 are even weakly conserved (68%)[1]. Comparatively, the sequence

---

[1] A fairly strict definition of conservation is used. A position is considered moderately conserved if only a few variations in amino acid identity are seen, and if the only variations seen are between residues with similar physico-chemical properties. A position is considered weakly conserved if slightly more variation is seen, both in terms of the number of amino acid identities observed at the site, or in terms of the degree of similarity of the residue types occupying the site. Even if a single variation to a substantially different residue type is seen (*e.g.* a single methionine at a position otherwise occupied by lysine), the position is considered unconserved.

	C34 Sequence	N36 Sequence
HV1-		
Z2	WMEWEREIDNYTGLIYRLIEESQTQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARIL
Z6	WMEWEREIDNYTGLIYRLIEESQTQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARIL
EL	WMEWEREIDNYTGLIYSLIEESQTQQEKNEKELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARIL
ND	WMEWEREIDNYTGLIYSLIEESQIQQEKNEKELL	SGIVHQQNNLLRAIEAQQHLLQLTVWGIKQLQARVL
Z8	WIEWEREIDNYTGVIYSLIENSQIQQEKNEQDLL	SGIVQQQNNLLRAIEAQQHMLQLTVWGIKQLQARVL
MA	WMQWEKEISNYTGIIYNLIEESQIQQEKNEKELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARVL
W1	WMEWEREIDNYTSLIYNLIEESQNQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARVL
W2	WMEWEREIDNYTSIIYSLIEESQNQQGKNEQELL	SGIVQQQNNLLRAIDAQQHLLQLTVWGIKQLQARVL
C4	WMEWDREIDNYTHLIYTLIEESQNQQEKNQQELL	SGIVQQQNNLLRAIKAQQHLLQLTVWGIKQLQARIL
A2	WMQWEREIDNYTNTIYTLLEESQNQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARVL
S1	WMEWEREIDNYTNLIYTLIEESQNQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARVL
B1	WMEWDREINNYTSLIHSLIEESQNQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARIL
PV	WMEWDREINNYTSLIHSLIEESQNQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARIL
H2	WMEWDREINNYTSLIHSLIEESQNQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARIL
H3	WMEWDREINNYTSLIHSLIEESQNQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARIL
B8	WMEWDREINNYTSLIHSLIEESQNQQEKNEQELL	SGIVQQQNNLLRAIEGQQHLLQLTVWGIKQLQARIL
LW	WMEWDREINNYTSLIHSLIEESQNQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARIL
BR	WMEWDREINNYTSLIHSLIEESQNQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARIL
MF	WMEWDREINNYTSLIHSLIDESQNQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARIL
J3	WMEWEREIDNYTSLIYTLIEESQNQQEKNEQELL	SGIVQQQNNLLRAIEGQQHLLQLTVWGIKQLQARIL
SC	WMEWEREIDNYTSLIYTLIEESQNQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARVL
JR	WMEWEKEIENYTNTIYTLIEESQIQQEKNEQELL	SGIVQQQNNLLRAIEAQQHMLQLTVWGIKQLQARVL
BN	WMEWEREIDNYTNLIYSLIEDSQIQQEKNEKELL	SGIVQQQNNLLMAIEAQQHMLELTVWGIKQLQARVL
MN	WMQWEREIDNYTSLIYSLLEKSQTQQEKNEQELL	SGIVQQQNNLLRAIEAQQHMLQLTVWGIKQLQARVL
KB	WMEWEREINNYTNLIYNLIEESQNQQEKNEQDLL	PGIVQQQNNLLRAIDAQQHLLQLTVWGIKQLQARVL
OY	WMQWEREIDNYTHLIYTLIEESQNQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARVL
Y2	WMKWEREIDNYTHIIYSLIEQSQNQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARVL
RH	WMQWEREIDNYTGIIYNLLEESQNQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARVL
S3	WMEWEREIDNYTSLIYTLLEESQNQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARVL
ZH	WLEWDKEVSNYTQVIYNLIEESQTQQEINERDLL	SGIVQQQNNLLRAIEAQQHLLKLTVWGIKQLQARIL
Consensus		
	WMEWEREIDNYT-LIY-LIEESQNQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARVL
	`*::*::*:.***  *: *::.** **  *:::**`	`.***:****** **..***:*:************:*`
Variation		
	-LQ-DK-VN---SI-HS-LDD--T--GI-QKD--	P---H------M--DG---M-E------------I-
	-IK-----S---GV--T---Q--I------R---	--------------K------K--------------
	--------E---NT--N---N-------------	------------------------------------
	------------H---R---K-------------	------------------------------------
	------------Q---------------------	------------------------------------
C34·N36		
	WMEWDREINNYTSLIHSLIEESQNQQEKNEQELL	SGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARIL

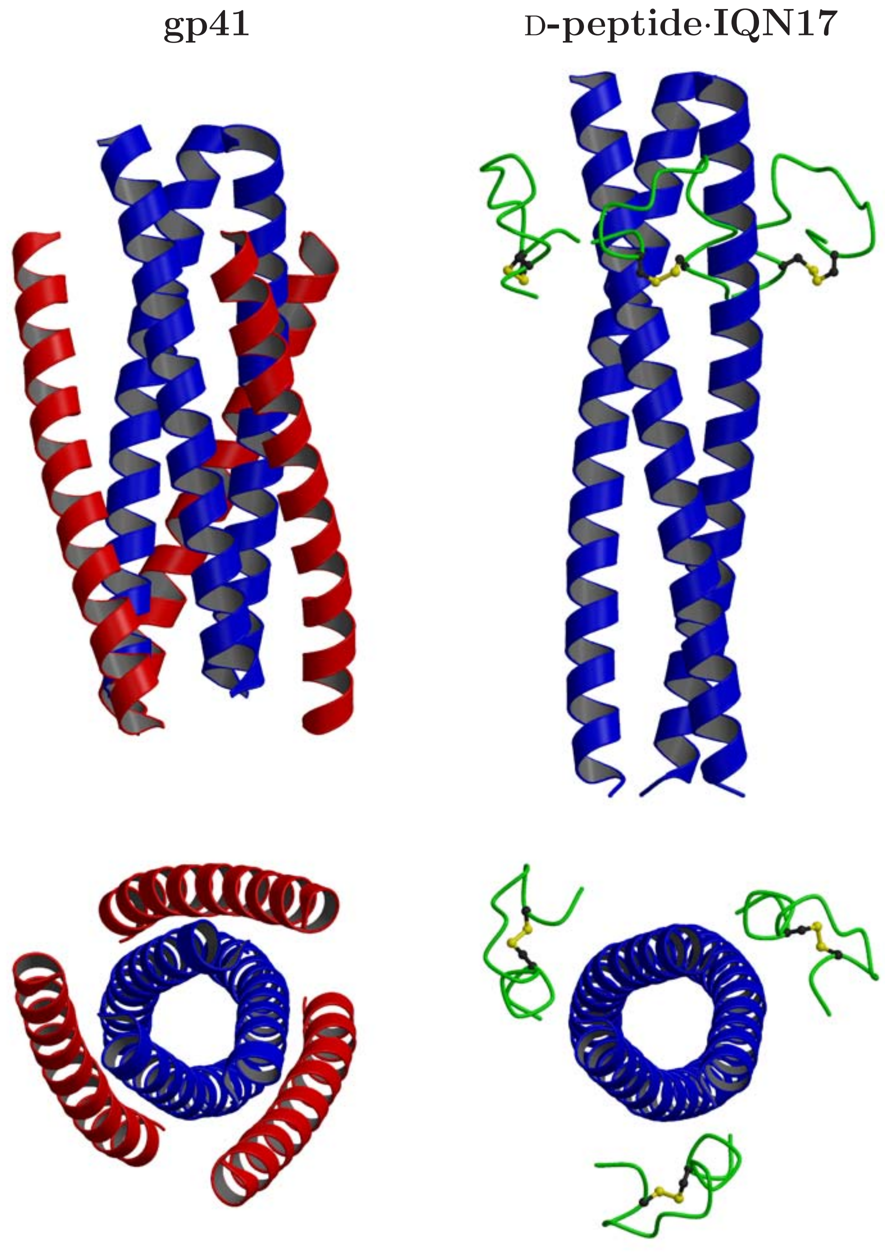
Table 4-1: **Conservation of gp41 six-helix bundle sequence among HIV-1 variants.** Variant sequences from the six-helix bundle region of gp41 are displayed along with the consensus sequence, the degree of conservation ([*] strictly conserved; [:] moderately conserved; [.] weakly conserved; [ ] unconserved), and the observed variations from the consensus. The sequence of C34·N36 is also displayed.

of the six-helix bundle region of gp41 is quite highly conserved among HIV-1 isolates (see Table 4-1). Among the same thirty isolates, the 36 residue N-terminal sequence present in the structure of the gp41 core ectodomain (N36) contains 28 absolutely conserved residues out of 36 total (78%), and of the eight other positions, four are highly conservative variations (89%). Even among the variable positions, none show a wide degree of variability. The 34 residue C-terminal sequence present in the ectodomain structure (C34) is somewhat less conserved than the N36 sequence, but still much more so than gp160 as a whole. Only fifteen of the 32 positions are strictly conserved (47%), but eleven of the other positions are only conservatively varied (81%). This relative conservation of sequence enhances the attractiveness of this area for targeted drug design.

Extensive structural information about the fusogenic conformation of gp41 and the six-helical bundle is available. The core domain has been crystallized in several constructs, and the structure has been solved to high resolution [21, 147, 161]. The structure solved by Chan *et al.* is displayed in Figure 4-2. The solution structure of the complete ectodomain of gp41 from the closely related Simian Immunodeficiency Virus (SIV) has also been solved by NMR spectroscopy [16]. In addition, atomic resolution structures with two inhibitory molecules (a D-peptide and a small molecule) bound to an N-terminal coiled-coil construct have been solved [43, 167].

4.1.2 Inhibitors targeting the N-terminal coiled coil of gp41

Molecules that bind to the N-terminal coiled coil of gp41 prevent association with the C-terminal sequence and thus prevent the formation of the fusogenic trimer-of-hairpins conformation. To date there have been several inhibitors of HIV-1 viral-cell membrane fusion identified which act by this mechanism. First, peptides with sequences corresponding to the C-terminal helical region have been identified as inhibitors [162]. When constrained to a helical form by a chemical linker, the potency of these peptides is enhanced [75], and several hydrophobic residues which occupy



gp41

D-peptide·IQN17

Figure 4-2: **Structure of the gp41 six-helical bundle and D-peptide–IQN17 complexes.** Two views of both the core structure of the fusogenic state of gp41 and the complex of the D-peptide with IQN17 (a gp41 coiled-coil model) are presented. In both cases, the central trimeric coiled coil is displayed in **blue** ribbon. The C34 helices of the six-helical bundle are displayed in **red** ribbon, and the D-peptide is displayed in **green** tube, with the disulfide linkage shown in atom-colored ball-and-stick representation. This figure was prepared with MOLSCRIPT [87] and RASTER3D [105].

a pocket on the surface of the N-terminal coiled coil have been identified as playing a key role in modulating the binding affinity [20]. Peptides from the analogous C-terminal region of SIV gp41 have also been shown to be inhibitors of HIV-1 infection [101]. Smaller inhibitors which bind to the same pocket on the coiled coil have also been identified. A D-peptide isolated by mirror-image phage display fills the binding pocket with residues similar to those of the C-peptides, albeit in a different orientation [43]. While the side chains of two tryptophan residues and an isoleucine from the C-terminal helix residue occupy this target pocket in the gp41 structure, the side chains of two tryptophan residues and a leucine residue from the D-peptide fill the same space in the complex of the D-peptide and the coiled coil. A second inhibitor consisting of a shorter C-peptide sequence linked to a non-peptidyl moiety was developed through screening of combinatorial libraries [46, 167], with the non-peptidyl portion of the molecule occupying the same target pocket on the coiled coil as the other studies identified as most significant. Two additional small molecule inhibitors, again targeting the same pocket, were identified through computational docking studies and subsequent experimental screening of the resulting matches [34, 71].

4.2 Database screening strategy

A side benefit of the charge optimization methodology is a framework in which the electrostatic binding free energies of a set of geometrically related ligands can be quickly estimated. In Section 2.4 it was shown that the electrostatic binding free energy of a set of ligands of variable charges, but fixed geometries and bound state conformations, is dependent only on the charges on the ligand. Thus, once the matrices involved in the optimization formulation have been computed, the calculation of the binding energy of any charge distribution on the ligand scaffold can be calculated at very little computational cost.

This fast evaluation of electrostatic energies can be used as the basis for a hier-

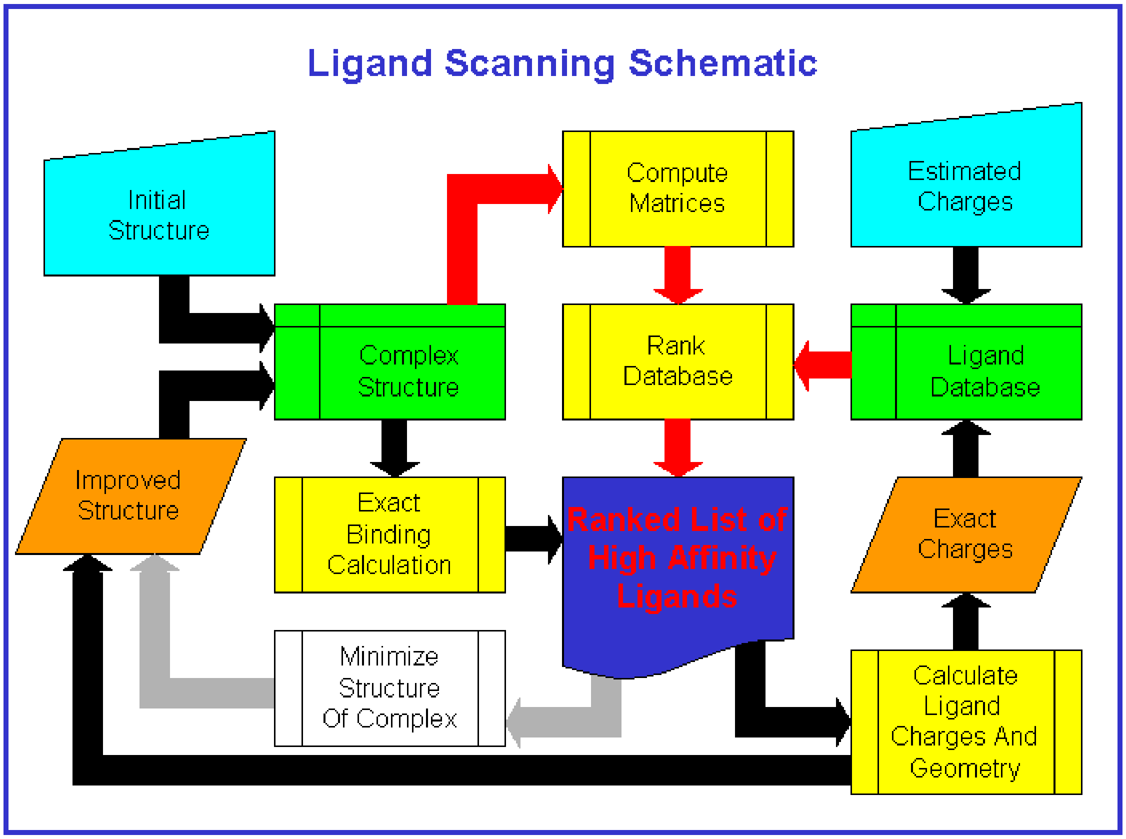

Ligand Scanning Schematic
Initial Structure
Compute Matrices
Estimated Charges
Complex Structure
Rank Database
Ligand Database
Improved Structure
Exact Binding Calculation
Ranked List of High Affinity Ligands
Exact Charges
Minimize Structure Of Complex
Calculate Ligand Charges And Geometry

Figure 4-3: **The ligand scanning procedure.** A flow chart of the ligand scanning procedure is outlined. The method centers around a ranked list of ligands, continuously re-ordered with increasingly accurate methods of computation.

archical scheme for the design of improvements to a known ligand. An optimization of different regions of the ligand of interest is used to pinpoint areas where significant improvements over the initial ligand are possible. A library of modified ligands is then generated by combinatorially substituting simple functional groups at select positions — due to combinatorial explosion, this library can easily be very large (four substitutions at six sites yields 4096 ligands and six substitutions at ten sites yields over 60 million). Initially, charges on the ligands may be estimated using a simple rule-based method at the level of functional groups, with no effects of the chemical environment taken into account. The library is ranked using the estimated charges on the initial ligand scaffold, and a set of top ranking ligands is selected for more detailed computation. The initial energy evaluation uses the approximate shape of the initial ligand and approximate charges, but more accurate energies can be attained using the actual partial atomic charges. Partial atomic charges on high ranking ligands can be computed by fitting to the electrostatic potential calculated using quantum mechanics. This procedure takes several minutes per (relatively small) ligand, and even longer if the quantum mechanical geometry must also be determined, and so is unfeasible for an overly large set of ligands but is easily applicable to a reasonably large subset. These more accurate charges are used to re-rank the ligands, again using the approximate shape of the initial ligand and the fast evaluation of binding free energies. More accurate evaluation of the binding energetics of a smaller subset of the top ranking ligands may then be done using the true shape of the new ligand. These calculations are somewhat more computationally expensive, often taking several hours, and thus only a relatively small selection of ligands can easily be analyzed at this level. As a further refinement, additional energy terms may be added, and as accurate, and costly, computations as are desired can be done as a final stage on the top-ranking set of ligands. Any of the top-ranking ligands may also be used as initial ligands to repeat the procedure, leading to successively more complicated derivatives of the starting ligand. The hierarchical procedure, using approximate calculations to successively reduce the size of the number of ligands being considered while simultaneously increasing the accuracy of the calculations, allows a large library of ligands to be evaluated in a reasonable time. Most importantly, those ligands of particular interest have their binding free energies calculated with as few approximations as is feasible. Figure 4-3 displays a schematic of the procedure. The key requirement for this method to be successful is for the lower levels of the hierarchy to be accurate enough that (1) ligands which are computed to bind tightly at the higher levels of the procedure are not eliminated by the most approximate calculations, and (2) substantial numbers of low affinity ligands are eliminated at a early stage.

4.3 Methods

Sequence analysis. Sequences of the *env* polypeptide (gp160) from HIV-1 variants were obtained from the Swiss-Prot sequence database [146]. Multiple sequence alignments were performed on the entire length of *env* using the ClustalX software package [149], using default parameters.

Preparation of structures. All structures used are contained in the Protein Data Bank (PDB) [125]. The gp41 core structure used was taken from the structure of three 36-residue peptides from the N-terminal region of gp41 (N36) in complex with three 34-residue peptides from the C-terminal region (C34) (PDB ID 1aik) [21]. The D-peptide structure was a complex of the D-peptide with a chimæric model system for the study of the N-terminal coiled coil of gp41 consisting of a portion of the GCN4 leucine zipper fused to the sequence of gp41 forming the target pocket (PDB ID 1czq) [43]. Hydrogen atom positions were added using the hbuild facility [14] within the charmm computer program [11]. The param19 parameter set [11] was used, with the addition of aromatic hydrogens on Phe, Tyr, Trp and His for consistency with the parameters used in the continuum electrostatic calculations. Several other structures, solved both by X-Ray crystallography and by NMR were used for visual comparison (PDB IDs: 1env [161], 1szt [147], 2ezo [16]).

Continuum electrostatic calculations. All continuum electrostatic calculations were performed using a locally modified version of the delphi computer program [55, 57, 134, 136] to solve the linearized Poisson–Boltzmann equation. An internal dielectric constant of 4 and an external dielectric constant of 80 was used unless otherwise specified, and the bulk ionic strength was set to 0.145 M. The molecular surface used to define the dielectric boundary was generated using a 1.4 Å radius probe, and an ion exclusion (Stern) layer [9] of 2.0 Å was also applied. Protein partial atomic charges and radii were taken from the parse parameter set [140] with a few minor changes. Charges on the bridging ring carbons of tryptophan were assigned to $0e$, charges for proline and for disulfide bridged cysteine residues were taken from the PARAM19 parameter set [11], and the charges from glutamate and lysine side chains were used for charged C and N termini respectively.

All computations on the six-helical bundle were performed using two-step focusing boundary conditions on a 167×167×167 unit cubic grid, in which the longest dimension of the molecule occupied first 23% and then 92% of one edge of the grid, resulting in final grid spacing of 0.34 Å. Boundary potentials for the more highly focused calculation were obtained from the lower focused calculation, and Debye–Hückel potentials were used at the boundary of the lower run. All calculations were averaged over ten translations of the structure on the grid in order to minimize artifacts from the the placement of the point charges and molecular boundaries onto the finite difference grid.

For the D-peptide, binding free energy calculations were performed using the same two-step focusing boundary conditions but on a 225×225×225 unit grid, resulting in final grid spacing of 0.29 Å. Calculations to determine the matrix elements for electrostatic optimization were done using a four-step focusing procedure on a 65×65×65 unit grid, with the molecule occupying 23%, 92%, 184%, and finally 276% of the grid, resulting in a final grid spacing of 0.33 Å. For the highest resolution calculations, the grid was centered on the region of interest, and interactions involving groups falling outside of this grid were computed from the 92% fill calculation.

Electrostatic optimization. Electrostatic optimization was performed using locally written software as previous described [23, 77–80, 92–94]. Singular value decomposition [119, 143] was used to remove all basis vectors with singular values smaller than 10^{-5} of the largest singular value, or with errors of more than 25% over 10 translations. Typically this involved the removal of 31 out of 84 basis vectors; several of residues most significantly removed from the interface pay almost no desolvation, leading to a number of very small eigenvalues in the desolvation matrix. When constraints were applied, the standard constraints were to limit individual partial atomic charges to a maximum magnitude of $0.85e$ and to limit total residue charges to -1, 0, or $+1e$; these constraints were chosen to limit the optimization to a chemically reasonable space. Constrained optimizations were performed using the LOQO software package [133, 154, 155].

Design of enhanced electrostatic interactions. The conformation of each of five N36 residues were considered individually, with all other residues fixed in their crystal structure conformation. For each position, the cardinal torsions for each dihedral were selected (±60° and 180° for sp^3–sp^3 bonds, and ±30°, ±90°, ±120°, ±150° and 180° for sp^3–sp^2 bonds). The side-chain atoms of the residue were then minimized to convergence using the adapted-basis Newton–Rhapson (ABNR) minimization algorithm in the CHARMM computer program [11] with the PARAM19 parameter set [11]. Electrostatics were treated in two ways: first using a distance-dependent dielectric constant of $\epsilon = 2r$, and secondly with all electrostatic interactions excluded. The minimum energy conformation was selected as favored for all residues for which this conformation was not clearly seen to be a minor variation of the crystal structure conformation upon visual inspection.

Ligand scanning. Partial atomic charges on the database of tryptophan derivatives were computed by fitting charges to the quantum mechanical electrostatic potential. All calculations were done on model compounds based on 3-methyl indole; a set of calculations on tryptophan alone and in the context of a short length of protein backbone showed that the charges computed for the side-chain atoms were largely independent of the context. Electrostatic potentials were calculated at the HF/6-31G* level of theory using the GAUSSIAN98 software package [50] with structures determined using the JAGUAR quantum chemistry program [130]. Restrained fits to the ESP were performed using the RESP computer program following the standard

Ligand	$\Delta G^{desolv.}_{rec.}$	$\Delta G^{desolv.}_{lig.}$	$\Delta G^{inter.}_{lig.-rec.}$	$\Delta G^{inter.}_{lig.-lig.}$	ΔG^{es}
C34	+17.14	+19.80	−11.47	–	+25.47
3 × C34 [a]	+17.28	+19.95	−11.66	+0.89	+26.46
D-peptide	+3.46	+9.24	−4.31	–	+8.39
3 × D-peptide [a]	+3.50	+9.26	−4.37	+0.12	+8.51

[a] per ligand

Table 4-2: **Electrostatic free energy of ligands binding to the gp41 coiled coil.** The electrostatic binding free energy terms (in kcal·mol^{-1}) for both C34 and the D-peptide binding to the gp41 N-terminal coiled coil are tabulated. Results for single ligand binding are accompanied by the results for the simultaneous binding of three ligands.

procedure [6, 28]. For calculations involving the real shape of tryptophan derivatives, the structure of the model compound was superimposed onto the target tryptophan in the context of the D-peptide–IQN17 complex. The PROFIT software package [102] was used for the fit, and only the heavy atoms within the rings were used. Even for derivatives with a partially aliphatic six-membered ring, this procedure was seen to perform well by visual analysis.

4.4 Results

4.4.1 Electrostatics of C34 and D-peptide binding

To gain a perspective on the binding energetics of the natural system, an overall electrostatic analysis of the gp41 core ectodomain structure was carried out. The electrostatic contribution to the free energy of binding of the C34 peptide to the N36 coiled coil was computed to be unfavorable by 25.47 kcal·mol^{-1}. Both the coiled coil and the peptide pay significant desolvation penalties on binding (+17.14 and +19.80 kcal·mol^{-1} respectively), but only make moderately favorable compensating interactions (−11.47 kcal·mol^{-1}), as detailed in Table 4-2.

Three C34 peptides can bind to one N36 trimer, and thus cooperativity effects are possible. The electrostatic contribution to binding of three isolated C34 helices was computed in the same way as were the previous results for the binding of a single C34 helix. The desolvation and interaction terms (per ligand) change very little between the two cases. The desolvation penalties are slightly higher (both by less than 0.2 kcal·mol^{-1}), and the interaction is slightly more favorable (also by about 0.2 kcal·mol^{-1}). Both these results may be expected due the increase of excluded solvent in the triply bound state; increased exclusion of solvent results in an increase in the desolvation energy, but the reduced screening effects of solvent in the bound state make the interactions stronger as well. In addition to this, each pair of C34 peptides interacts unfavorably by +0.89 kcal·mol^{-1}, and it is this direct repulsion that contributes most of the 1.0 kcal·mol^{-1} per ligand calculated anti-cooperativity. While there was a small anti-cooperative effect computed, it is very small compared to the overall unfavorable electrostatics of binding.

A similar analysis was then performed on the complex of the D-peptide with IQN17, a chimæric construct in which a region of the N-terminal coiled coil including the target pocket is fused to a trimeric form of the GCN4 leucine zipper. Due to the much smaller size of the D-peptide relative to the C34 helix, the magnitudes of the electrostatic energy terms computed for the D-peptide binding to an N-terminal coiled-coil construct are much smaller in magnitude than the corresponding terms for the C-peptide binding. The overall electrostatic contribution to binding is still unfavorable (by +8.39 kcal·mol^{-1}), but the details are a little different. The receptor pays only a small desolvation penalty of +3.56 kcal·mol^{-1} upon binding, while the D-peptide pays a much larger cost (+9.24 kcal·mol^{-1}). The favorable interactions made between the ligand and receptor in the bound state contribute only −4.31 kcal·mol^{-1}, not nearly enough to compensate the desolvation costs. Again, these results are detailed in Table 4-2.

As was the case for C34 binding, three D-peptides can bind simultaneously to the target coiled coil, and thus the electrostatic contributions to cooperative binding were computed. The small enhancements of desolvation penalties and interaction energy due to a more solvent excluded bound state that were observed with the C34 peptide are also seen here, although all are below 0.1 kcal·mol^{-1} in magnitude. The direct interaction between the peptides is also very small, only +0.12 kcal·mol^{-1}, and this tiny effect fully accounts for the net computed cooperativity. In this system it is clear that the D-peptides interact very weakly even with three present in the bound state, and thus a model of a single peptide binding is wholly adequate.

4.4.2 Optimization of D-peptide binding

Since the overall electrostatic contribution of the D-peptide to binding was computed to be unfavorable, modifications to the D-peptide that may electrostatically enhance the binding affinity were considered. The partial atomic charges on all the side-chain atoms of the D-peptide were varied so as to optimize the electrostatic binding free energy (see Table 4-3). When the charges were allowed to vary freely, the optimal improvement in binding free energy was 4.30 kcal·mol^{-1}, although this required an unphysical charge distribution, with a net charge of $-13.58e$. With individually optimized side chains (fixing all other partial atomic charges to their wild-type values), three residues (Glu4, Arg6, and Trp10) showed an optimal improvement of over 1.0 kcal·mol^{-1}, although again for Glu4 and Arg6, this required unphysical charges, with the total charge on each residue near $-10e$. Trp10 had a more reasonable optimal charge of $-0.63e$, as did Trp12 (net charge $-0.92e$) which had the fourth best optimal improvement (0.85 kcal·mol^{-1}). When constraints were added to limit the optimal partial atomic charges charges to chemically reasonable values and to limit total residue charges to integers between -1 and $+1e$, only Trp10 and Trp12 showed optimal improvements of above 0.5 kcal·mol^{-1}. Both residues optimized to be neutral in total charge, with Trp10 improving by 0.86 kcal·mol^{-1} and Trp12 improving by 0.66 kcal·mol^{-1} relative to natural tryptophan. The global constrained optimum has a net charge of $-2e$ and binds 2.01 kcal·mol^{-1} better than wild type.

		X-Ray Structure				Enhanced Structure			
		Free		Constrained		Free		Constrained	
Residue		$Q_{tot.}^{opt.}$	$\Delta\Delta G^{es}$	$Q_{tot.}^{opt.}$	$\Delta\Delta G^{es}$	$Q_{tot.}^{opt.}$	$\Delta\Delta G^{es}$	$Q_{tot.}^{opt.}$	$\Delta\Delta G^{es}$
Gly[a]	1	–	–	–	–	–	–	–	–
Ala[b]	2	−0.18	−0.23	0	0.00	−0.21	−0.30	0	0.00
Cys[c]	3	–	–	–	–	–	–	–	–
Glu	4	−10.23	−1.16	−1	−0.08	−10.65	−1.27	−1	−0.09
Ala[b]	5	−0.65	−0.13	0	0.00	−0.85	−0.22	0	0.00
Arg	6	−9.83	−1.04	−1	−0.36	−12.99	−1.72	−1	−0.52
His	7	−4.09	−0.26	−1	−0.21	−3.97	−0.60	−1	−0.52
Arg	8	−26.80	−0.67	−1	−0.07	−13.01	−1.75	−1	−0.23
Glu	9	−1.05	−0.14	−1	−0.09	−3.31	−0.38	−1	−0.25
Trp	10	−0.63	−1.12	0	−0.86	−0.97	−1.84	−1	−1.63
Ala[b]	11	−3.42	−0.26	0	0.00	−8.04	−1.51	0	0.00
Trp	12	−0.92	−0.85	0	−0.66	−2.19	−2.54	−1	−1.44
Leu	13	+0.10	−0.09	0	−0.06	−0.10	−0.12	0	−0.09
Cys[c]	14	–	–	–	–	–	–	–	–
Ala[b]	15	−0.25	0.00	0	0.00	−1.56	−0.10	0	0.00
Ala[b]	16	+0.38	−0.55	0	0.00	+0.36	−0.48	0	0.00
All		−13.58	−4.30	−2	−2.01	−13.78	−6.71	−5	−3.36

[a] Since glycine has no side-chain heavy atoms, this residue was not optimized.

[b] Due to the constraints used, alanines are forced to be completely hydrophobic in the constrained optima.

[c] Due to the requirement for the formation of a disulfide linkage between Cys 3 and Cys 14, these residues were not optimized.

Table 4-3: **Electrostatic optimization of D-peptide side chains individually and together.** Results of the optimization of the charges of the side-chain atoms of the D-peptide (total optimal charge and optimal improvement over wild type) are tabulated for both the crystal structure and the structure designed for enhanced electrostatic interactions. Each entry corresponds to the optimization of the atoms of the side chain of the specified residue alone, with the charges of all other atoms fixed at their wild-type values. The "All" entry corresponds to the simultaneous optimization of the side-chain atoms of all residues, with the charges of all backbone atoms fixed to their wild-type values. All free energies are in kcal·mol^{-1}.

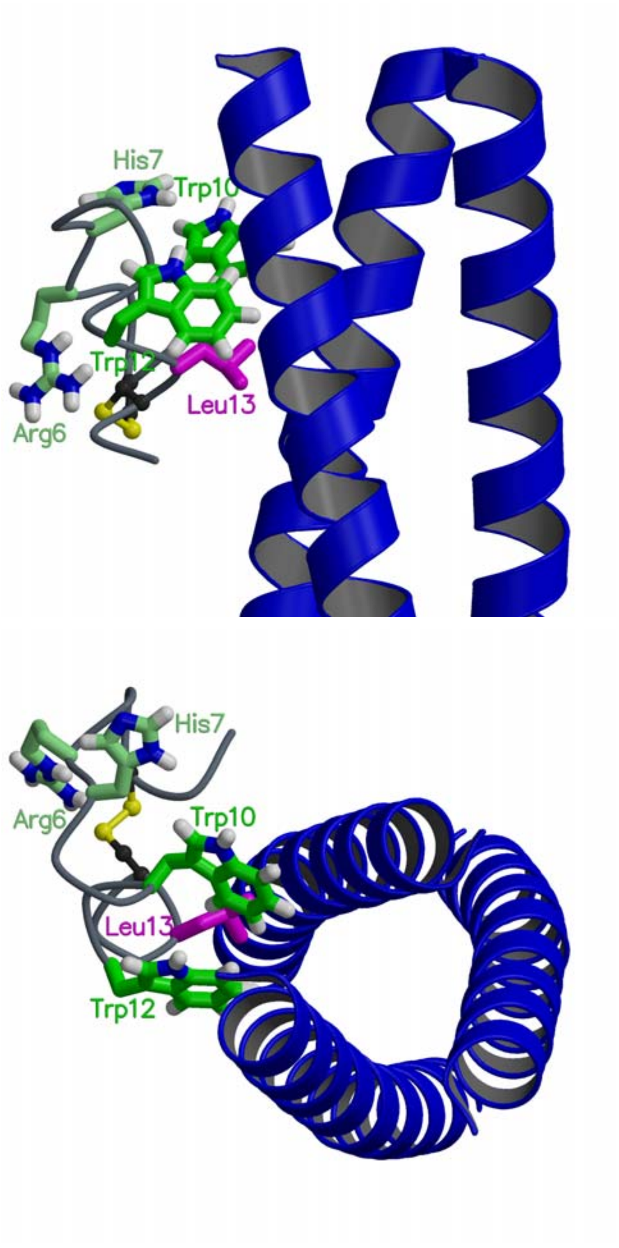


Figure 4-4: **Structural details of the D-peptide bound to IQN17.** The key side chains involved in D-peptide binding to IQN17 are displayed. Trp10, Trp12 and Leu13 make direct contact with the target pocket on the receptor. Of these, Trp10 and Trp12 show significant possibilities for improvement by electrostatic optimization (**green**), while the hydrophobic Leu13 shows little room for improvement (**magenta**). Two residues further removed from the receptor (Arg6 and His7) show reasonable opportunity for improvement, although less than the two tryptophans (**light green**). This figure was prepared with MOLSCRIPT [87] and RASTER3D [105].

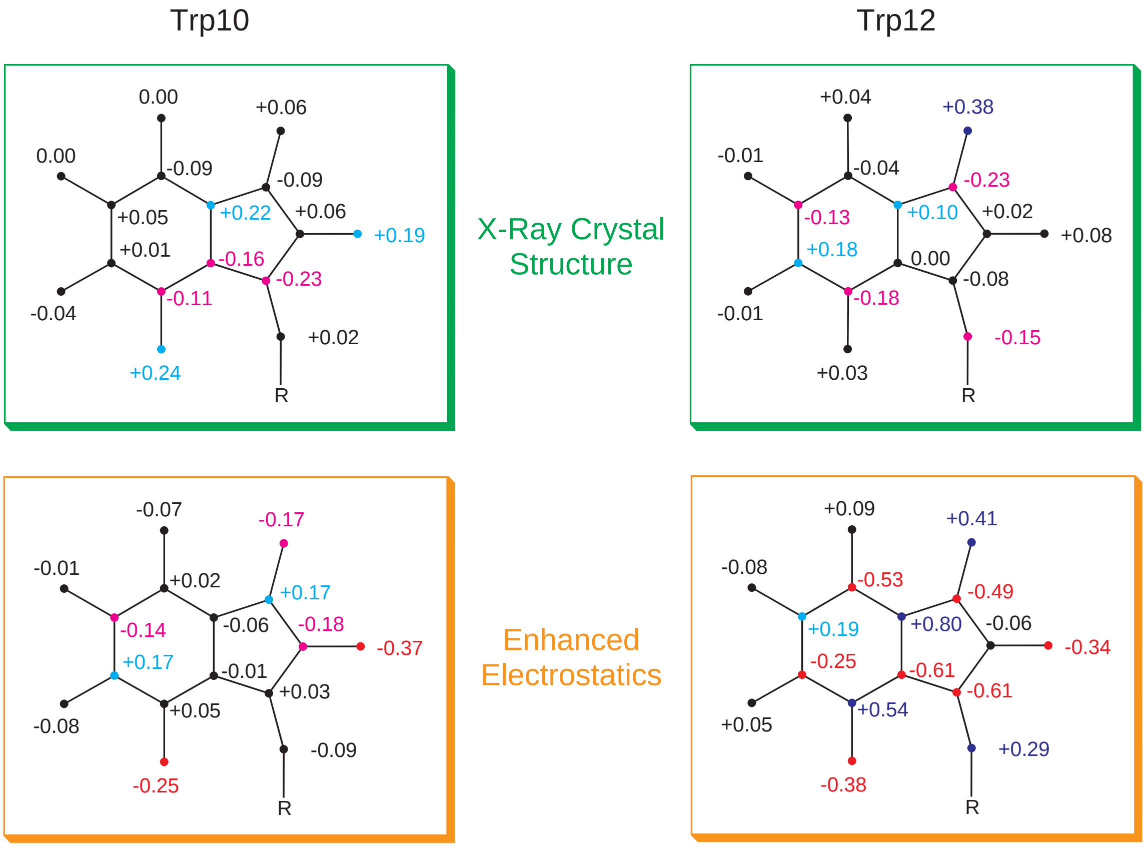


Figure 4-5: **Optimal charges on Trp10 and Trp12 of the D-peptide.** The partial atomic charges at the atom centers of Trp10 and Trp12 that optimize the electrostatic free energy of binding to IQN17 are shown. The positions are colored by charge ($q \leq -0.25$; $-0.25 < q \leq -0.10$; $-0.10 < q < +0.10$; $+0.10 \leq q < -0.25$; $+0.25 \leq q$). Results for both the crystal structure and the structure designed for enhanced electrostatic interactions are displayed.

The optimal charge distributions on both tryptophans show a quite hydrophobic character (see Figure 4-5). On Trp10 the largest magnitude partial charge in the optimum is 0.24e, and the optimal charge distribution is not immediately suggestive of any modifications. The most notable aspect of the Trp10 optimum is that the NH group of the indole takes on near neutral charges. The Trp12 optimum is also primarily hydrophobic, with the exception, in this case, of the indole NH, which optimizes to a dipole similar to the naturally occurring NH group.

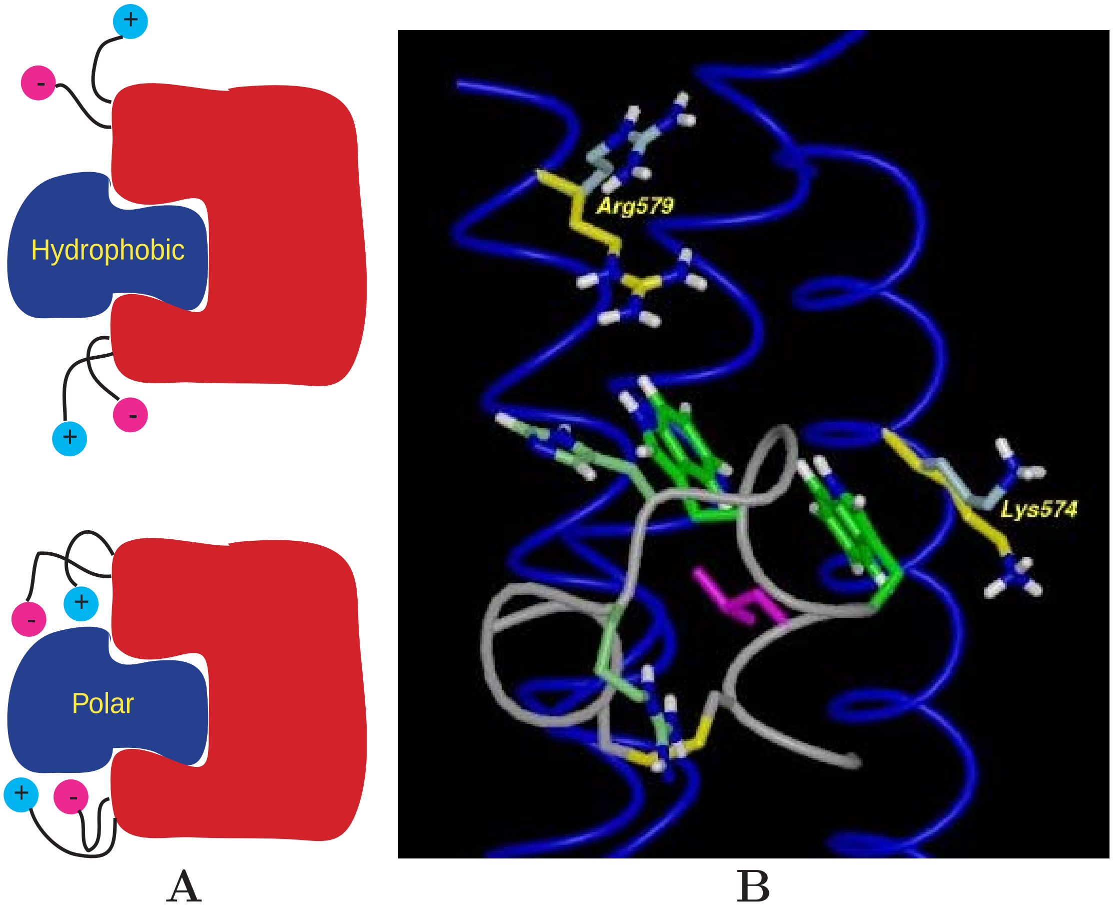


Figure 4-6: **Design of enhanced electrostatic interactions. (A) Top:** When polar or charged residues located on the edge of the binding interface are oriented to interact with solvent, the optimal ligand may be largely hydrophobic. **Bottom:** However, when these peripheral residues are poised to interact with the ligand in the bound state, the optimal ligand for the same receptor may be polar, or even charged. **(B)** Two positively charged residues on the receptor (Lys574 and Arg579, using standard gp41 numbering) are located on the periphery of the binding site, and can easily make closer contact with the D-peptide. Using the structure with closer polar contacts should allow for the design of a ligand which can make more favorable electrostatic interactions.

4.4.3 Design of enhanced electrostatic interactions

While the residues that line the D-peptide binding pocket of the N-terminal coiled coil are largely hydrophobic, several polar and charged residues are located around the edge of the pocket. Since many these are surface residues, it is likely that these groups could adopt alternate conformations, particularly if functional groups on the D-peptide were poised to interact with them. With a hydrophobic ligand, polar residues on the receptor may prefer to make interactions with solvent rather than interact with the ligand, but from a design perspective it may be useful to target an alternate conformation in which a polar ligand may make direct interactions with polar groups on the receptor (see Figure 4-6-A). The possibility of alternate conformations was investigated for all conserved polar residues around the D-peptide binding pocket. Five such residues were considered (Trp571, Lys574, Gln575, Gln577, and Arg579, using standard gp41 numbering). Of these five, only Lys574 and Arg579 were found to have alternate low energy conformations which make much closer contact with the D-peptide than are observed in the crystal structure (see Figure 4-6-B). For the other three residues, the crystal structure conformation was the minimum energy conformation both with an electrostatic energy term calculated with a distance-dependent dielectric constant and with no electrostatic energy term used. For Arg579, this alternate conformation is the same as that found in the C34–N36 X-ray crystal structure. While the alternate Lys574 conformation is not identical to those found in other structures, a great deal of structural variability is seen for this residue over all the known structures.

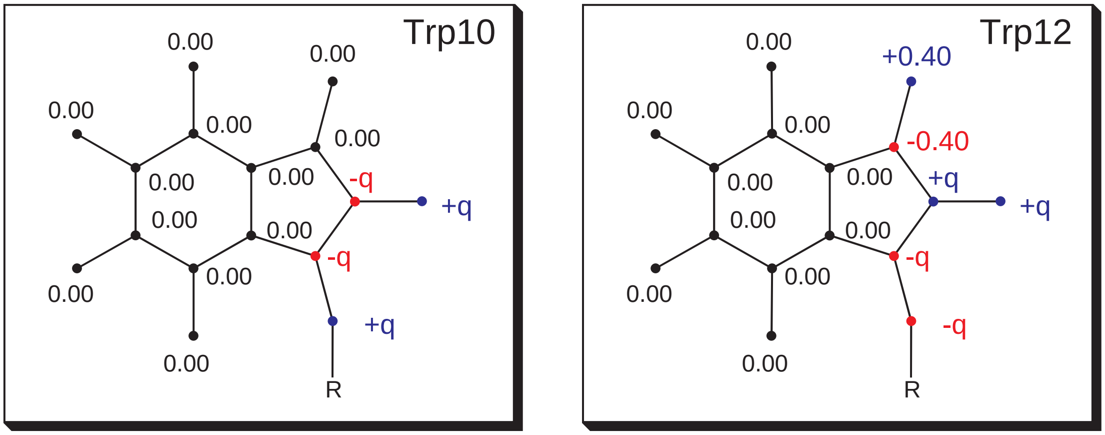
The partial charges of the side-chain atoms of the D-peptide were optimized for binding to this alternate receptor structure, and the expected enhancement of the role of electrostatics was observed. In the unconstrained optimization, the global optimum binds 6.71 kcal·mol^{-1} better than wild type, compared to a 4.30 kcal·mol^{-1} improvement for the crystal structure. In addition, six residues show individual improvements of more than 1.0 kcal·mol^{-1}, and the greatest improvement is 2.54 kcal·mol^{-1}. These enhancements are also seen in the constrained optimizations. The global constrained optimum has a net charge of $-5e$ and binds 3.36 kcal·mol^{-1} better than wild type. Individually, four residues show an improvement of over 0.5 kcal·mol^{-1} (Arg6, His7, Trp10 and Trp12), with both tryptophans showing optimal improvements of above

1.0 kcal·mol^{-1}.

The optimal charge distributions of both tryptophans have a negative overall charge and have correspondingly increased polarity compared with the optimal charges from the crystal structure. In both distributions, the majority of the charge is located on the five-membered ring of the indole, with the six-membered ring being largely hydrophobic. Again, the optimal charges on the NH group on Trp10 do not resemble the natural charges, having a slight dipole of opposite sign, while the same charges on Trp12 show remarkable similarity to the natural NH. When constrained to a neutral net charge, both Trp10 and Trp12 still show an optimal improvement in binding free energy of over 1.0 kcal·mol^{-1} (1.13 and 1.24 kcal·mol^{-1} respectively). Again, in both cases the six-membered ring optimizes to be largely hydrophobic, and the NH group takes on near natural charges only in the case of Trp12, while being near neutral for Trp10. In other areas, the charges are more varied, but again do not suggest any obvious modifications.

Using the optimal charge distributions as a guide, two "chemical-like" charge distributions were generated for the tryptophan scaffold (see Figure 4-7). These charge distributions consisted of completely hydrophobic six-membered rings, and paired positive and negative charges of equal value on a total of four atoms of the five-membered ring. For the Trp10 candidate, the NH charges were also zeroed, while for the Trp12 candidate these charges were left at the natural charges. The free energy of binding for these tryptophan replacements were evaluated with charges in the range of 0.0 to 0.5e. Individually, the substitution at Trp10 is computed to improve binding by upwards of 1.1 kcal·mol^{-1}, while that at Trp12 is computed to provide up to a 0.6 kcal·mol^{-1} improvement in binding free energy. With both substitutions made simultaneously, an improvement of up to 1.8 kcal·mol^{-1} over two natural tryptophans is calculated. This maximal improvement is seen for charge magnitudes of 0.3 or 0.4e on both derivatives. While these charges do not correspond to any chemically realizable molecule, they do indicate that significant improvements

A



B

	Trp10		Trp12	
q	X-Ray	Enhanced	X-Ray	Enhanced
0.00	-0.64	-0.79	-0.45	-0.49
0.10	-0.54	-1.00	-0.51	-0.56
0.20	-0.50	-1.13	-0.52	-0.61
0.30	-0.53	-1.18	-0.49	-0.64
0.40	-0.62	-1.15	-0.41	-0.63
0.50	-0.50	-1.05	-0.29	-0.64

C

		Trp12					
	q	0.00	0.10	0.20	0.30	0.40	0.50
Trp10	0.00	-1.24	-1.30	-1.33	-1.34	-1.33	-1.31
	0.10	-1.45	-1.51	-1.56	-1.56	-1.56	-1.54
	0.20	-1.61	-1.65	-1.69	-1.70	-1.70	-1.69
	0.30	-1.64	-1.70	-1.75	-1.79	-1.77	-1.76
	0.40	-1.62	-1.68	-1.73	-1.75	-1.76	-1.75
	0.50	-1.51	-1.58	-1.63	-1.66	-1.67	-1.60

Figure 4-7: **Chemical-like charges on Trp10 and Trp12 of the D-peptide.** The improvement in binding free energy (relative to wild type) for regularized "chemical-like" charges on Trp10 and Trp12 are shown. **(A)** The charge arrangements on the scaffold of each residue. **(B)** The improvement in binding free energy for individual substitutions on each position. **(C)** The improvement for simultaneous substitutions at both positions in the enhanced structure.

Derivative type	N	Example molecule
Fluoro	63	1,2,5-trifluoro-3-methyl indole
Chloro	6	4-chloro-3-methyl indole
Methyl	6	3,6-dimethyl indole
Heterocyclic	6	1-methyl indole
Oxy	2	3-methyl-5-oxy indole
Aliphatic	14	4,5,6,7-tetrahydro-3-methyl indole
Natural Trp	1	3-methyl indole
Total	98	

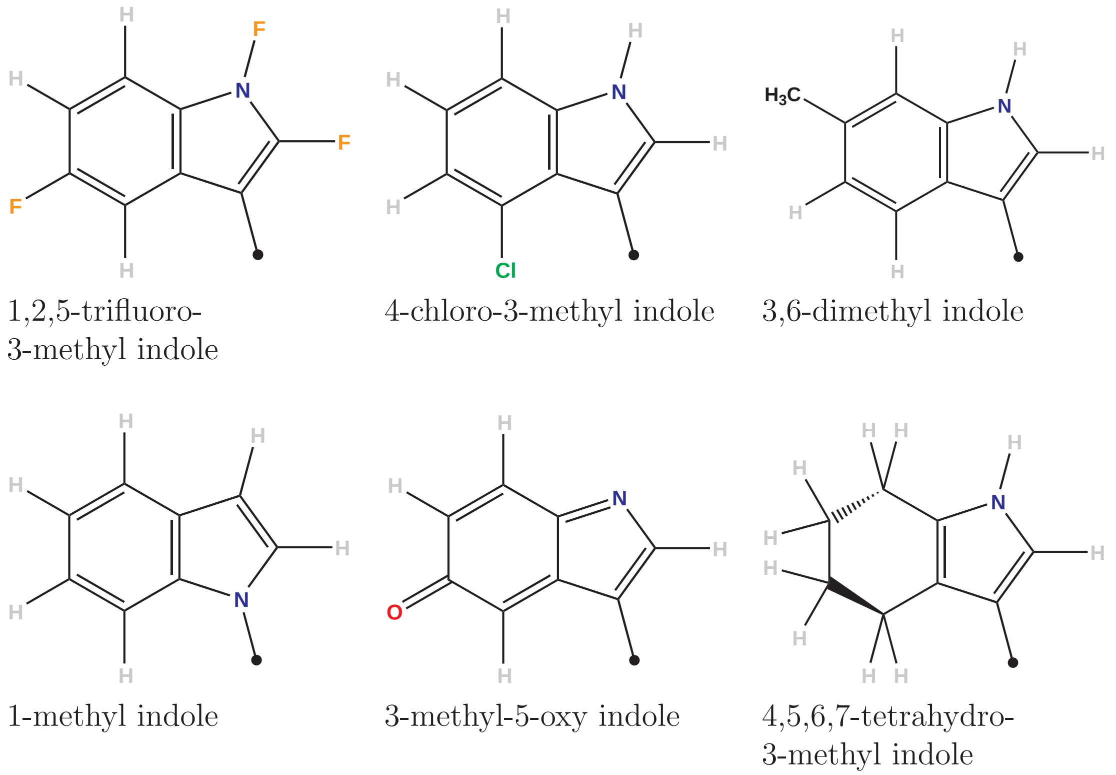
Table 4-4: **Summary of tryptophan derivatives in ligand scanning library.** The derivatives present in the library used for the ligand scanning procedure are classified into general types, and the number of each type in the library is indicated.

in binding free energy can be made even when the system is highly constrained to conform to chemical norms.

4.4.4 Ligand scanning of D-peptide tryptophans

The ligand scanning procedure outlined in Section 4.2 was applied to the screening of a database of tryptophan derivatives for substitution at the two tryptophan positions (10 and 12) on the D-peptide. The library, outlined in Table 4-4 and Figure 4-5 consisted of 98 unique derivatives of tryptophan which, applied to two sites, yields 9604 possible combinations. Ultimately, the charges for the complete database were computed quantum mechanically, and the scanning was done with these charges. However, for a large subset (the 63 fluoro derivatives) the computations were initially done with a rule-based charge determination method.

The top ten ligands predicted by the ligand scanning procedure (Table 4-6) have computed binding free energies between 0.6 and 0.8 kcal·mol^{-1} better than wild type, as computed using the quick energy evaluation. All have the same derivative at position 12, 4,5,6,7-tetrahydro-3-methyl-indole, which differs from tryptophan by the replacement of the four non-bridging carbons of the six-membered ring with aliphatic carbons. This molecule is significantly less polar than its aromatic precursor, with



1,2,5-trifluoro-3-methyl indole

4-chloro-3-methyl indole

3,6-dimethyl indole

1-methyl indole

3-methyl-5-oxy indole

4,5,6,7-tetrahydro-3-methyl indole

Table 4-5: **Structures of representative tryptophan derivatives in library.** The structures of one member of each class of derivatives present in the library used for the ligand scanning procedure are displayed.

near zero charges on the aliphatic atoms (compared to slightly above $+0.1e$ on aromatic hydrogens, and slightly below $-0.1e$ on aromatic carbons). More diversity is seen in the substitutions at position 10; however, all but one of the substitutions at this position have either two or four of the aromatic carbons on the six-membered ring replaced by aliphatic groups. The sole exception contains three fluorines on the six-membered ring, which results in a similar reduction in the polarity of this region. In addition, three of the top scoring derivatives at position 10 (including the absolute top scorer), have halogen substitutions at N_1, effectively reducing the NH dipole, as was observed in the optimal charge distributions. The commonalities seen in the top scoring ligand derivatives all correspond well to the optimal charges seen at the atom centers of Trp10 and Trp12. Similarly, the worst performing ligands have charge

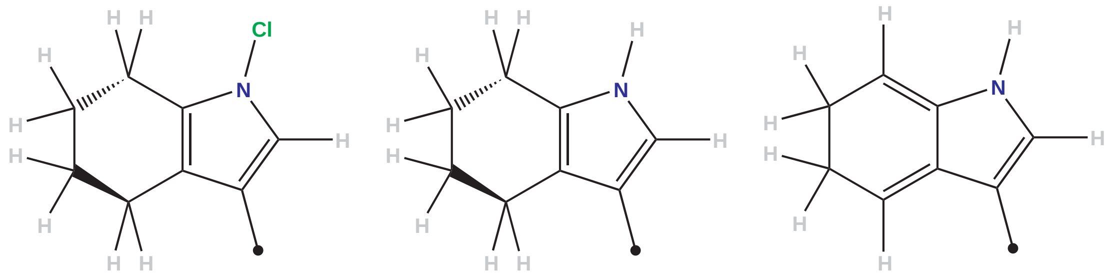
Rank	Position 10	Position 12	$\Delta\Delta G^{es}$
1	1-chloro-4,5,6,7-tetrahydro	4,5,6,7-tetrahydro	−0.77
2	5,6-dihydro	4,5,6,7-tetrahydro	−0.71
3	4,5,6,7-tetrahydro	4,5,6,7-tetrahydro	−0.70
4	1-fluoro-4,5,6,7-tetrahydro	4,5,6,7-tetrahydro	−0.66
5	2-fluoro-4,5,6,7-tetrahydro	4,5,6,7-tetrahydro	−0.65
6	2-fluoro-5,6-dihydro	4,5,6,7-tetrahydro	−0.63
7	7-chloro-4,5,6,7-tetrahydro	4,5,6,7-tetrahydro	−0.61
8	1,5,6,7-tetrafluoro	4,5,6,7-tetrahydro	−0.61
9	7-fluoro-5,6-dihydro	4,5,6,7-tetrahydro	−0.59
10	2-chloro-5,6-dihydro	4,5,6,7-tetrahydro	−0.59
666	Natural	Natural	0.00
9601	6-oxy	1,2,4,7-tetrafluoro	+4.85
9602	5-oxy	5-oxy	+5.35
9603	5-oxy	6-oxy	+5.59
9604	6-oxy	6-oxy	+6.03
9605	6-oxy	5-oxy	+6.19

Table 4-6: **Energetics of ligand scanning results.** The improvements in binding free energy (relative to wild type) are shown for the ten best and five worst scoring ligands as determined by the ligand scanning procedure, using quantum mechanically determined charges. All energies are in kcal·mol^{-1}.

distributions that differ from natural tryptophan in a manner opposite that of the optimum. Most of the worst binding ligands have oxy substitutions at positions 5 or 6, making the six-membered ring *more* polar. In addition, the fifth worst ligand includes a fluoro substitution at N_1 of Trp12, eliminating the NH dipole which, in contrast to that of Trp10, was strongly reproduced in the optimal charge distribution.

The electrostatic binding free energies of the top three ligands were recalculated using the correct shape for each ligand. With the more exact calculation, the top ligand (a 1-chloro-4,5,6,7-tetrahydro substitution at position 10 and a 4,5,6,7-tetrahydro substitution at position 12) remains the ligand with the highest computed improvement, with a computed binding free energy 0.9 kcal·mol^{-1} better than the initial ligand, and the second and third ligands, as ranked by the approximate shape calculations, have electrostatic binding free energies of 0.5 and 0.8 kcal·mol^{-1} better

		$\Delta\Delta G^{es}$	
Position 10	Position 12	Mid. Res.	High Res.
1-chloro-4,5,6,7-tetrahydro	4,5,6,7-tetrahydro	−0.77	−0.94
5,6-dihydro	4,5,6,7-tetrahydro	−0.71	−0.49
4,5,6,7-tetrahydro	4,5,6,7-tetrahydro	−0.70	−0.76



1-chloro-4,5,6,7-tetra-hydro-3-methyl indole

4,5,6,7-tetrahydro-3-methyl indole

5,6-dihydro-3-methyl indole

Table 4-7: **Highest resolution binding free energies for top scoring ligands.** The improvements in electrostatic binding free energy (in kcal·mol^{-1}) for the top scoring ligands from the ligand scanning procedure were calculated more accurately (High Res.), using both the quantum mechanically determined shape and charge distribution, with the results displayed, along with the results using an approximate shape (Med. Res.) and structures of the three tryptophan derivatives involved. The results obtained with the approximate ligand shape are all within 0.25 kcal·mol^{-1} of those computed with the more accurate ligand shape.

than wild-type. The greatest deviation between the exact and approximate shape calculations is 0.22 kcal·mol^{-1}, for a 5,6-dihydro substitution at position 10 and a 4,5,6,7-tetrahydro substitution at position 12.

While the above results were based on quantum mechanically derived charges, a rule-based method for assigning charges as a preliminary step can reduce the number of quantum mechanical calculations that must be done, by eliminating the worst binding ligands at an early stage. This rule-based procedure was used to rank a database consisting of the 63 fluoro derivatives, plus the natural tryptophan. Similar results were obtained with this rule-based method as with the exact charges. The relative time scales of each level of the hierarchy are displayed in Table 4-8. The first stage is very fast, taking only thirty seconds to screen the 64 possible ligands from substi-

	Single Ligand	Single Site	Full Database
Number of Ligands	1	64	4096
Level 1			
Estimating Charges	0.02s[a]	1.2s	1.2s[b]
Scanning DB	0.5s[a]	33s	32m
Total	**0.5s**[a]	**34s**	**32m**
Level 2			
Computing QM Geometries	2h48m	7d11h	7d11h[b]
Computing QM Charges	7m	7h28m	7h28m[b]
Scanning DB	0.5s[a]	33s	32m
Total	**2h55m**	**7d18h**	**7d19h**
Level 3			
Continuum ES Calculation	2h45m	7d8h[a]	469d[a]
Total	**2h45m**	**7d8h**[a]	**469d**[a]

[a] Estimated

[b] Charges only need to be determined for 64 molecules.

Table 4-8: **Timing of database ranking at different levels of ligand scanning.** All timings are for a single 1000 MHz Intel PIII processor.

tution at a single site, and only thirty minutes to evaluate the 4096 possible ligands arising from substitution at both sites. Computing charges quantum mechanically is substantially more costly, taking over seven hours on a single processor even if a geometry optimization is not performed, and taking over seven days on a single processor when the quantum mechanical geometry is first computed. However, since the charge determination only needs to be done on the set of 64 derivatives, the screening of the 4096 substitutions at both sites takes almost no more time than the screening of a single site. Furthermore, since the quantum mechanical computations for each molecule are independent, this second stage can be reduced to less than a day of computing time with as few as eight processors. The highest level of the hierarchical procedure, however, requires that an individual continuum electrostatic calculation be done for every ligand. For a single site, this would take a similar amount of time as the second stage (with geometry optimization). However, when substitutions are done at both sites, calculations on the entire database fo 4096 ligands would have

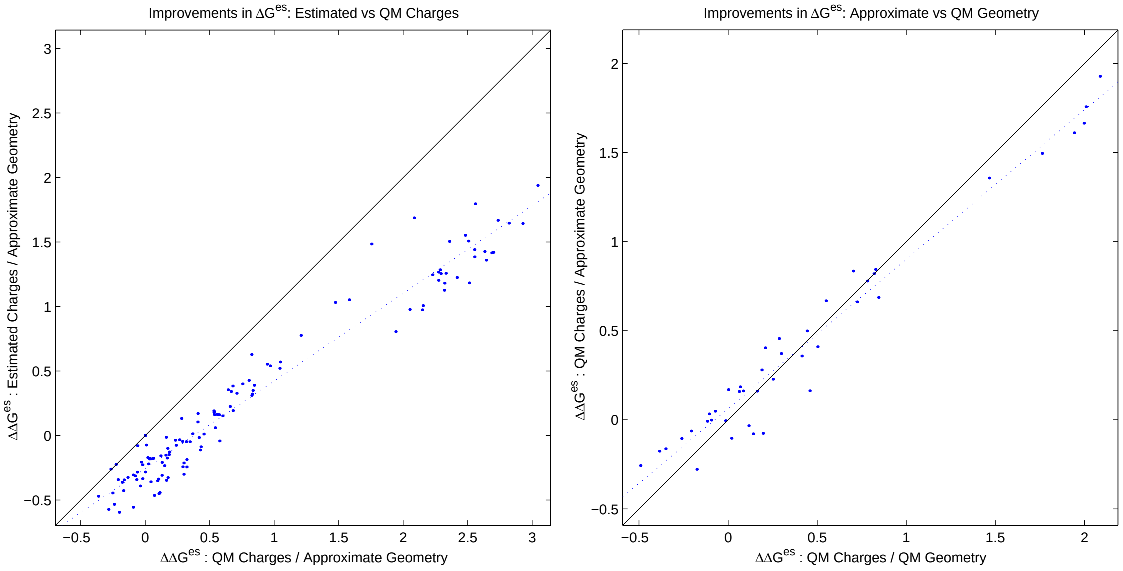


Figure 4-8: **Performance of multiple levels of the ligand scanning hierarchy. Left:** The relative binding free energies of all fluoro substitutions on both Trp10 and Trp12 computed with both estimated charges and an approximate shape are compared with those computed with quantum mechanically derived charges and the same approximate shape. The correlation between the two methods is quite strong. **Right:** The relative binding free energies of all monofluoro and difluoro substitutions on both Trp10 and Trp12 computed with quantum mechanically derived charges and an approximate shape are compared with those computed using the same charges but a more accurate geometry. The two methods show very strong correlation. All $\Delta\Delta\mathrm{G}^{\mathrm{es}}$ values are in kcal·mol^{-1}, and are relative to the binding free energy (computed with the same energy function) of the ligand with natural tryptophan residues.

to be done, taking over a year on a single processor! Even with 64 processors, this computation would take a week, and thus is infeasible as a general procedure for screening a database of any reasonable size.

In order for any hierarchical procedure to be effective, the rankings obtained for each successive stage must be similar to those obtained for the preceding stage. In Figure 4-8, two comparisons are made. First, the relative binding free energies (compared to wild type) for all fluoro derivatives both position 10 and 12 as calculated using estimated charges with the approximate shape were compared with those calculated using both the quantum mechanically derived charges and the same approximate shape. The results show good correlation between the two procedures, and thus validate the lowest level of the hierarchy — the approximate charges can be used with the approximate shape to eliminate the ligands with the worst predicted affinities without eliminating ligands which would be computed to bind tightly at the second level in the hierarchy. The highest level of the hierarchical procedure involves performing an individual continuum electrostatic calculation for each ligand, using both quantum mechanically derived charges and geometry. The relative binding free energies of all monofluoro and difluoro substitutions at either position 10 or 12 computed using the quantum mechanically derived charges and the approximate shape (the second stage in the procedure) were compared to the results of this computationally costly highest level. The results are very promising, showing very strong correlation. Once again, the quantum mechanically derived charges can be used with the approximate geometry to efficiently pick out those ligands likely to be computed to bind tightest using a less approximate energy function, eliminating those predicted to bind more poorly with out inadvertently also eliminating high affinity ligands. It thus seems that the multi-stage procedure, given a reasonable rule set for the first stage, can effectively speed up the screening process for a large database of ligands.

4.4.5 Stabilization of the D-peptide bound conformation

The D-peptide is quite small (only sixteen residues), and thus although it is constrained into a cyclic form by a disulfide linkage, it is likely relatively unstructured when isolated in solution. However, in order for productive binding, a single conformation with a helical structure formed to arrange the key residues (Trp10, Trp12, and Leu13) appropriately in the target pocket is necessary. This requirement for structure formation on binding provides an opportunity to enhance the binding affinity by a less direct mechanism. A modification which makes the bound state conformation of the D-peptide more stable while in isolation, but which makes no direct contribution to the stability of the complex, will enhance the binding affinity by reducing the energetic penalty associated with forming the “binding-active” conformation of the peptide.

Even the smallest of proteins has a core made up of hydrophobic residues which provide a driving force for protein folding and which contribute significantly to the stability of the folded state. In contrast, the D-peptide lacks any sort of core, likely due both to limits in abilities of the natural amino acids to form such a core with the added constraints of the size of the peptide and the requirement to form an effective binding interface, and to the artificial environment in which the peptide was developed. However, it may be possible to enhance the stability of the bound state conformation of the D-peptide by building in some degree of hydrophobic core using non-standard amino acids.

The structure of the D-peptide was visually analyzed for possible substitutions that would fill in the core of the folded state. In particular, two types of modifications were considered. Firstly, residues with a $C_\alpha \rightarrow H_\alpha$ bond vector directed toward the center of the peptide would be good candidates for replacement with the corresponding α-methyl derivative. Secondly, residues with any $C_\beta \rightarrow H_\beta$ bond vector directed into the core would be a good choice for replacement with a β-methyl derivative. Since the side chains of all residues in the D-peptide are oriented toward the outside of the folded state (with the exception of the two disulfide-linked cysteines), substitution at any other positions would not result in the appropriate placement of the methyl group. Two residues satisfy each possible design motif. The H_α atoms of both Arg6 and Ala11 are located on the inside of the D-peptide, facing the empty core, and both Trp10 and Cys12 have H_β atoms similarly positioned (see Figure 4-9). Methyl substitutions at any of these four positions would likely add to the formation of a hydrophobic core, and possibly favor the folded, and active for binding, conformation of the D-peptide.

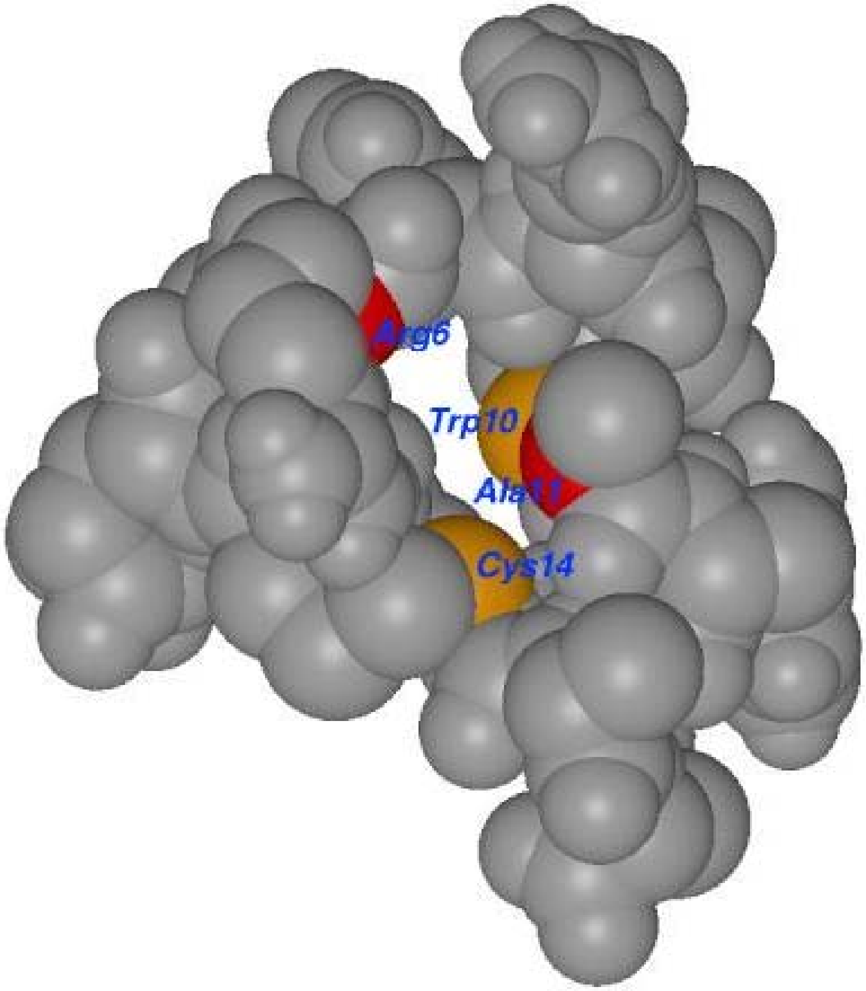


Figure 4-9: **Building a hydrophobic core into the D-peptide.** Several locations on the D-peptide were identified at which a methyl substitution is likely to fill in the core, stabilizing the bound state conformation. **Red:** A substitution at the α-carbon is likely to fill the core. **Orange:** A substitution at the β-carbon is likely to fill the core.

4.5 Discussion

The electrostatic contributions to binding of both the C34 peptide and the smaller D-peptide to the N36 trimeric coiled coil of gp41 are computed to be significantly unfavorable. In both cases, the interactions each ligand makes with the coiled coil in the bound state only compensate the desolvation penalty paid by the ligand by roughly half. The remaining half of the ligand desolvation cost, plus the full cost of desolvating the target receptor, results in a large unfavorable net electrostatic binding free energy. As for the optimal ligand the favorable interactions made in the complex are equal to twice in magnitude the cost of desolvating the ligand [92], it is clear that there is significant room for improvement of the electrostatic binding free energy in this system.

The initial optimization of the partial atomic charges on the side-chain atoms of each residue reveals an important aspect of this system — despite the natural ligand's unfavorable binding free energy of 8.4 kcal·mol^{-1}, only 4.3 kcal·mol^{-1} can be gained by optimizing the side-chain charges. The theoretical optimum ligand is required to have a net favorable binding free energy [78], but the constraints imposed by requiring the peptide backbone, as well as the cysteines involved in the disulfide linkage, to remain at their wild-type charges make this unattainable, even with non-physical charges allowed on the variable atoms. The addition of further constraints to limit the search to chemically reasonable charge distributions (requiring residues to have integral net charges between -1 and $+1e$, and not allowing any individual partial atomic charge to exceed $0.85e$ in magnitude) further reduces the possible improvement in binding free energy to 2.0 kcal·mol^{-1}.

While only a 2.0 kcal·mol^{-1} improvement in binding free energy is seen with the optimization of all residues, the majority of this effect is localized to two residues. Optimizing only the charges on Trp10 yields an improved binding free energy of 0.9 kcal·mol^{-1} with a net neutral charge, and optimizing only Trp12 produces a similarly neutral residue with a 0.7 kcal·mol^{-1} more favorable binding free energy than wild type. Both these optima are largely hydrophobic, even more so than an natural tryptophan. Trp10 makes no direct electrostatic interactions either with the receptor or with other groups on the D-peptide. Thus, the cost of desolvating the slightly polar aromatic system, and the more polar NH group, is not one worth paying, since no compensating favorable interactions are made. Trp12 makes a single hydrogen bond through the NH group, and no other direct electrostatic interactions. The optimum has partial charges on the NH very similar to the natural values, but is otherwise largely hydrophobic. Again, if no compensating interactions can be made, the cost of desolvating an aromatic system is a strictly unfavorable contribution to the binding free energy.

Since the residues lining the target binding pocket on the N36 trimer are largely hydrophobic, it is not entirely surprising that the optima are similarly hydrophobic, particularly for the residues (such as the two tryptophan side chains) that occupy the pocket. However, several polar and charged residues are located on the edge of the pocket, and it is reasonable to assume that a ligand as large as the D-peptide could interact with these residues. Lys574 and Arg579 in particular make no close contacts with the D-peptide in the crystal structure, but being surface residues would be expected to have large number of conformations accessible. In the crystal structure of the C34–N36 six-helical bundle, both these residues are in conformations which, when reconstructed in the context of the D-peptide–IQN17 structure, make much closer contact, without any major steric clashes. In several other structures of gp41, including a structure of the SIV-1 protein, Arg579 is always observed in the closer contacting conformation. This conformation is also the minimum energy structure determined by molecular mechanics, both with no electrostatic component, and with a distance-dependent dielectric Coulombic treatment of electrostatics. Lys574 is seen in several different conformations in various structures, including both that seen in the complex with the D-peptide and the closer contacting conformation seen in the six-helical bundle. In addition, Lys574 has been implicated in forming a salt-bridge with an acidic group on a small molecule inhibitor bound to the same position [71]. The minimum energy conformations of this lysine, in the context of the D-peptide, resemble the close contacting conformation seen in the gp41 structures, but are slightly different, with the ammonium group making even closer contact with the D-peptide. With a distance-dependent dielectric Coulombic electrostatic treatment, the minimum energy structures were significantly strained, and eliminating the electrostatic component produces minima which are similar in structure, but less internally strained. Including the electrostatic term may not be desirable in any case, as it may tend to over emphasize attractive electrostatics in the wild-type structure. The aim here, however, is to have close contacting polar residues for which the optimal electrostatic interactions will be *designed*; minimizing van der Waals and covalent energies will produce reasonable geometries, which will be easily accessible when the appropriate electrostatic interactions are designed into the ligand.

Electrostatically, the ligand binds to the modified structure somewhat worse than to the crystal structure. This is not surprising, since more of both the receptor and the ligand are buried on binding, and thus the desolvation penalties will be higher, and electrostatic interactions were not optimized in the conformational search. However, the higher amount of buried surface, and the increased favorable van der Waals interactions made in the modified structure, act to more than adequately counter balance the increased unfavorable electrostatic energy. Thus, this modified structure seems a wholly reasonable target for design.

Optimization of the D-peptide side-chain charges for binding to the structure designed for enhanced electrostatic interactions indicates that the procedure was generally successful. The possible improvement in binding is found to be 6.7 kcal·mol^{-1} when all atomic charges were allowed to vary freely, and 3.4 kcal·mol^{-1} with the imposition of constraints to ensure chemically reasonable charges. For the constrained optimum, this is a 1.4 kcal·mol^{-1} greater improvement than was seen previously. Perhaps more significantly, the individual residue optimizations also provide greater improvement. Both Trp10 and Trp12 are found to give improvements of approximately 1.5 kcal·mol^{-1}, with optimal net charges of $-1e$ on both, and both give improvements of over 1.0 kcal·mol^{-1} when constrained to be neutral. In addition, Arg6 and His7 show optimal net charges of $-1e$, both with improvements of 0.5 kcal·mol^{-1} over the wild-type charge distribution. The optimal tryptophan charge distributions are still quite hydrophobic, with a completely hydrophobic isostere of Trp10 showing a 0.8 kcal·mol^{-1} more favorable binding free energy than natural tryptophan, and an isostere of Trp12 which is hydrophobic everywhere except for the NH binding 0.5 kcal·mol^{-1} better. However, adding some degree of polarity to the five-membered ring of both the tryptophans, even when added in a highly constrained manner and maintaining an overall neutral charge, improves the binding by up to 1.8 kcal·mol^{-1} compared to the natural ligand.

While the optimization procedure did indicate that it may be possible to improve the binding free energy of the D-peptide by modifying the two tryptophans which occupy the target binding pocket, no clear indications of a chemical substitution to make were obtained. However, since the indole ring system is very rigid, many chemical modifications to the tryptophan all have very similar geometries. As a result, it is possible to take advantage of the pre-calculation of the desolvation and interaction matrices required for electrostatic optimization to rapidly screen a large database of indole derivatives. Since the geometries are all similar, simply replacing the charges of tryptophan with those of each derivative should give a reasonable estimate of the differences in the electrostatic binding free energy of the members of the library. The derivatives chosen were selected for a number of reasons. Firstly, since the replacement of a hydrogen with a fluorine generally results in a very small geometry change, but a significant change in polarity, the core of the library consisted of the replacement of every hydrogen on the indole ring with fluorine, in all combinations. Single hydrogen to chlorine and hydrogen to methyl substitutions were included for each hydrogen position, in order to lightly to sample these modifications, which involve larger geometry differences. In order to produce differences in the charge distribution of the aromatic π-system, several replacements of non-bridging carbon atoms with nitrogen were included, as were replacements of two or four non-bridging carbons of the six-membered ring of the indole with aliphatic carbons. Hydrogen to oxygen replacements, which significantly affect both the σ- and π-systems of the ring were considered for all aromatic CH groups, but only two were found to be stable under quantum mechanical analysis. Several further modifications of the aliphatic derivatives, with fluoro- and chloro- substitutions on the five-membered ring were also included, in order to sample differences in the polarity in this region with a more hydrophobic six-membered ring. This was done due to the observation in the optimal charge distributions of a largely hydrophobic six-membered ring but some tendency toward polarity in the five-membered ring. This library spans a relatively large region of the charge distributions possible for chemical derivatives of tryptophan.

As may have been expected given the results of the optimization, the top scoring ligands primarily contain aliphatic substitutions on the six-membered ring. For position 12, all of the top scoring ligands contain the same 4,5,6,7-tetrahydro derivative, which effectively depolarizes the six-membered ring while maintaining the NH group, which makes a hydrogen bond in the bound state, corresponding reasonably well to the optimal charge distribution. For position 10, both the 5,6-dihydro and the 4,5,6,7-tetrahydro derivatives are sampled in the best ligands, along with numerous derivatives of these. Interestingly, 1-fluoro-4,5,6,7-tetrahydro, 1-chloro-4,5,6,7-tetrahydro and 4,5,6,7-tetrahydro all have computed binding free energies within 0.1 kcal·mol^{-1} of each other, despite substantial differences in charge at the NH. Thus it seems that, although the NH group has the largest charges in the natural ligand and near zero charges in the optimum, this group contributes relatively little to the binding free energy. The smaller magnitude change of reducing the polarity of the six-membered ring is energetically much more significant.

The central benefit of the ligand scanning procedure revolves around the ability to rapidly rank a list of charge distributions. For a moderately sized database, or one such as is used here where the same library of derivatives can be used at multiple positions, the charge distributions can be derived at a relatively high level, such as fitting to quantum mechanical potentials. However, for a larger database, a more rapid method of determining charges may be useful. A rule-based method, replacing charges based on functional group substitutions would be particularly fast and efficient. For hydrogen to fluorine substitutions this procedure should be relatively straightforward. Considering the charges obtained by fitting to the quantum mechanical electrostatic potential on a few test systems, a general rule was devised where the replacement of an aromatic C–H bond with an aromatic C–F bond changes the charges by $+0.25e$ on the carbon and by $-0.25e$ on the "hydrogen". Starting with the PARSE charges on tryptophan, this transforms a C–H with a $-0.125e$ C and a $+0.125e$ H to a C–F with a $+0.125e$ C and a $-0.125e$ F. In order for such a method to be effective, the energy rankings obtained with the rule-based charges should correlate reasonably well with those obtained with more accurate charges. The correspondence does not have to be exact, but choosing a reasonable cutoff for selecting high ranking ligands at the lowest level should not eliminate high scoring ligands at the next level.

The results obtained using the rule-based charges correlate well with the more accurately computed binding energies, including those computed both with more exact charges and a more exact shape. The results using the approximate shape but the quantum mechanically derived charges correlate even better. Thus, even though the best ligand found was computed to improve the binding free energy by less than 1.0 kcal·mol^{-1} relative to the initial D-peptide, the procedure works well, rapidly selecting the best set of ligands without sacrificing the accuracy of the end result.

4.6 Conclusions

The N-terminal coiled coil of HIV-1 gp41 provides an attractive target for the design of inhibitors of viral–cell membrane fusion. The D-peptide inhibitor developed by Eckert *et al.* [43], which binds to a relatively hydrophobic pocket surrounded by several polar and charged residues with a net unfavorable electrostatic contribution to binding, further seemed a viable starting point for application of the electrostatic optimization procedure as part of a design protocol. However, the initial optimization results showed only small gain in binding free energy, particularly when constraints enforcing reasonable chemical limits were implemented.

Analysis of the structure pinpointed two charged residues on the periphery of the binding site which made no close contacts with the bound D-peptide, but which have been identified as making close interactions in other complexes with ligands of the coiled coil. Performing the optimization in the context of a receptor structure in which these residues were poised to make close contacts with the D-peptide provided much more significant improvements in binding affinity, with two tryptophans showing optimal improvements of over 1.0 kcal·mol^{-1}. Nonetheless, no clear chemical modifications to enhance binding affinity were apparent in the optimization.

A hierarchical procedure to computationally screen a library of derivatives of a starting molecule was developed around the charge optimization methodology as a means to screen a database of modified tryptophan replacements at the two D-peptide tryptophans buried on binding. With all combinations of substitutions at both positions, for a total of over 9000 distinct molecules, the greatest improvement found was computed to bind just under 1.0 kcal·mol^{-1} better than wild type. All the high scoring molecules contained derivatives that were significantly less polar than the original tryptophans — the moderate polarity of the aromatic system of natural tryptophan pays a energetic penalty for desolvation, but makes no interactions in the bound state to compensate.

While the results of the computations did not lead to predictions of major improvements in the binding affinity of the D-peptide, the methods outlined here, including the design toward an "electrostatically enhanced" target conformation and the ligand scanning procedure, can readily be applied to other systems. The binding pocket targeted by the D-peptide is largely hydrophobic, despite several peripheral polar residues, making highly favorable electrostatic interactions infeasible. In other systems with more highly polar binding sites, it is likely that the procedures described here would lead more easily to substantial improvements.

Chapter 5

Designing Improved Protein Inhibitors: HIV-1 Cell Entry Inhibitors Targeting the C-Terminal Heptad Repeat of gp41

Abstract

Previous work in our laboratory and others has resulted in the development of methodologies for the detailed analysis of the electrostatic contributions to binding affinities, as well as a procedure to calculate charge distributions that optimizes the electrostatic contribution to the binding free energy of a ligand of given geometry to a target receptor, in the context of a continuum model of solvation. We have applied these methods to the design of improved inhibitors of HIV-1 cell membrane fusion.

In order for HIV to infect a cell, the viral membrane must fuse with that of the target cell. This membrane fusion event is mediated by the viral membrane glycoprotein gp41, which is thought to undergo a conformational change involving the docking of three helices from the C-terminal region of gp41 against a trimeric coiled coil from the N-terminal region as a prerequisite for membrane fusion. Recently a protein inhibitor of membrane fusion (5-Helix) was developed that, by binding to an isolated C-terminal helix, blocks the formation of the fusogenic structure. A detailed energetic analysis of the binding of 5-Helix to a C-terminal helix was performed using the X-ray crystal structure of the core of the HIV-1 gp41 ectodomain as a structural

model. The overall electrostatic binding free energy was computed to be significantly unfavorable, and several residues on 5-Helix which make substantial contributions to binding, both favorable and unfavorable, were identified. The electrostatic affinity optimization methodology was applied to the side chains of 5-Helix, with the results showing that significant improvements in binding affinity are possible if the electrostatic contribution to the binding free energy is optimized. Several mutations accessible by experimental methods are suggested, with calculated improvements in binding affinity of up to 500-fold.

5.1 Introduction

As outlined in Section 4.1.1, an essential step in the infection of cells by human immunodeficiency virus (HIV) is the fusion of the viral membrane with that of the target cell [42]. This membrane fusion event is facilitated by gp41, an HIV viral envelope glycoprotein. It is believed that gp41 must undergo a major conformational change into a fusogenic form in order to mediate viral–cell membrane fusion. This conformational change involves the docking of a sequence of residues from the C-terminal region of three gp41 chains against a trimeric coiled coil pre-formed from the N-terminal region of the three chains, resulting in a "trimer-of-hairpins" with a six-helical bundle as a primary structural element [21, 42, 147, 161].

The pre-hairpin intermediate in which both the N-terminal coiled coil and the C-terminal region are exposed has been studied and validated as a target for inhibition of membrane fusion. Molecules that bind either to the N-terminal or to the C-terminal regions of gp41 have been shown to block the formation of the fusogenic trimer-of-hairpins conformation and thus inhibit membrane fusion. One class of inhibitors consists of peptides from the C-terminal and N-terminal regions of gp41 that are active inhibitors in membrane fusion assays and appear to act by these mechanisms [20, 75, 138]. Additional classes of inhibitors of HIV viral–cell membrane fusion, targeting both the N-terminal coiled coil and the C-peptide, have also been developed. These include both D-peptide [43] and small molecule [34, 46, 167] inhibitors which bind to the N-terminal coiled coil, as well as protein constructs based around the coiled coil

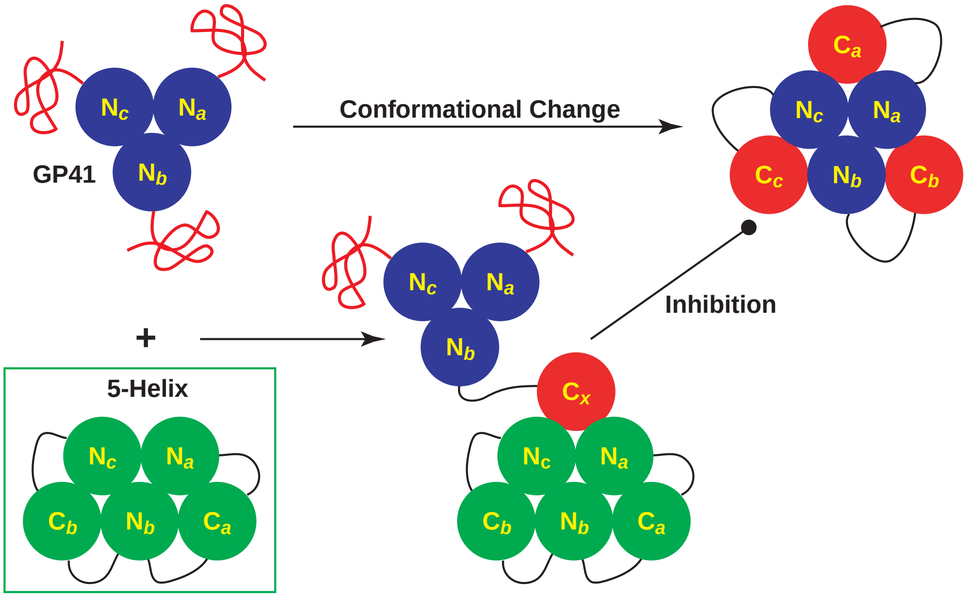


Figure 5-1: **Inhibition of the gp41 conformational change by 5-Helix.** 5-Helix binds to the C-terminal region of gp41, preventing the docking of the C-terminal helix against the N-terminal coiled coil required for the formation of the fusogenic structure.

which bind to the C-terminal region [41, 97, 127].

One protein inhibitor of HIV cell entry that has recently been developed is 5-Helix [127]. This construct consists of five helical sequences, three with a sequence equivalent to the N-terminal region of gp41 which forms a trimeric coiled coil, and two with a sequence equivalent to the C-terminal region of gp41, which dock against the coiled coil. A six-helical bundle consisting of three N-terminal and three C-terminal peptides is known to be a stable structure, being a key characteristic of the fusogenic conformation of gp41. 5-Helix is able to form such a structure by binding to a free C-terminal peptide, and in doing so sequesters the C-terminal region away from the N-terminal coiled coil of native gp41, thus inhibiting the conformational change in gp41 which is required for viral–cell membrane fusion (see Figure 5-1). In both cell–cell fusion and viral infectivity assays, 5-Helix has been determined to inhibit membrane fusion with a low nanomolar IC_{50} [127].

Over the past several years, our laboratory has developed a set of methodologies for analyzing the electrostatic energetics of protein–ligand binding and of protein stability using a continuum model of solvation. These include methods both for the analysis of structures and for the design of structures with improved affinity and specificity properties. Component analysis provides a dissection of all electrostatic contributions to binding (or folding) into an additive set of contributions from various groups in the system (amino acid side chains, backbone carbonyl and amino groups, etc.) considering solvation effects as well as direct electrostatic interactions [69]. Electrostatic affinity optimization provides a framework for varying the partial atomic charges on a ligand so as to minimize the electrostatic contribution to the binding free energy [23, 77–80, 92–94]. Here these methods were applied to the analysis of the binding of 5-Helix to an isolated C-terminal helix with the explicit goal of identifying regions of 5-Helix that are not fully complementary to the C-peptide and of predicting mutations to 5-Helix with higher computed affinity to the C-peptide. Several mutations to 5-Helix that are predicted to improve binding affinity resulted from the analysis.

5.2 Methods

Preparation of structures. No crystal structure of 5-Helix alone or bound to C-peptide was available for this work. However, 5-Helix consists of three 40-residue N-terminal sequences, and two 38-residue C-terminal sequences, linked by five-residue, glycine-rich linkers. The crystal structure of the gp41 ectodomain core region solved by Chan *et al.* (Protein Data Bank [125] ID 1aik) [21] should be an excellent model of the complex because it consists of a six-helical bundle of three 36-residue N-terminal sequences and three 34-residue C-terminal sequences. These sequences reside wholly inside the sequence of 5-Helix, and thus the use of this structure as a model for 5-Helix seems reasonable; only four helical residues from each chain are not considered (three from the N-terminal end, one from the C-terminal end), and the linker is not expected to play a major role in binding.

Hydrogen atoms were added using the HBUILD facility [14] of the CHARMM computer program [11], using the PARAM19 parameter set [11] with the addition of aromatic hydrogens on Phe, Tyr, Trp and His for consistency with the parameters used in the continuum electrostatic calculations. Visual analysis of structure suggested no reason for the ionizable residues to be in their non-standard states, and thus all histidines were left in their neutral state, and all acidic residues were left charged. Binding was considered as the rigid binding of a C34 helix to the 5-Helix model. While this is likely to be an accurate representation of 5-Helix, which forms an exceptionally stable structure in isolation (5-Helix remains helical up to 100° C in the absence of denaturant, and does not unfold until nearly 90° C in 3.7 M GuHCl [127]), the C-peptide is believed to be disordered in the unbound state. This will affect the desolvation penalties for C-peptide residues in the component analysis which might be somewhat underestimated by the pre-formed structure. However, because the C-peptide desolvation does not enter the charge optimization, these results will be unaffected by the C-peptide pre-configuration.

Continuum electrostatic calculations. All continuum electrostatic calculations were done using a locally modified version of the DELPHI computer program [55, 57, 134, 136] to solve the linearized Poisson–Boltzmann equation. An internal dielectric constant of 4 and an external dielectric constant of 80 was used unless otherwise specified, and the ionic strength was set to 0.145 M. The molecular surface (used to define the dielectric boundary) was generated using a probe radius of 1.4 Å, and an ion exclusion (Stern) layer [9] of 2.0Å was applied around all molecules. Protein partial atomic charges and radii were taken from the PARSE parameter set [140] with a few minor changes. Charges on the bridging ring carbons of tryptophan were assigned to $0e$, charges for proline and for disulfide bridged cysteine residues were taken from the PARAM19 parameter set [11], and the charges from glutamate and lysine side chains were used for charged C and N termini respectively. Binding and solvation free energy calculations were performed using two-step focusing boundary conditions on a 191×191×191 unit cubic grid, in which the longest dimension of the molecule occupied first 23% and then 92% of one edge of the grid (final grid spacing of 0.31 Å). Boundary potentials for the more highly focused calculation were obtained from the lower focused calculation, and Debye–Hückel potentials were used at the boundary of the lower run. Calculations for component analysis and to determine the matrix elements for electrostatic optimization were done using a three-step focusing procedure on a 129×129×129 unit grid, with the molecule occupying 23%, 92%, and finally 184% of the grid (final grid spacing of 0.23 Å). For the highest resolution calculations, the grid was centered on the region of interest, and interactions involving groups falling outside of this grid were computed from the 92% fill calculation. All calculations were averaged over ten translations of the structure on the grid in order to minimize artifacts from the the placement of the point charges and molecular boundaries onto the finite difference grid.

Electrostatic affinity optimization, in which the "ligand" charge distribution is allowed to vary so as to produce the most favorable electrostatic binding free energy, were performed as previously described [23, 77–80, 92–94] using locally written software. In this case, 5-Helix was treated as the ligand and C-peptide as the receptor. Singular value decomposition [119, 143] was used to remove all basis vectors with singular values of less than 10^{-5} of the largest singular value, or for which the standard error over ten translations was greater than 25% of the value. Typically this involved the removal of 773 out of 990 basis vectors; the majority of residues significantly removed from the interface pay almost no desolvation, leading to a large number of very small eigenvalues in the desolvation matrix. Basis vectors in the null space were allowed to be populated only when required to satisfy imposed constraints, and were penalized by a harmonic penalty with a coefficient of 10.0 kcal·mol^{-1}·e^{-2} in the optimization, but not in the final energy evaluation. Constrained optimizations were performed using the computer program LOQO [133, 154, 155]. Typical constraints applied to all optimizations were that all residues must have an integral net charge, that no residue may have a net charge of greater that $1.0e$ in magnitude, and that no individual partial atomic charge may have a charge of greater than $0.85e$ in magnitude. These constraints were chosen to limit the optimization to regions of charge space reasonably attainable in the context of natural amino acids.

Design and modeling of mutations. Mutant structures were built using the CHARMM computer program [11] with the PARAM22 all-atom parameter set [100] and using a distance dependent dielectric of $\epsilon = 4r$ for Coulombic electrostatic interactions. For each mutated residue, the lowest energy conformation was found using the following procedure. Each side-chain torsion angle was sampled at 30° intervals, followed by 100 steps of adapted-basis Newton–Rhapson (ABNR) minimization of the side chain with the rest of the protein structure held fixed. In cases where van der Waals clashes were observed in the minimum energy structure by energetic and visual analysis, the side chains involved in the clash were also allowed to move during the minimization. In all, four additional side chains on the C-peptide were allowed to move: Glu 22, Ser 23, Gln 27, and Glu 31. Before any further computations were performed, all seven mobile side chains (the three variable positions on 5-Helix and the four mobile residues on the C-peptide) were minimized to convergence (typically around 1000 steps). Repeating the same procedure with wild-type 5-Helix produces a structure very similar to the crystal structure, and minimization from the crystal structure geometry produces the same structure as the conformational sampling procedure described above.

Calculation of free energies of binding. Free energies of binding in solution were calculated by adding the difference in solvation free energies of the complex and the two ligands to the vacuum binding free energy. *In vacuo* binding free energies were calculated using CHARMM [11] with the PARAM22 all-atom force field [100]. Solvation energies were calculated using a Poisson–Boltzmann/Surface Area (PBSA) procedure, using PARSE radii and charges [140], with the same changes as detailed previously. The electrostatic component was calculated using a locally modified version of the DELPHI computer program [55, 57, 134, 136] as described above. The non-polar component was calculated from the solvent accessible surface area using the relationship, $\Delta G = \gamma A + b$ with $\gamma = 5.4$ cal·mol^{-1}·Å^{-2} and $b = 0.920$ kcal·mol^{-1} [140]. Solvent accessible surface areas (using a probe radius of 1.4 Å) were calculated using CHARMM [11].

5.3 Results

5.3.1 5-Helix–C-peptide electrostatic binding free energy

To gain an initial perspective on the role of electrostatic interactions in the 5-Helix–C-peptide complex, the electrostatic contributions to the free energy of 5-Helix binding to a single C34 helix were computed. 5-Helix pays a 17.6 kcal·mol^{-1} dehydration penalty, and the C34 helix pays a 19.5 kcal·mol^{-1} dehydration penalty, but they only recover 10.5 kcal·mol^{-1} of favorable intermolecular interactions, resulting in a net electrostatic contribution to binding of +26.7 kcal·mol^{-1}. Thus, electrostatics are significantly destabilizing to complex formation in this system.

5.3.2 Electrostatic contributions to 5-Helix binding

In order to gain further insight into the basis for the unfavorable contribution that electrostatics make to the free energy of association in this system, an electrostatic component analysis was carried out on the 5-Helix–C-peptide complex. As described in Section 2.4.1, previous work has described the methodology by which the contribution of various groups in a protein to the electrostatic binding free energy may be calculated [69]. For the purpose of this work, each residue was considered as the

Energy Term	Energy
$N36_{abc}$ Desolvation	$+23.59$
$N36_{abc}$ Indirect	-5.80
$C34_a$ Desolvation	$+0.03$
$C34_a$ Indirect	$+0.01$
$C34_b$ Desolvation	$+0.02$
$C34_b$ Indirect	$+0.01$
$N36_{abc}$–$C34_a$ Indirect	-0.08
$N36_{abc}$–$C34_b$ Indirect	-0.18
$C34_a$–$C34_b$ Indirect	-0.01
Total 5-Helix Desolvation	$+17.59$
$C34_x$ Desolvation	$+24.51$
$C34_x$ Indirect	-4.97
Total $C34_x$ Desolvation	$+19.54$
$N36_{abc}$–$C34_x$ Interaction	-12.24
$C34_a$–$C34_x$ Interaction	$+0.88$
$C34_b$–$C34_x$ Interaction	$+0.88$
Total Interaction	-10.47
Net Electrostatic Energy	$+26.65$

Table 5-1: **Helical contributions to the 5-Helix–C34 binding free energy.** The contributions of the components of each helix (in kcal·mol^{-1}) to the electrostatic binding free energy of 5-Helix to an isolated C34 helix are detailed.

union of three chemical groups: backbone carbonyl, backbone amino and side chain. For each group all of its energetic contributions to the binding free energy were calculated. These are: (i) the desolvation penalty, which is the energetic cost of moving the group from the region of low dielectric in the unbound state to the (larger) region of low dielectric in the bound state; (ii) the indirect interactions, which are the energetic change in interactions between different groups in the same molecule when the dielectric boundary is changed from that of the unbound state to that of the bound state (intramolecular interactions); (iii) the direct interactions, which are the interactions between a group on one molecule and groups on the other molecule in the bound state (intermolecular interactions).

The breakdown of the energetic components on a helix-by-helix basis is detailed in Table 5-1. The total desolvation penalty for a set of groups (such as a helix) is the sum of desolvation contributions for the component groups, while those intramolecular interactions between the component groups within a set sum to give the indirect interaction of a single set. Intramolecular interactions between component groups of different sets are grouped together into an "indirect" interaction between each pair of sets. The large desolvation penalty of 5-Helix results almost exclusively from contributions from the N36 trimer, including +23.6 kcal·mol^{-1} of direct group desolvation penalties, and −5.8 kcal·mol^{-1} of indirect interactions between groups within the N36 trimer. This is consistent with the C34 peptide binding in a groove between a pair of N36 helices. The desolvation penalty of the C34 peptide can be broken down into +24.5 kcal·mol^{-1} of direct group desolvation terms and −5.0 kcal·mol^{-1} of indirect interactions between groups within the C34 helix. The C34 helices of 5-Helix pay essentially no desolvation penalty upon binding, but each make slightly unfavorable interactions of +0.9 kcal·mol^{-1} with the bound C34. The majority of the total interaction free energy of −10.5 kcal·mol^{-1} consists of direct interactions (−12.2 kcal·mol^{-1}) between the N36 trimer and the bound C34.

The interactions between the C34 helices of 5-Helix and the additional bound C34 were considered in more detail. The 0.9 kcal·mol^{-1} repulsion between each helix could be due either to general electrostatic repulsions, since each C34 helix has a net charge of −6, or to a few specific unfavorable interactions. Each C34 helix in 5-Helix contains eight acidic residues which could make unfavorable interactions with the additional C34 helix. However only two direct interactions between side chains are greater in than 0.1 kcal·mol^{-1} in magnitude; these are the two symmetry related interactions between Glu22 on one helix and Glu31 on another, each of which is unfavorable by 0.2 kcal·mol^{-1}. Thus, the unfavorable interaction of the C34 helices seems to be a general electrostatic effect spread out over the many acidic residues of each helix.

A detailed analysis of the contributions to binding of various groups on the N36 trimer was also done, and the most significant contributions are outlined in Table 5-2.

Helix	Group	ΔG_{mut}	ΔG_{desolv}	ΔG_{indir}	$\Delta G_{int}^{C34_x}$	$\Delta G_{int}^{Specific}$	
$N36_c$	Asn10	+2.45	+1.67	−0.70	+1.49	+1.61	Glu22
$N36_a$	Glu16	+2.42	+2.52	+0.04	−0.14	−1.49	Gln24
$N36_c$	Arg13	−1.75	+0.81	−0.77	−1.79	−1.67	Glu22
$N36_c$	CO3	−1.37	+0.79	−0.88	−1.29	−1.22	Asn30
$N36_a$	Gln19	+1.03	+1.45	−0.12	−0.30	–	–
$N36_a$	CO5	−1.01	+0.43	−0.58	−0.86	−0.96	Gln27

Table 5-2: **Most significant group contributions from 5-Helix.** All components of 5-Helix with mutation energies greater than 1.0 kcal·mol^{-1} in magnitude are shown, identified both by the helix on which the group is located and by the group identity. All energies are in kcal·mol^{-1}.

Six components have mutation terms greater than 1 kcal·mol^{-1} in magnitude: three side chains make overall unfavorable contributions to binding (two by more than 2 kcal·mol^{-1}), one side chain contributes favorably to binding, and two carbonyls make overall favorable contributions to binding.

Asn10 on $N36_c$ pays a significant desolvation penalty only partially offset by indirect interactions and also makes an unfavorable direct interaction of +1.6 kcal·mol^{-1} with Glu22 on C34, resulting in a net contribution to binding of +2.5 kcal·mol^{-1}. Glu16 on $N36_a$ also pays a significant desolvation penalty, but makes almost no overall indirect or direct interactions, despite a favorable direct interaction of −1.5 kcal·mol^{-1} with Gln24. Gln19 on $N36_a$ pays a significant desolvation penalty, but makes little back in indirect interactions, despite a favorable indirect interaction with Gln18 on $N36_c$ of −0.7 kcal·mol^{-1}, and, since it makes almost no direct interactions, is unfavorable over all.

Arg13 on $N36_c$ regains most of its desolvation penalty from indirect interactions, and makes a strong favorable interaction with Glu22, resulting in an overall contribution to binding of −1.8 kcal·mol^{-1}. Two carbonyls gain more than all of their direct desolvation energy back from indirect interactions, and also make direct favorable interactions with an amide side chain on C34, thus having a net favorable contribution to binding (−1.4 and −1.0 kcal·mol^{-1}). The overall picture thus seems to be that

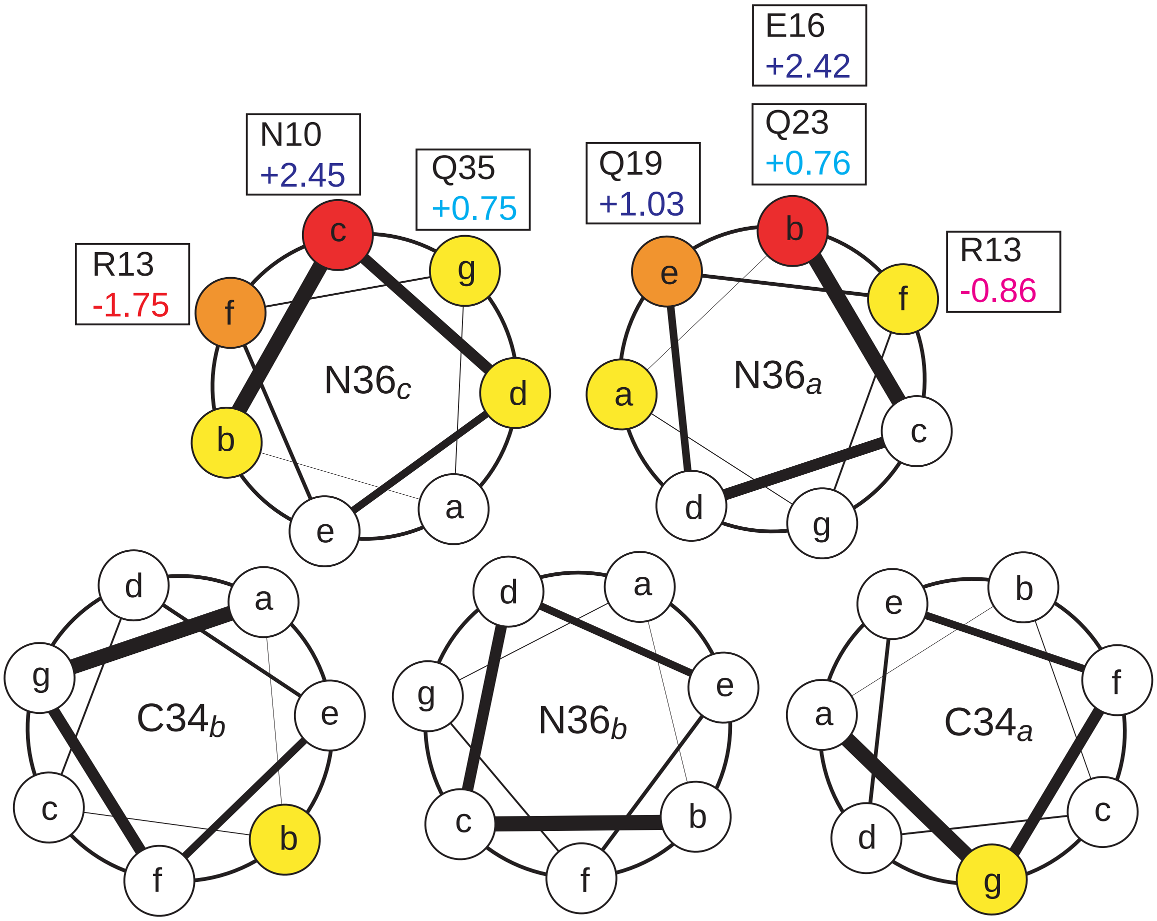


Figure 5-2: **Electrostatic contributions of 5-Helix side chains to C34 binding.** The positions of the helical wheel figure are colored according to the mutation energy of side chains at that position. 🟥 indicates a residue with a mutation energy of more than 2.0 kcal·mol^{-1} in magnitude, 🟧 indicates a residue with a mutation energy of at least 1.0 kcal·mol^{-1} in magnitude, and 🟨 indicates a residue with a mutation energy of over 0.20 kcal·mol^{-1} in magnitude. The value of the mutation free energy (in kcal·mol^{-1}) for all side chains for which this value is above 0.50 kcal·mol^{-1} in magnitude is also shown. A helical position is colored if *any* residue at that position has a substantial contribution.

polar and charged residues play a substantial role in the binding of 5-Helix to the C34 helix, but that there is considerable room for improvement.

The spatial arrangement of side chains with significant mutation terms — the relative free energy of binding of the natural complex and that of a hypothetical mutant complex in which the side chain of the residue in question, and only that side chain, is replaced with a hydrophobic isostere — is shown in Figure 5-2. Negative mutation free energies correspond to side chains which contribute favorably to binding, relative to a hydrophobic replacement, while positive mutation free energies indicate residues which contribute unfavorably to binding, again relative to the hydrophobic isostere. The most strongly contributing positions are all located directly at the binding interface: all residues with mutational energies greater than 1.0 kcal·mol^{-1} in magnitude are located at positions b and e on N36$_a$ and positions c and f of N36$_c$, using a helical wheel representation for the structure, and all residues with mutational energies greater than 0.5 kcal·mol^{-1} are found either at these positions or at position f on N36$_a$ and position g on N36$_c$. A few residues on the layer immediately removed from the interface, as well as several glutamates on the C34 helices have mutational energies in the range of 0.2 to 0.5 kcal·mol^{-1}.

5.3.3 Optimization of 5-Helix binding

While component analysis is very useful in identifying residues which contribute favorably or unfavorably to the binding free energy, and thus can suggest places where mutations are likely to stabilize or destabilize the complex, such an analysis can not give very much insight into what mutations (besides mutation to a hydrophobic residue) should be made. In addition, by considering only the wild-type charge distributions, component analysis has no predictive power in suggesting non-polar residues whose replacement by a polar or charged residue may enhance binding affinity. Electrostatic affinity optimization overcomes these shortfallings. The procedure involves varying the charge distribution on one member of a binding complex so as to obtain the best possible electrostatic binding free energy. Constraints limiting maximal atomic charges, total residue charges, and limiting the variable charges to a subset of the total ligand charge distribution are all easily incorporated into the optimization procedure.

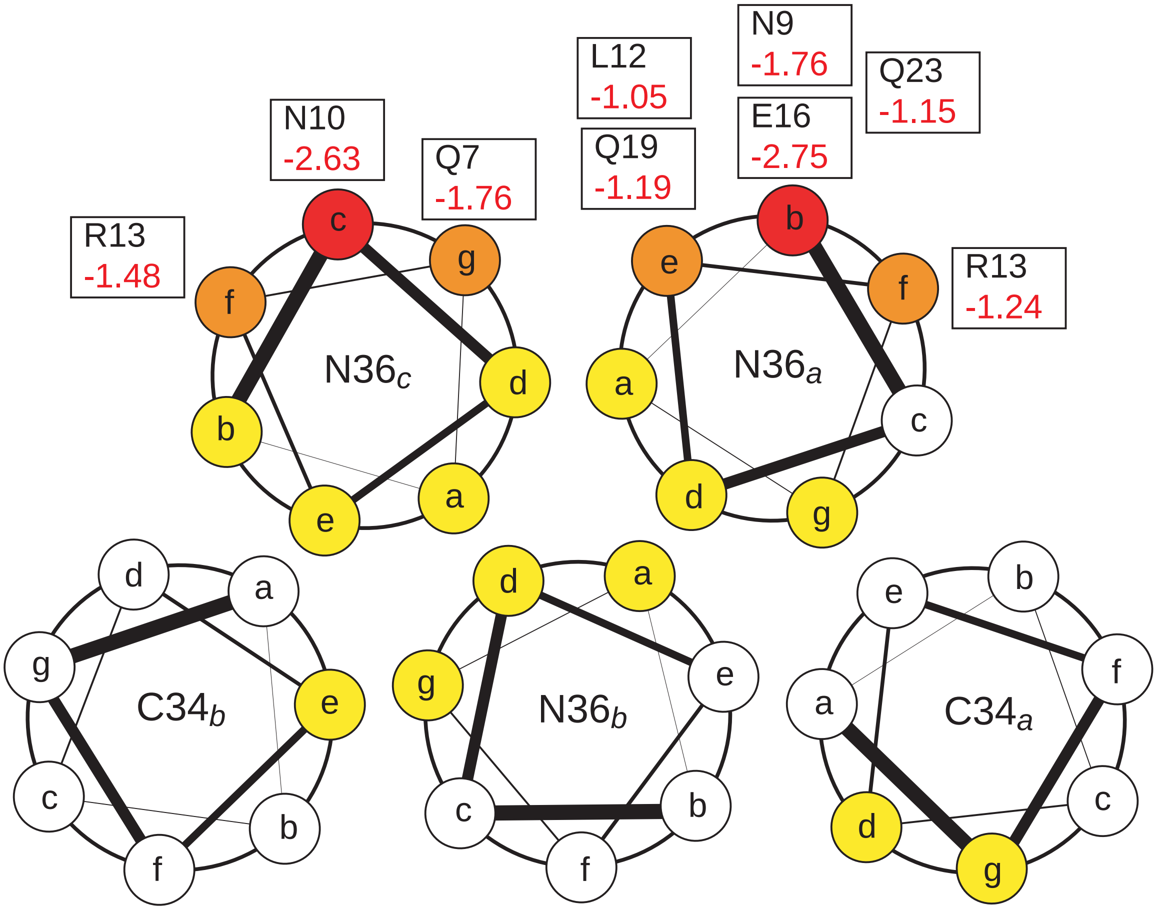


Figure 5-3: **Optimization of the contribution of 5-Helix side chains to C34 binding.** The positions of the helical wheel figure are colored according to the optimal improvement in bining free energy of side chains at that position. 🟥 indicates a residue with an optimized binding contribution at least 2.0 kcal·mol^{-1} better than wild type, 🟧 indicates a residue with an improvement of at least 1.0 kcal·mol^{-1} over wild type, and 🟨 indicates a residue with an improvement of at least 0.5 kcal·mol^{-1}. The value of the improvement (in kcal·mol^{-1}) is detailed for all positions with imporovements above 1.0 kcal·mol^{-1}. A helical position is colored if *any* residue at that position show a substantial improvement.

In order to investigate the possibility of mutations which may enhance 5-Helix's affinity for the C-terminal sequence of gp41, the partial atomic charges on each side chain of 5-Helix were varied in turn, keeping all other residues at their wild-type identities, so as to optimize the free energy of 5-Helix binding to an isolated C34

Helix	Residue	$\Delta\Delta G_{mut}$	$\Delta\Delta G^{-1}_{opt}$	$\Delta\Delta G^{0}_{opt}$	$\Delta\Delta G^{+1}_{opt}$	Mutation
$N36_a$	Glu16	−2.38	−1.50	−2.47	−2.76	Gln
$N36_c$	Asn10	−2.45	+8.38	−2.62	−0.64	Leu
$N36_c$	Gln7	+0.31	+6.22	−1.76	+3.71	–
$N36_a$	Asn9	−0.33	+1.20	−0.61	−1.76	His/Lys
$N36_c$	Arg13	+1.70	+1.97	−0.37	−1.48	–
$N36_a$	Arg13	+0.82	−0.16	−0.73	−1.25	–
$N36_a$	Gln19	−1.03	+2.24	−1.19	+1.62	Hydrophobic
$N36_a$	Gln23	−0.76	−0.25	−0.82	−1.15	–
$N36_a$	Leu12	0.00	+4.94	−0.39	−1.06	Positive

Table 5-3: **Greatest improvements in binding on optimization of 5-Helix side chains.** All 5-Helix side chains with optimal improvements in binding free energy of over 1.0 kcal·mol^{-1} (relative to wild type) are displayed, with the improvement in binding free energy with the total residue charge constrained to −1, 0, and +1*e* listed, as well as the relative energy of a hydrophobic isostere. For several residues, suggested properties or identities of amino acid substitutions likely to improve binding are also listed.

helix. The results of these optimizations are summarized in Figure 5-3.

Nine residues gave optimal improvements over wild type of over 1.0 kcal·mol^{-1}, and two gave improvements of more than 2.0 kcal·mol^{-1}. All these most significant residues were located directly at the binding interface, at the b, e and f positions of helix $N36_a$ and at the c, f and g positions of helix $N36_c$. These residues consist of all types of functionalities, positively and negatively charged, polar, and hydrophobic. In addition, there were many positions one or two layers removed from the interface which gave improvements of between 0.5 and 1.0 kcal·mol^{-1} over wild type on optimization.

The nine residues which gave optimal improvements of over 1.0 kcal·mol^{-1} were examined in more detail. The charges on each of these side chains were optimized constraining the total charge on the residue to be −1, 0, or +1*e*, and the binding energetics examined (see Table 5-3). In addition, the optimal atomic charges on all residues which gave optimal improvements of over 1.5 kcal·mol^{-1} were analyzed (see Figure 5-4).

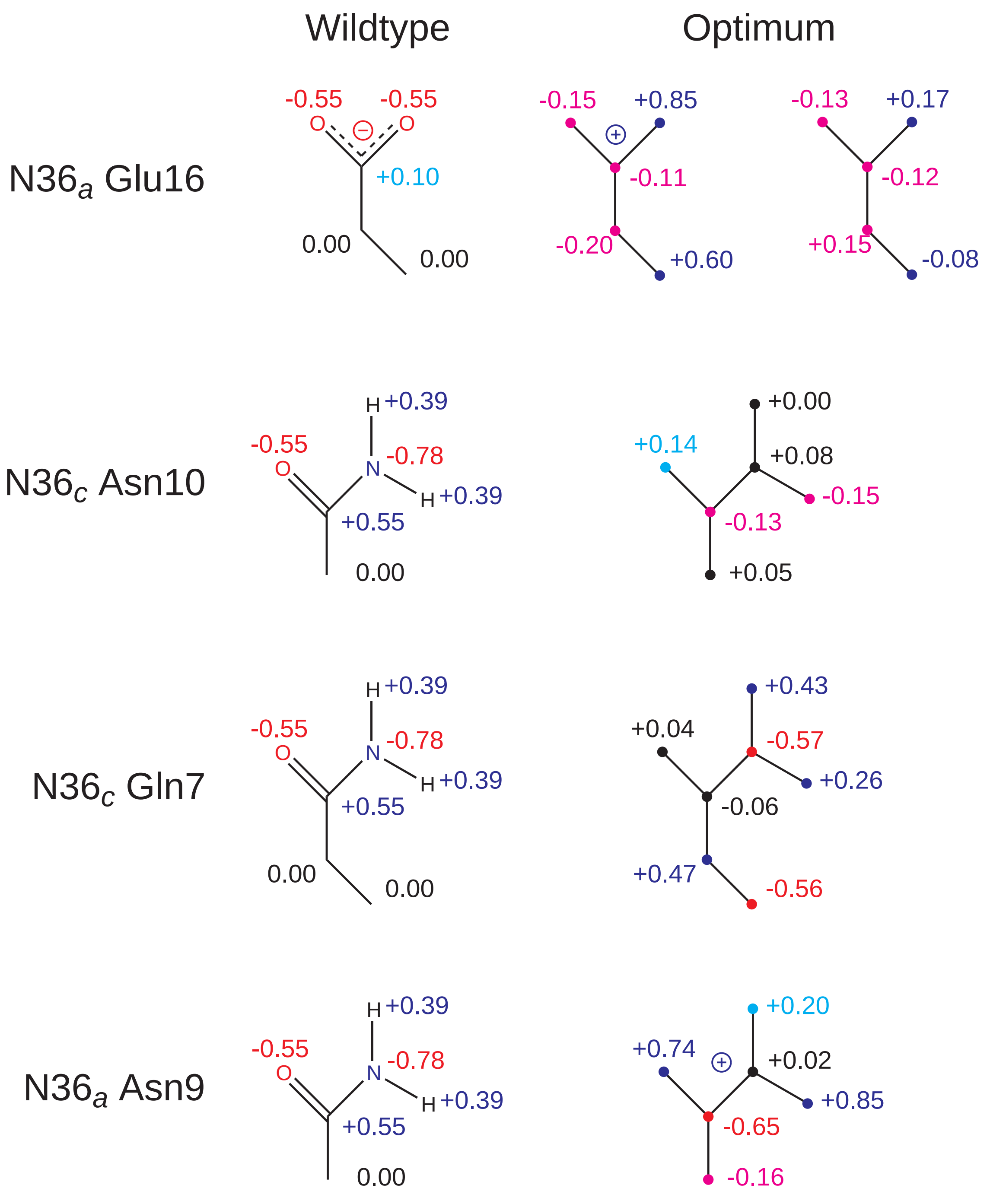


Figure 5-4: **Partial charges on side chains with greatest improvements on optimization.** The optimal partial charges at the side-chain atoms of residues whose optimization shows the greatest improvement in binding free energy are displayed. For each residue the wild-type charge distribution is shown along with the optimal charge distribution.

Glu16 on the N36_a helix was seen in the component analysis to contribute unfavorably to binding, relative to a hydrophobic isostere, by 2.4 kcal·mol^{-1}. Optimizing the partial atomic charges on this residue gives an improvement in binding affinity (relative to wild type) of 2.8 kcal·mol^{-1} when the net charge on the residue is $+1e$, and 2.5 kcal·mol^{-1} when the residue is neutral. However, when the residue is fixed at a total charge of $-1e$, as it is in the wild type, the optimal improvement is only 1.5 kcal·mol^{-1}. Thus it seems likely that a mutation at this position would enhance the binding affinity. While the optimal net charge is $+1e$, fixing the net charge to $0e$ costs only 0.3 kcal·mol^{-1} in the optimization, and replacing the residue with a hydrophobic isostere results in a binding free energy only 0.4 kcal·mol^{-1} below optimal. A likely replacement which preserves the geometry of the wild-type residue, but is neutral rather than negatively charged, is glutamine.

N36_c Asn10 was also seen to contribute unfavorably to binding, relative to a hydrophobic isostere, by 2.4 kcal·mol^{-1}. While optimization of the partial atomic charges on this residue gives an improvement in binding free energy of 2.6 kcal·mol^{-1} over wild type when the residue is neutral, fixing the net charge at $+1e$ reduces the improvement to 0.6 kcal·mol^{-1}, and fixing the net charge at $-1e$ leads to an optimal binding free energy 8.4 kcal·mol^{-1} *worse* than wild type. The optimal charges on the neutral side chain are all very low in magnitude ($\max|q_i| = 0.15e$), suggesting strongly that a hydrophobic group at this position is most favorable for binding. Leucine, with the same number of heavy atoms as asparagine and a similar topology, would seem to be a good replacement.

Gln7 on the N36_c is a particularly interesting residue. The component analysis results show that this residue contributes favorably to binding by 0.3 kcal·mol^{-1} relative to a hydrophobic isostere, suggesting that polar interactions are important at this position. However, the affinity optimization shows a strong preference for a neutral residue; fixing the net charge at $-1e$ leads to an optimal binding free energy 6.2 kcal·mol^{-1} worse than wild type, and fixing the net charge at $+1e$ gives an optimum whose binding free energy is 3.7 kcal·mol^{-1} worse than wild type. The optimal charge distribution with no overall charge, on the other hand, has a binding free energy 1.8 kcal·mol^{-1} better than wild type. Examination of the partial atomic charges of the optimum shows a remarkable similarity to wild type at the amide NH$_2$ group, but near hydrophobic charges at the carbonyl. Significant charges are also found at the C$_\beta$ and C$_\gamma$ atoms in the optimum, although constraining these charges to zero reduces the optimal binding free energy only slightly, to a 1.4 kcal·mol^{-1} improvement. Unfortunately, none of the twenty common amino acids have a charge distribution similar to this, although an unnatural amino acid substitution here may substantially improve the binding affinity.

N36$_a$ Asn9 contributes only slightly unfavorably (0.3 kcal·mol^{-1}) to binding relative to a hydrophobic isostere. Optimization of the partial atomic charges of this residue leads to an improvement in binding affinity of 1.8 kcal·mol^{-1}, with a net charge of $+1e$. With a net charge of 0, the optimal improvement is reduced to 0.6 kcal·mol^{-1}, and when the net charge is $-1e$ the optimal binding free energy is 1.2 kcal·mol^{-1} worse than wild type. These results indicate a strong preference for a positively charged residue at this position. Two substitutions are thus suggested. Histidine is of similar shape and size to asparagine, and has a pK$_a$ only slightly below 7, thus being quite easy to protonate at neutral pH. Lysine, while significantly different in structure to asparagine, is flexible and could possibly adopt a favorable conformation if placed at this position.

Five additional residues show optimal improvements in binding free energy between 1.0 and 1.5 kcal·mol^{-1}, relative to the wild-type side chain. Arg13 on both the N36$_a$ and N36$_c$ helices contribute favorably to binding relative to hydrophobic isosteres by 0.8 and 1.7 kcal·mol^{-1}. The results of the electrostatic affinity optimization at these positions show that 1.3 to 1.5 kcal·mol^{-1} can be gained from varying the charge distribution but also indicate that a positive charge, as found in the wild type arginine, is strongly favored. Gln19 on N36$_a$ contributes unfavorably to binding by 1.0 kcal·mol^{-1} relative to a hydrophobic isostere, and the optimal improvement in binding free energy is only slightly better than this (1.2 kcal·mol^{-1}, for a neutral residue). The optimal binding free energy for varying this residue is worse than wild type when the net charge is constrained to either $-1e$ (by 2.2 kcal·mol^{-1}) or to $+1e$ (by 1.6 kcal·mol^{-1}). A hydrophobic residue at position 19 on N36$_a$ would thus seem to favor binding. Mutation of N36$_a$ Gln23 to a hydrophobic isostere is computed to improve binding by 0.8 kcal·mol^{-1}. The affinity optimization results at this position show a slight preference for a positive residue (1.2 kcal·mol^{-1} improvement) over a neutral residue (0.8 kcal·mol^{-1} improvement), but show limited room for improvement with a negative residue (0.2 kcal·mol^{-1}). No information can be gleaned from component analysis for hydrophobic residues such as Leu12 on helix N36$_a$. However, the optimization shows that only a slight improvement (0.4 kcal·mol^{-1}) can be made in the context of a neutral residue, but that with a positively charged residue as much as 1.1 kcal·mol^{-1} could be gained in the binding free energy. A negatively charged residue is excluded from this position, with the optimal binding energy in this case being 4.9 kcal·mol^{-1} worse than wild type.

5.3.4 Binding energetics of 5-Helix mutants

Four mutations to common amino acids at three positions on 5-Helix were suggested by the optimization procedure. Model structures of the proposed mutants were constructed and the binding energetics analyzed in detail. In addition to each single mutant, all combinations of two and three mutations were also considered. Binding free energies were calculated from the difference in solvation free energies of the complex and the isolated components, combined with the computed rigid body *in vacuo* binding energy. All residues involved in the mutations, either directly or due to close contacts in any structure, were allowed to minimize their geometries, but all other residues and the backbone were kept in their crystal structure positions — in all, four residues on the C34 peptide, as well as the three variable positions on 5-Helix, were

Mutations	$N36_a$-9	$N36_a$-16	$N36_c$-10	$\Delta\Delta G^{WT}_{binding}$	K_d^{WT}/K_d^{mut}
0	N	E	N	0.00	1.0
1	K	E	N	−0.90	4.6
1	H	E	N	−1.13	6.7
1	N	Q	N	−1.45	12
1	N	E	L	−1.69	17
2	K	Q	N	−1.90	25
2	K	E	L	−2.73	100
2	H	Q	N	−1.98	28
2	H	E	L	−2.95	150
2	N	Q	L	−2.95	150
3	K	Q	L	−3.51	380
3	H	Q	L	−3.72	530

Table 5-4: **Relative binding free energies of mutant structures.** The computed differences in binding free energy (relative to wild type, in kcal·mol^{-1}), including both electrostatic and non-electrostatic terms, are shown for all modeled mutant structures. Also listed is the equivalent improvement in K_d, computed at 298 K.

mobile. The wild-type sequence was subjected to the same procedure so as to make comparisons of the energetics more meaningful. The results of the mutation design studies are summarized in Table 5-4.

The single mutants all have calculated binding affinities better than wild type, ranging from five-fold to seventeen-fold improvement in the dissociation constant. The two weakest effects were seen for the mutation of $N36_a$ Asn9 to lysine or histidine, and the largest improvement resulted from the mutation of $N36_c$ Asn10 to leucine. The double mutants have calculated improvements in affinity of 25-fold to 150-fold, again with the largest effects predicted for the mutants including the $N36_c$ Asn10→Leu mutation. The two triple mutants have predicted improvements in binding affinity of 380-fold and 530-fold. The improvements in binding affinity for each mutation are roughly additive, with only about 0.5 kcal·mol^{-1} lost in the triple mutants over the sum of the single mutant improvements.

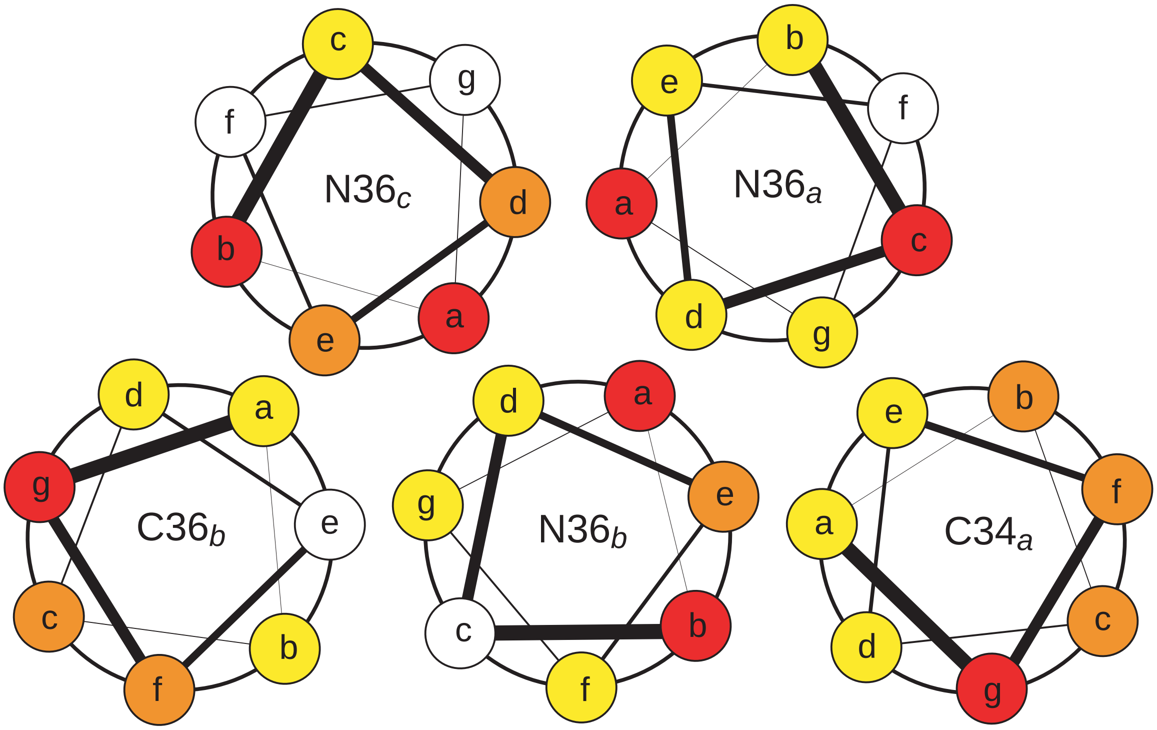


Figure 5-5: **Electrostatic contributions of 5-Helix side chains to 5-Helix stability.** The positions of the helical wheel figure are colored according to the mutation energy of side chains at that position. 🟥 indicates a residue with a mutation energy of over 3.0 kcal·mol^{-1} in magnitude, 🟧 indicates a residue with a mutation energy of at least 2.0 kcal·mol^{-1} in magnitude, and 🟨 indicates a residue with a mutation energy of 1.0 kcal·mol^{-1} or above in magnitude. A helical position is colored if *any* residue at that position has a substantial contribution.

5.3.5 Electrostatic contributions to 5-Helix stability

Mutations which enhance the binding affinity of a static structure, but which substantially destabilize that structure, will not lead to higher affinity. Rather, the resulting deformation penalty that must be paid to form the structure required for binding may lead to significantly decreased affinity. To estimate the destabilizing effects of the mutations suggested here, the electrostatic contribution of all the side chains on 5-Helix to the stability of the five-helical bundle structure was computed, using the isolated side chains in solution as a model of the unfolded state. Mapping the results onto the structure (see Figure 5-5) reveals, as expected, a much different distribution

Sequence	$\Delta G_{\mathrm{es}}^{\mathrm{stab.}}$	$\Delta G_{\mathrm{N36_a9}}^{\mathrm{mut.}}$	$\Delta G_{\mathrm{N36_a16}}^{\mathrm{mut.}}$	$\Delta G_{\mathrm{N36_c10}}^{\mathrm{mut.}}$
NEN (WT)	+147.5	−0.7	+1.3	+0.9
KQL	+148.5	+0.4	+0.2	+0.0
HQL	+150.1	+2.9	+0.2	+0.0

Table 5-5: **Electrostatic contributions to stability of 5-Helix mutants.** The relative electrostatic contributions to stability of the key mutant residues (in kcal·mol^{-1}) are displayed in the context of the wild-type structure as well as in that of the two triple mutant structures. The total electrostatic contribution of 5-Helix side chains to the stability is shown, as is the contribution of each mutated residue relative to a hydrophobic isostere.

than was seen for contributions to binding. Many of the most significant contributors are located in the core of the trimeric coiled coil, with several additional large contributions from acidic residues on the C34 helices. Few significant contributions are seen along the binding interface for the additional C34 helix, where the contributions to binding were localized.

In addition to the analysis of the wild-type structure, the contribution to stability of all side chains in the two triple mutant structures was also determined. The overall electrostatic contribution to stability for the two mutants is slightly more unfavorable than wild type, with the KQL mutant computed to have a 1.0 kcal·mol^{-1} more unfavorable contribution, and the HQL mutant computed to be more unfavorable by 2.6 kcal·mol^{-1}. The greatest contribution to this increased destabilization is from the N36$_a$ position 9. In the wild-type structure, the asparagine contributes favorably by 0.7 kcal·mol^{-1} relative to hydrophobic isostere, whereas both a lysine or a histidine contribute unfavorably, lysine only by a little (0.4 kcal·mol^{-1}) and histidine by more (2.9 kcal·mol^{-1}), at least in the charged state. On the other hand, the mutants at both other positions, N36$_a$ 16 and N36$_c$ 10, contribute more favorably to stability than do the wild-type residues. The wild-type glutamate at N36$_a$ 16 contributes unfavorably by 1.3 kcal·mol^{-1} relative to a hydrophobic group, and replacement by a glutamine reduces this unfavorable contribution to 0.2 kcal·mol^{-1}. An asparagine at position $N36_c$ 10, as is found in the wild-type sequence, contributes unfavorably by 0.9 kcal·mol^{-1}, with the Leu replacement improving this by making no electrostatic contributions to stability.

5.4 Discussion

The electrostatic contribution to the free energy of binding of 5-Helix to a C34 helix is significantly unfavorable. The direct interactions across the binding interface contribute only -10.5 kcal·mol^{-1}, with the -12.2 kcal·mol^{-1} of interactions between the C34 ligand and the groups on the inner coiled coil of 5-Helix partially offset by $+0.9$ kcal·mol^{-1} of unfavorable interactions between groups on the outer helices of 5-Helix and the bound C34. This is barely more than half of the desolvation penalty of either 5-Helix or the C34 helix, and thus the overall electrostatic contribution to binding is unfavorable by 26.6 kcal·mol^{-1}. Thus, it is clear that significant improvements in binding free energy are possible if the electrostatic interactions in this system are optimized.

Breaking down the binding free energy into the contributions made by the side chain, the backbone carbonyl, and the backbone amino group of each residue allows hot spots of electrostatic contributions to binding to be pinpointed. All the most significantly contributing side chains are located along the binding interface, with several amide and acid groups directly at the binding interface contributing unfavorably, and arginines located on either side of the interface making favorable contributions. Relative to hydrophobic isosteres, only six groups on 5-Helix contribute over 1.0 kcal·mol^{-1}, and two of these are backbone carbonyls. Only one of the four most significant side chains contributes favorably to the binding free energy, making a favorable direct interaction across the interface, and nearly fully compensating its desolvation penalty with indirect interactions. The three unfavorably contributing side chains all do so for different energetic reasons. Asn10 makes a significant unfavorable interaction with Glu22 on the bound C34, an unfavorable contribution augmented by the desolvation penalty. Glu16 makes a significant favorable interaction with Gln24, but smaller unfavorable direct interactions almost completely negate this effect, leaving the residue paying a desolvation penalty but gaining nothing in return. Gln19 does not make any significant interactions, favorable or unfavorable, but still pays a desolvation penalty. A mutation to any of these residues to their hydrophobic isostere would improve the binding affinity, but the differences in why the natural residues are unfavorable suggests that the best way to improve binding may not be the same in each case.

By considering each residue's effect on binding relative to the electrostatic optimum, rather than relative to a hydrophobic reference state, a greater amount of information useful for design can be obtained. This can clearly be seen by considering the energetics of all side chains for which optimization of the electrostatic contribution to the binding free energy results in an improvement of greater than 1.0 kcal·mol^{-1}. While the three residues identified as particularly unfavorable in the component analysis are of course included in this list, so is Arg13, which contributes favorably by 1.7 kcal·mol^{-1} relative to a hydrophobic isostere. This arginine may make a favorable contribution, but an even more favorable interaction is possible. In addition, several residues whose contributions relative to a hydrophobic isostere are both favorable and unfavorable by less than 1.0 kcal·mol^{-1} can make more significant gains upon optimization. Finally a leucine is identified as making significant improvements on optimization. As leucine is a completely hydrophobic residue, analyzing the wild-type system will never give information about the modification of this position. This demonstrates one of the key benefits of the optimization procedure; rather than identifying unfavorable interactions in the natural system and attempting to improve these, the optimization methodology allows the design targets to be chosen based on their absolute possibility of improvement, regardless of the sign or magnitude of the wild-type contribution.

While the optimization procedure identifies the sites most susceptible to improvement in electrostatic interactions, the optimal charge distributions do not correspond to precise chemical modifications. For a design procedure to be truly effective, applicable modifications must be able to be proposed, which in the case of 5-Helix requires the design of natural amino acid substitutions which improve the binding affinity. Considering the optimal charge distributions at the top four sites based on the optimal improvement in binding free energy, potential modifications were able to be suggested in three cases. While only in the case of Asn10, whose optimal charges clearly indicate a preference for a hydrophobic residue, did the optimal charges compare very closely to those of an amino acid side chain, substitutions were easily suggested for two other positions. The significant preference for a neutral or positive charge over a negative charge at position 16 on the $N36_a$ helix suggests a Glu→Gln modification which is supported by the structure — Glu16 makes a hydrogen bond with one carboxylate oxygen, but the other makes no direct interactions, thus making the glutamine NH_2 group easily accommodated. The preference for a positive charge at position 9 on the $N36_a$ helix limits the proposed modifications to histidine, lysine or arginine, with the bulky head group of arginine making this seem the least plausible substitution.

The case of Gln7 on helix $N35_c$, as well as that of Arg13 on both the $N36_a$ and $N36_c$ helices, highlights an important point regarding the optimization procedure. A significant improvement may be seen as possible in the optimization while no chemical modification matches the charge distribution required for the improvement. These three residues all clearly favor the overall charge of the wild-type residue, unlike the case for two of the positions discussed earlier. For the two arginines, the only other residues which could possibly be substituted are histidine, whose much smaller size would eliminate the ability to interact across the interface, and lysine, whose charge distribution does not seem to be a better match to the optima than does that of arginine. These arginines do contribute favorably, they just do not do so optimally, but given the limited scope of positively charged amino acids, arginines at these positions are likely better than any other choice. Similarly, the optimal charges on Gln 7 do not suggest any possible modifications. The natural and optimal charges on the NH_2 group are very similar, while the optimal charges on the carbonyl are near zero, and the aliphatic portion of the side chain is polar in the optimum. No amino acid has polarity in the aliphatic portion of the side chain, and eliminating the carbonyl would leave a amine which would clearly protonate and take on a positive charge. Thus, for these positions, although it may be possible to generate improvements with a greater scope of chemical functionality, the wild-type residues are near optimal in the limited charge space of the twenty standard amino acids.

The optimization procedure, as well as the component analysis, considers only electrostatic interactions, and strictly applies only to variations in charge within the context of the same shape and atom locations. To more accurately evaluate the effect of the suggested mutations on the binding affinity, an energetic analysis of a model of each mutant structure strengthens the predicted effects of the mutants. The roughly additive effects of the mutations suggests a lack of major interactions between any of the mutated residues. Since none of the residues are directly contacting one another, a lack of steric interactions is not surprising, but the longer range of electrostatic interactions could lead to (anti-)cooperativity. In particular, since the mutations at two positions resulted in an increase in net charge (Asn→His/Lys and Glu→Gln), it would not be surprising to see these mutations become somewhat less effective when both are simultaneously made. The combined effect of these mutations is indeed about 0.4 $\text{kcal}\cdot\text{mol}^{-1}$ less than the sum of the individual mutations, slightly greater than the up to 0.2 $\text{kcal}\cdot\text{mol}^{-1}$ difference seen for the pairs of mutations including the charge conserving Asn→Leu mutation. However, while there is a slight reduction in the efficacy of these mutations when made in combination, the effects are not large, even with mutations which both increase the charge on the inhibitor.

Mutations that improve binding affinity are meaningless if they also destabilize the folded state by a large amount. However, of the three positions mutated, only one residue contributes favorably to stability, as compared to a hydrophobic isostere, and this only by 0.7 kcal·mol^{-1} — both other positions contributed unfavorably, by 1.2 and 0.9 kcal·mol^{-1}. This suggests that these residues are not important, at least electrostatically, as stabilizing structural elements, as may be expected for residues located on the surface of the protein in the unbound state. However, analysis of the wild-type structure does not directly give information about the mutants. Repeating the analysis on the two triple mutant structures shows that the two unfavorably contributing positions become more stabilizing upon mutation. In both cases the mutation is to a less polar residue (Glu→Gln and Asn→Leu), with the more polar wild-type residue paying a larger desolvation penalty than it regains in interactions. Mutation to a less polar group reduces the desolvation penalty, with a corresponding increase in the stabilizing effect. In the case of the asparagine to leucine mutation, the favorable interactions are also eliminated, although the net change in contribution is still favorable. On the other hand, in the case of the glutamate to glutamine mutation, the favorable interactions are dominated by the interactions of a single carboxyl oxygen, which is maintained in the mutant, resulting in a slightly *more* favorable interaction in the mutant. At the third position, in which the wild-type residue contributed favorably to stability, both possible mutants contribute unfavorably. In the case of the asparagine to lysine mutation, the effect is quite small, but a much greater effect is seen for the asparagine to histidine mutant, with the histidine contributing unfavorably by 2.9 kcal·mol^{-1}. Overall, both triple mutants are computed to be slightly less stable, electrostatically, but only by a maximum of 2.6 kcal·mol^{-1}, with the majority of the larger number resulting from the Asn→His mutation. This value only includes electrostatic effects, and no large sources of strain were seen in the mutant structures. Thus these differences in stability could easily be shifted slightly in either direction by the inclusion of additional energy terms. The most important result, though, is that none of the mutations are computed to severely destabilize the protein, and only one mutation is computed to destabilize the protein by any significant amount. As 5-Helix is a highly stable protein, remaining helical to 100° C in the absence of denaturant, and not unfolding until nearly 90° C in 3.7 M GuHCl [127], even mutants with moderate reductions in stability should remain stably folded, and thus active in binding to the C-terminal peptide.

5.5 Conclusions

Continuum electrostatics provides a useful tool for the dissection of the energetics of binding of biologically important systems. Two methodologies based on continuum electrostatics, component analysis and electrostatic affinity optimization, were applied to the system of 5-Helix, a protein construct which inhibits HIV-1 viral–cell membrane fusion by binding to a peptide from the C-terminal region of HIV-1 gp41.

Component analysis revealed several residues located along the binding interface whose electrostatic interactions were unfavorable; replacement of these residues with hydrophobic isosteres was computed to stabilize the complex. In addition several residues making significantly favorable electrostatic interactions were identified.

Electrostatic affinity optimization provides a means to investigate further the locations and types of mutations most likely to improve binding. This procedure involves varying the charge distribution so as to maximize the favorable interactions in the bound state relative to the unfavorable desolvation penalty. Applying constraints on the total charge of a side chain during the optimization provides a means to quickly determine, in general, the feasibility of a mutation of each residue to a negative, neutral, or positive replacement. In addition, the optimal binding free energy, corresponding to the *best possible* electrostatic binding free energy given that geometry, provides a quantitative measure of the degree to which a mutation will be able to improve binding. Finally, by analyzing the optimal charge distributions of individual side chains, further insight into how binding may be improved is possible.

Using this technique, three residues on 5-Helix were identified as the best candidates for mutation, and four changes to naturally occurring amino acids were suggested (two mutations seemed equally promising at one site). Modeling of the mutant structures and evaluation of their relative binding free energies show calculated improvements in binding for each single mutant of five-fold to seventeen-fold. The improvement gained by each mutant was roughly additive when multiple mutants were considered, and an improvement of over 500-fold is calculated for one of the triple mutants.

While the calculations presented here are based on 5-Helix, several other constructs based around the trimeric coiled coil from gp41 have been made and are active inhibitors of HIV-1 viral–cell membrane fusion [41, 97]. Although the details of the energetics are likely to vary somewhat with the specific design of the construct, it is probable that the mutations suggested here for 5-Helix would have similar effects in other systems of related structure.

Chapter 6

"Action-at-a-Distance" Interactions: Enhancement of Binding Affinity Through Through Long-Range Electrostatic Interactions

Abstract

The electrostatic contributions to the free energy of binding of β-lactamase inhibitor protein (BLIP) to TEM1 β-lactamase were considered in detail using a continuum solvation model. In addition to several interfacial residues identified as playing an important role in stabilizing the complex, a number of charged residues somewhat removed from the interface were also found to contribute significantly to the binding free energy, with both favorable and unfavorable interactions observed as far as 10 Å away from the interface. Optimization of the side-chain partial atomic charges on BLIP gave similar results. While interfacial residues can make large contributions to the binding free energy, the wild-type residues are near optimal; the greatest opportunities for improving the binding affinity relative to wild type are located somewhat more removed from the interface. The results of the energetic analysis identified ten residues, all exposed on the surface in both the bound and unbound states, whose

mutation to a positively charged residue was computed to improve the binding affinity. The energetic effects can be quite significant, with the optimal charges on all ten residues computed to yield an improvement in binding free energy greater than 15.0 kcal·mol^{-1} over wild type, and individual side chains yielding optimal improvements as high as 7.7 kcal·mol^{-1}. The results are a promising indication of a novel avenue for the design of tight binding protein–protein complexes, namely, the improvement of complementary electrostatic interactions at surface patches outside of the binding interface, where packing restrictions might be small.

6.1 Introduction

Over the last twenty years, many advances have been made in the field of protein design, largely as a result of phrasing the appropriate inverse problem and developing methods capable of addressing inverse design [40, 114]. Many of the current protocols for protein design involve the construction of stabilizing protein side-chain arrangements by methods including dead-end elimination [33, 37, 58, 88, 90, 96], self-consistent mean-field theory [84–86], simulated annealing [61, 91], and genetic algorithms [35, 72]. In all these approaches, successful design is achieved by the consideration of detailed atomic interactions and their effects on the geometry and energetics of packing.

While the bulk of the work to date has focused on the design of protein cores, the design of protein binding interfaces can, in principle, be addressed by a similar overall approach. However, while the hydrophobic cores of proteins can reasonably be treated by methods which greatly simplify, or even completely neglect, the effects of electrostatic interactions, both between protein groups and with solvent, protein interfaces generally contain polar and charged residues [27], and thus an appropriate treatment of electrostatics is necessary. While the additional requirement to treat solvation and electrostatic interactions adds a further layer of complexity to an already difficult problem in these cases, recent work has begun to address some of these issues [17, 93].

In many cases, an alternative strategy, and one that does not demand the same detailed packing together of side chains into an exquisite three-dimensional jigsaw puzzle, may be desirable. One potential method of this type involves the enhancement of affinity through the creation of favorable, relatively long-range electrostatic interactions by the mutation of surface residues located somewhat outside of the protein–protein binding interface. When the residues being considered are not located directly at the binding interface, but rather remain on the protein surface even in the bound state, a detailed consideration of the packing of residues may be unnecessary. Futhermore, as a result of the relatively long range over which such mutations project their electrostatic effects, such a design strategy should be more tolerant of local imperfections in structural models.

While it seems that these "action-at-a-distance" electrostatic interactions may be a useful tool in the design of high-affinity protein–protein complexes, it is less apparent how effective these types of mutations can be. Since these interactions may be highly screened by solvent, the energetic contributions could be too small to be of any relevance in design. Another important question is whether the sites where such mutations will be most effective are localized on the structure, and, if so, how these locations might be determined.

An important consideration in any type of design involving electrostatic interactions is the counterplay of favorable direct electrostatic interactions and unfavorable desolvation effects, which has been shown to be incredibly important in understanding the energetics of electrostatic interactions in biological systems. Buried salt-bridges in proteins have been found in general to contribute relatively little to the stability of proteins, and in many cases contribute unfavorably, due to the large desolvation penalty outweighing the favorable interactions made in the folded state [68]. Similar results have been seen in both protein–protein and protein–DNA complexes, with unfavorable desolvation effects being greater than the favorable interactions made on complex formation, and thus leading to an unfavorable electrostatic contribution to binding [17, 64]. However, much of this work has focused on individual, short-range electrostatic interactions such as salt-bridges and hydrogen bond networks, and the lessons learned from detailed analyses of these systems may or may not prove to be extendable in a straightforward manner to longer-range electrostatic interactions.

We have begun to address these issues by analyzing the electrostatic contributions to binding in the complex of the β-lactamase inhibitor protein (BLIP) with TEM1 β-lactamase (TEM1) [132]. Using methods based on continuum electrostatics, we were able to consider in detail the electrostatic contributions to the energetics of binding for all residues in the complex, with a particular focus on those residues situated at the periphery of the binding interface. In addition, an electrostatic optimization procedure was applied to all residues on BLIP. This procedure explicitly identifies molecular fragments whose electrostatics are undercompensated and might be improved through the design of mutations. In a number of instances, mutations to take advantage of these types of peripheral interactions were identified.

6.2 Methods

Preparation of structures. All calculations were done using the X-ray crystal structure the BLIP–TEM1 complex (Protein Data Bank [125] ID 1jtg) [144]. Hydrogen atoms were added using the HBUILD facility [14] within the CHARMM computer program [11] using the PARAM22 all-atom parameter set [100]. An analysis of hydrogen-bonding patterns suggested no reason for the ionizable residues to be in their non-standard protonation states, and thus all histidines were left in their neutral state, and all acidic residues were left charged.

Continuum electrostatic calculations. All continuum electrostatic calculations were performed using a locally modified version of the DELPHI computer program [55, 57, 134, 136] to solve the linearized Poisson–Boltzmann equation. An internal dielectric constant of 4 and an external dielectric constant of 80 were used unless otherwise specified, and the ionic strength was set to 0.145 M, with a 2.0 Å ion exclusion (Stern) layer [9]. The dielectric boundary was specified by the molecular surface generated with a 1.4 Å radius probe. Protein partial atomic charges and radii were taken from the PARSE parameter set [140] with a few minor changes. Charges on the bridging ring carbons of tryptophan were assigned to $0e$, charges for proline and for disulfide bridged cysteine residues were taken from the PARAM19 parameter set [11], and the charges from glutamate and lysine side chains were used for charged C and N termini respectively. Binding was considered in the rigid-body docking approximation.

Calculations for the component analysis were done using a three-step focusing procedure on a 161×161×161 unit cubic grid, with the longest dimension of the molecule occupying first 23%, then 92%, and finally 184% of one edge of the grid, resulting in a final grid spacing of 0.22 Å. Boundary potentials for the more highly focused calculations were obtained from the previous calculation, and Debye–Hückel potentials were used at the boundary of the lowest resolution (23%) calculation. For the highest resolution calculations, the grid was centered on the region of interest, and interactions involving groups falling outside of this grid were computed from the 92% fill calculation. All calculations were averaged over ten translations of the structure on the grid in order to minimize artifacts from the the placement of the point charges and molecular boundaries onto the finite difference grid. Calculations to determine matrix elements for the electrostatic affinity optimization were done using the same procedure, but with a 129×129×129 grid (final grid spacing of 0.28 Å). All other calculations were done using a two-stage focusing procedure (the molecule occupying first 23% then 92% of the grid) on a 257×257×257 grid (final grid spacing of 0.28 Å). While the component analyses were done with finer grid spacing, the potentials are converged with respect to the spacing of the grid at all values used. The three-step focusing methods were used to reduce the computational cost of the calculations in cases where the potential produced by only a small subset of charges is of interest. In this case, the results of the three-step focusing on a smaller grid give equivalent results as a two-step procedure using a grid of twice the size.

Electrostatic affinity optimization. The electrostatic affinity optimizations, in which the ligand charge distribution is allowed to vary so as to produce the most favorable electrostatic contribution to the binding free energy, were performed as previously described [23, 77–80, 92–94] using locally written software. Singular value decomposition [119, 143] was used to remove all basis vectors with singular values of less than 10^{-5} of the largest singular value or for which the error over ten translations was greater than 25% of the value. Typically this involved the removal of 1131 out of 1436 basis vectors; the majority of residues significantly removed from the interface pay almost no desolvation, leading to a large number of very small eigenvalues in the desolvation matrix. Basis vectors in the null space were allowed to be populated only when required to satisfy imposed constraints, and were penalized by a harmonic penalty with a coefficient of 10.0 $\text{kcal}\cdot\text{mol}^{-1}\cdot e^{-2}$ in the optimization, but not in the final energy evaluation. Constrained optimizations were performed using the computer program LOQO [133, 154, 155]. Typical constraints used in all optimizations were that all residues must have an integral net charge, that no residue may have a net charge of greater that $1.0e$ in magnitude, and that no individual partial atomic charge may exceed $0.85e$ in magnitude. These constraints were chosen to limit the optimization to regions of charge space reasonably attainable in the context of amino-acid chemistry.

6.3 Results

6.3.1 Electrostatic contributions to BLIP–TEM1 binding

Previous work has described a methodology by which the contribution of various groups in a protein to the electrostatic binding free energy may be calculated [69] (see Section 2.4.1). For the purpose of this work, each residue was considered as the union of three chemical groups: backbone carbonyl, backbone amino and side chain. For each group all of its energetic contributions to the binding free energy were calculated. These are: (i) the desolvation penalty, which is the energetic cost of moving the group from the region of low dielectric in the unbound state to the (larger) region of low dielectric in the bound state; (ii) the indirect interactions, which are the energetic change in interactions between different groups in the same molecule when the dielectric boundary is changed from that of the unbound state to that of the bound state (intramolecular interactions); (iii) the direct interactions, which are the interactions between a group on one molecule and groups on the other molecule in the bound state (intermolecular interactions).

A variety of analyses were carried out to understand the balance of electrostatics involved in binding of the BLIP–TEM1 complex. The change in electrostatic binding free energy due to turning on the partial atomic charges in a chemical group in the context of all other partial atomic charges (called the "mutational free energy") is a calculation similar in spirit to a set of alanine scanning experiments. Rather than measuring the effect of each side chain relative to alanine in the context of all others, this procedure calculates the electrostatic effect of each set of charges (backbone groups and side-chain groups) in the context of all others. Because the calculation only varies the partial atomic charges but not the shape of the group, it corresponds to a comparison of the effect of the actual group to the effect of its hydrophobic isostere. It has been pointed out that mutational free energies, whether from computations of this sort or scanning experiments, do not correspond even approximately to additive free energy contributions, since their addition double counts pairwise interactions [69].

The mutational free energy was computed for all chemical groups in the BLIP–TEM1 complex. For TEM1 they spanned a range from -7.7 kcal·mol^{-1} (a favorable effect on binding) for the Lys208 side chain to $+2.7$ kcal·mol^{-1} (an unfavorable effect) for the Glu213 side chain; for BLIP they spanned -14.3 kcal·mol^{-1} (Lys74 side chain) to $+4.6$ kcal·mol^{-1} (Asp163 side chain). All groups with mutational free energy greater in magnitude than 0.5 kcal·mol^{-1} are displayed in Table 6-1 for TEM1 and in Table 6-2 for BLIP. The largest mutational components on both binding partners correspond to charged side chains. Backbone groups have mutational free energies of

		$\Delta\Delta G^{\text{des.}}$	$\Delta\Delta G^{\text{dir.}}$	$\Delta\Delta G^{\text{ind.}}$	$\Delta\Delta G^{\text{con.}}$	$\Delta\Delta G^{\text{mut.}}$	$d_{\text{min.}}$
Lys	208	1.91	−11.10	1.46	−2.91	−7.73	2.76
Lys	48	2.19	−4.26	−3.45	−1.67	−5.52	4.78
Arg	217	1.99	−8.68	1.24	−1.73	−5.45	1.80
Lys	86	1.96	−2.88	−2.19	−0.58	−3.11	2.19
Glu	213	3.19	−0.43	−0.01	2.97	2.75	2.55
Asp	106	1.36	3.34	−2.37	1.84	2.33	6.60
*Glu	146	0.17	1.12	0.54	1.00	1.83	5.64
Glu	143	1.78	−0.46	0.26	1.68	1.58	2.26
$^{\circ}$Asp	207	0.09	2.45	−1.24	0.70	1.30	9.06
Glu	141	1.96	2.37	−3.09	1.60	1.24	4.40
Asp	90	0.10	0.95	0.09	0.62	1.14	7.67
*Arg	139	0.05	−0.56	−0.55	−0.51	−1.06	6.94
$^{\circ}$Arg	196	0.04	−1.43	0.49	−0.43	−0.90	8.70
Glu	85	5.76	−4.03	−0.84	3.33	0.89	1.71
Arg	214	0.19	−0.31	−0.70	−0.32	−0.82	3.55
*Asp	154	0.04	0.58	0.15	0.41	0.78	10.77
Glu	79	10.40	−10.16	−0.84	4.90	−0.60	1.66
Glu	96	0.08	0.50	−0.04	0.32	0.55	5.60
Arg	153	0.01	−0.27	−0.25	−0.25	−0.51	8.52
Ser	209	0.96	−5.30	1.47	−0.97	−2.88	1.86
Ser	105	1.06	−4.77	1.47	−0.59	−2.24	1.73
Hsd	87	0.64	1.13	−0.17	1.12	1.60	2.82
Asn	107	0.46	−0.16	0.57	0.67	0.87	3.46
NH	80	0.89	−3.06	0.43	−0.42	−1.73	1.86
CO	75	0.84	−2.29	0.09	−0.26	−1.37	1.84
NH	81	0.88	−2.77	0.77	−0.12	−1.13	1.90
CO	104	0.56	0.08	0.10	0.65	0.74	3.43
CO	141	0.23	−0.49	−0.38	−0.21	−0.64	4.22
CO	79	0.22	−1.10	0.28	−0.20	−0.61	3.70
CO	85	0.08	−0.13	−0.55	−0.26	−0.60	4.91
CO	211	0.54	0.06	−0.01	0.56	0.59	2.86
CO	80	0.23	−1.45	0.64	−0.18	−0.58	3.61
CO	212	0.26	−0.32	−0.51	−0.16	−0.57	3.54

Table 6-1: **Greatest TEM1 contributions to BLIP–TEM1 complex formation.** All components on TEM1 (in kcal·mol^{-1}) with a mutational energy of greater than 0.5 kcal·mol^{-1} in magnitude are listed, grouped into charged side chains, polar side chains, and backbone groups. Highlighted in yellow are those components identified as acting through an "action-at-a-distance" mechanism. * and $^{\circ}$ mark groups of charged residues interacting with each other through intramolecular salt bridges.

		$\Delta\Delta G^{\text{des.}}$	$\Delta\Delta G^{\text{dir.}}$	$\Delta\Delta G^{\text{ind.}}$	$\Delta\Delta G^{\text{con.}}$	$\Delta\Delta G^{\text{mut.}}$	$d_{\text{min.}}$
Lys	74	6.81	−13.22	−7.89	−3.74	−14.30	1.66
Asp	49	12.81	−25.36	1.43	0.85	−11.11	1.73
Asp	163	1.74	3.18	−0.35	3.15	4.57	3.61
Arg	160	2.41	−5.90	0.36	−0.36	−3.14	1.84
Glu	73	10.98	−6.79	−5.98	4.60	−1.79	1.87
Arg	144	0.45	−1.58	−0.61	−0.64	−1.74	2.84
Asp	133	0.04	0.92	0.09	0.54	1.05	10.57
Asp	68	0.15	0.58	0.08	0.48	0.80	7.41
Arg	43	0.26	0.48	−1.45	−0.23	−0.72	4.47
Asp	135	0.02	0.56	0.05	0.32	0.62	8.83
Ser	71	0.94	−3.68	0.31	−0.74	−2.42	1.71
Ser	113	0.61	−2.11	0.32	−0.28	−1.18	1.95
Thr	55	1.12	−0.70	0.24	0.89	0.66	2.69
Phe	142	0.76	−0.25	0.03	0.65	0.54	2.55
CO	35	0.96	1.10	0.20	1.60	2.25	2.82
NH	143	0.70	−3.38	0.71	−0.64	−1.97	2.08
CO	142	0.32	−1.89	0.37	−0.44	−1.20	3.54
CO	49	0.98	−0.86	0.99	1.04	1.11	2.79
CO	36	0.76	−0.78	−0.90	−0.08	−0.92	2.19
NH	142	0.43	0.50	−0.02	0.66	0.90	2.24
CO	141	0.74	−0.20	−1.36	−0.04	−0.82	1.85
NH	48	0.38	0.28	−0.04	0.50	0.62	2.88
CO	138	0.03	0.88	−0.30	0.32	0.61	5.26
CO	76	0.02	0.38	0.13	0.28	0.53	10.32
CO	144	0.01	−0.43	−0.10	−0.26	−0.53	7.05
NH	145	0.01	−0.38	−0.15	−0.25	−0.52	6.64
CO	71	0.02	−0.42	−0.11	−0.24	−0.51	5.41

Table 6-2: **Greatest BLIP contributions to BLIP–TEM1 complex formation.** All components on BLIP with a mutational energy of greater than 0.5 kcal·mol^{-1} in magnitude are listed, grouped into charged side chains, polar uncharged side chains, and backbone groups. Highlighted in yellow are those components identified as acting through an "action-at-a-distance" mechanism. All energies are in kcal·mol^{-1}.

as high as 2.3 kcal·mol^{-1} in magnitude, and a handful of neutral-polar side chains have mutation energies of up to 2.9 kcal·mol^{-1}. However, four charged side chains on both TEM1 and BLIP have mutation energies of more than 3.0 kcal·mol^{-1} in magnitude, with two groups on BLIP having the charged state favored by over 10.0 kcal·mol^{-1} over the corresponding hydrophobe. All of these most favorably contributing residues are located near the center of the binding interface.

Three residues on TEM1 and six residues on BLIP are particularly interesting due to their location on the periphery of the binding interface (see Figure 6-1). On TEM1, Asp90, Glu96, and Glu146 have unfavorable mutational free energies (of up to 1.8 kcal·mol^{-1}), and all are at least 5.5 Å from the nearest atom on BLIP. On BLIP, Asp68, Asp133, and Asp135 have unfavorable mutational free energies (by up to 1.0 kcal·mol^{-1}) despite none being closer than 7.4 Å to any TEM1 atom. All these residues are located on the periphery of the binding interface, too far removed from TEM1 to make any direct contact across the interface. In comparison, polar residues making direct interactions at the binding interface typically make contacts with atom-to-atom distances (including hydrogen atoms) of less than 3.0 Å, with distances below 2.0 Å not uncommon (see Tables 6-1 and 6-2). The BLIP residues Arg144, Arg160 and Asp163 occupy a slightly different location somewhat closer to the binding interface. While still distinctly peripheral, these residues make contacts with atoms on TEM1 at a distance of between 1.8 and 3.6 Å. The contributions of these residues relative to a hydrophobic replacement are more significant, due to their closer interactions; the two arginines have mutational free energies of -1.7 and -3.1 kcal·mol^{-1}, while that of Asp163 is $+4.6$ kcal·mol^{-1}. Thus, all the positively charged residues in this set contribute favorably, while all the negatively charged residues contribute unfavorably. Looking in more detail at the component contributions for this set of side chains reveals a somewhat unexpected pattern. In all cases the group desolvation energy is quite low — between 0.5 and 2.4 kcal·mol^{-1} for the closer contacting set and below 0.2 kcal·mol^{-1} for those residues further removed; in addition, the total of

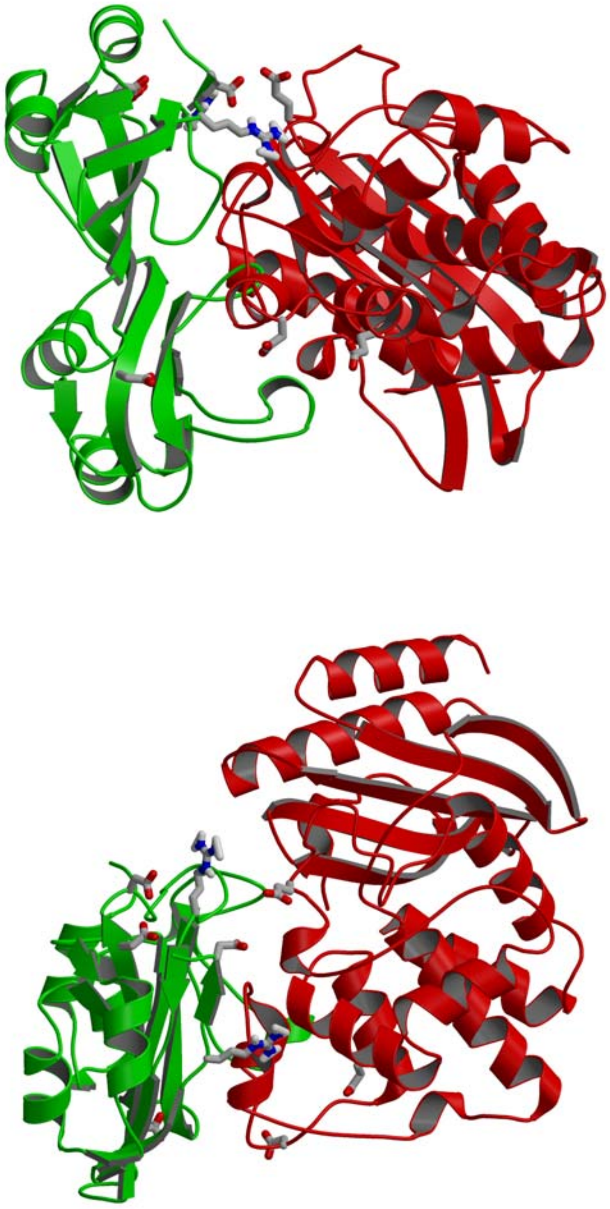


Figure 6-1: **"Action-at-a-distance" components in the BLIP–TEM1 complex.** The residues involved in "action-at-a-distance" interactions are displayed on the structure of the BLIP–TEM1 complex. TEM1 is displayed in **red** and BLIP in **green** The bottom view is rotated 90° out of the page relative to the top view. These figures was prepared with MOLSCRIPT [87] and RASTER3D [105].

the indirect interactions is at most 0.6 kcal·mol^{-1} in magnitude in all cases. The direct interactions, however, are more significant and dominate; the direct interactions accounted for up to 1.1 kcal·mol^{-1} in magnitude for the more peripheral set and up to 5.9 kcal·mol^{-1} in magnitude for the closer contacting group of residues. These data suggest that reasonably strong effects on binding occur through interactions involving residues near the interface that are not buried yet participate in strong intermolecular effects. It is interesting that the calculated energetics show very low desolvation penalties (because the side chains remain solvent exposed in the bound state) yet intermolecular interactions that are relatively strong despite solvent screening. If this is indeed the case, as a class such interactions may provide a convenient and attractive mode for altering molecular binding affinity.

6.3.2 Variation of results with internal dielectric

Since a dielectric constant of 4.0 may not be a good model for the surface of a protein, where the increased motion of protein atoms may lead to a higher effective dielectric constant [1, 2, 129], the effect of the value of the internal dielectric constant on the component energies was evaluated. The component analysis was repeated using an internal dielectric constant of 20.0 as well as using a uniform dielectric of 80.0, with an ionic strength of 0.0 M in the latter case (see Table 6-3). While for most groups, the mutational binding free energy was strongly reduced in magnitude even in changing the internal dielectric constant from 4.0 to 20.0 (for BLIP Lys74, $\Delta\Delta G^{\mathrm{mut.}}$ is −14.3 kcal·mol^{-1} with $\epsilon_{int} = 4.0$, but is only −3.7 kcal·mol^{-1} with $\epsilon_{int} = 20$), in the case of the six peripherally acting residues on BLIP the effect of the internal dielectric constant is much less. The greatest effect is seen for the relatively closely contacting Asp163, whose mutational energy changes from 4.6 to 2.2 kcal·mol^{-1} (barely a two-fold reduction) as ϵ_{int} changes from 4.0 to 20.0. This small variation with internal dielectric extends to the calculation in a uniform dielectric constant of 80.0. In fact, due to the lack of screening by mobile ions, the computed interaction energy in

		Mutation			Interaction	
	ϵ_{int} =	4	20	$80^{\dagger\ddagger}$	20	4
Lys	74	−14.30	−3.73	−2.07	−3.35	−13.22
Asp	49	−11.11	−3.37	−0.79	−6.09	−25.36
Asp	163	4.57	2.20	1.95	1.89	3.18
Arg	160	−3.14	−2.07	−2.16	−2.55	−5.90
Glu	73	−1.79	0.16	0.90	−0.82	−6.79
Arg	144	−1.74	−1.16	−1.57	−0.98	−1.58
Asp	133	1.05	0.68	1.26	0.63	0.92
Asp	68	0.80	0.42	0.86	0.30	0.58
Arg	43	−0.72	−0.23	−0.71	0.28	0.48
Asp	135	0.62	0.51	1.22	0.46	0.56
Ser	71	−2.42	−0.77	−0.33	−0.94	−3.68
Ser	113	−1.18	−0.48	−0.23	−0.61	−2.11
Thr	55	0.66	−0.03	−0.08	−0.21	−0.70
Phe	142	0.54	0.04	−0.04	−0.07	−0.25
CO	35	2.25	0.52	0.12	0.34	1.10
NH	143	−1.97	−0.55	−0.19	−0.72	−3.38
CO	142	−1.20	−0.33	−0.07	−0.38	−1.89
CO	49	1.11	0.08	0.07	0.24	−0.86
CO	36	−0.92	−0.20	−0.06	−0.18	−0.78
NH	142	0.90	0.12	−0.03	0.04	0.50
CO	141	−0.82	−0.12	−0.04	−0.06	−0.20
NH	48	0.62	−0.10	0.02	−0.23	0.28
CO	138	0.61	0.19	0.06	0.22	0.88
CO	76	0.53	0.12	0.03	0.09	0.38
CO	144	−0.53	−0.21	−0.07	−0.19	−0.43
NH	145	−0.52	−0.10	−0.04	−0.09	−0.38
CO	71	−0.51	−0.15	−0.06	−0.14	−0.42

† [Salt] = 0 M for constant dielectric.

‡ Mutation and interaction are equal in constant ϵ.

Table 6-3: **Variation of BLIP components with internal dielectric.** The mutation and interaction energies (in kcal·mol^{-1}) on all components on BLIP identified in Table 6-2 are tabulated for various internal dielectric constants. Again, those components identified as acting through an "action-at-a-distance" mechanism are highlighted in yellow.

dielectric 80.0 is *larger* in magnitude than that computed with an internal dielectric constant of 4.0 in several cases[1]. The two largest components, both of which make direct interactions with TEM1 and are well buried in the middle of the binding interface, show a dramatic reduction in interaction with increasing internal dielectric constant, with the direct interactions of Lys74 being reduced more than six-fold, and those of Asp49 being reduced by well over twenty-fold, upon moving from the standard conditions of an internal dielectric constant of 4.0 and an ionic strength of 0.145 M to the uniform dielectric constant 80.0 with no mobile ions. Two other charged groups buried at the binding interface (Arg43 and Glu73) show a change in sign of the interaction energy in the uniform dielectric case relative to the calculations with a lower internal dielectric constant. This is due to the the change in the distance dependence of the interactions as salt is removed from the system (reducing the screening of long-range interactions) and high dielectric solvent is allowed inside the molecules (increasing the screening of short-range interactions). The energetics of the most significant non-charged groups are also all greatly reduced with increased internal dielectric constant, with all terms below 1.0 kcal·mol^{-1} in magnitude at an internal dielectric constant of 20.0, and all below 0.4 kcal·mol^{-1} in magnitude in a uniform dielectric constant of 80.0.

6.3.3 Optimization of BLIP binding

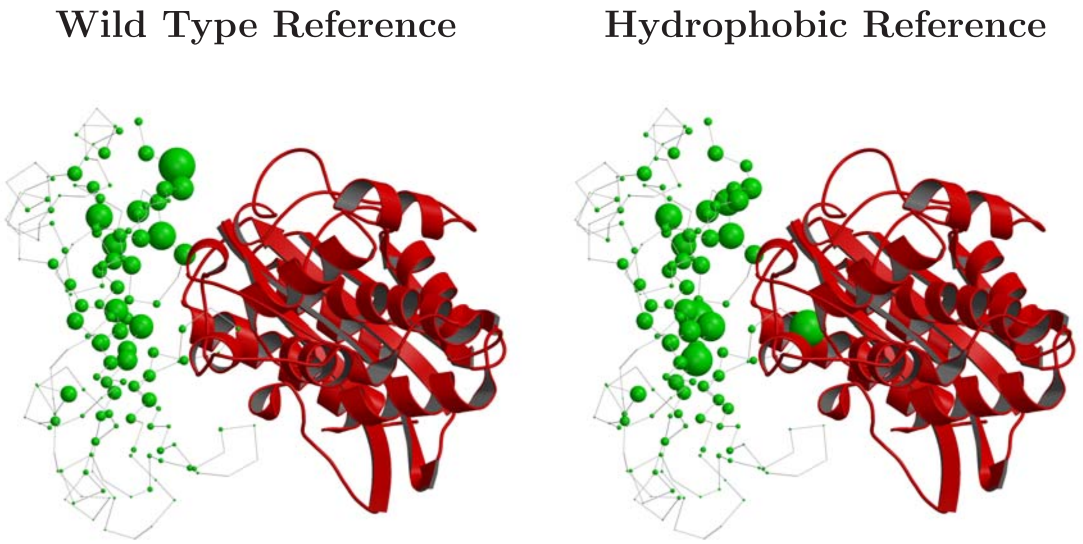
In addition to the component analysis, the partial atomic charges on the side-chain atoms of every residue on BLIP were optimized so as to yield the best possible binding free energy to TEM1. This was done for each residue in turn, with the charges of all other residues, and those of the protein backbone, fixed at their natural values. Three sets of optimizations were performed, constraining the total charge on each residue to -1, 0 and $+1e$ (see Table 6-4). Many residues show significant optimal

---

[1] In a uniform dielectric with 0.0 M ionic strength, all electrostatic energies reduce to Coulomb's Law, and thus there is no desolvation.

	Q_{res}	-1	0	$+1$
Asp	163	−3.45	−5.98	−7.70
Tyr	143	3.01	−2.48	−6.00
Phe	132	−3.31	−4.83	−5.70
Tyr	137	−3.45	−4.38	−5.16
Glu	73	−1.33	−5.09	−3.41
Trp	162	5.04	−1.46	−5.08
Trp	112	2.72	−2.20	−5.06
Ser	146	5.24	−1.16	−4.72
Phe	142	6.61	−1.98	−4.61
Ser	71	8.43	−1.16	−4.40
Gln	72	2.40	−1.36	−4.01
Ser	138	0.06	−2.21	−3.95
Hsd	148	3.79	−1.18	−3.85
Leu	75	−0.60	−2.37	−3.73
Ser	130	4.46	−1.21	−3.71
Arg	144	0.53	−1.67	−3.62
Leu	76	−1.74	−2.71	−3.59
Phe	9	−1.06	−2.38	−3.54
Gln	161	−0.72	−2.19	−3.53
Ser	113	5.16	−0.42	−3.48
Leu	164	−0.31	−1.95	−3.35
Leu	83	−0.98	−2.14	−3.20
Thr	55	3.84	−1.22	−3.14
Tyr	115	−0.24	−2.01	−3.10
Leu	129	−0.83	−2.03	−3.10
Leu	149	0.09	−1.70	−3.05
Ser	128	5.69	−0.16	−3.02
Ser	69	2.12	−1.20	−3.00

Table 6-4: **Greatest optimal improvements on BLIP side chains for binding TEM1 (relative to wild type).** All BLIP side chains whose optimal improvement in binding free energy relative to a *wild type* reference state is greater than 3.0 kcal·mol^{-1} are tabulated. Results for optimizations constrained to −1, 0, and +1e total residue charge are shown. Highlighted in yellow are those components identified as acting through an "action-at-a-distance" mechanism. All energies are in kcal·mol^{-1}.



Wild Type Reference **Hydrophobic Reference**

Figure 6-2: **Optimization of BLIP side chains for binding to TEM1.** The results of the electrostatic optimization of BLIP side chains for binding to TEM1 are shown mapped onto the structure of the complex. TEM1 is displayed in **red** and BLIP in **green**. The radius of the sphere at each C_α on BLIP is proportional to the energetic improvement of that residue on optimization. The left hand figure displays the results relative to wild type, and largest radius corresponds to a 7.7 kcal·mol^{-1} improvement. The right hand figure shows the results relative a hydrophobic reference state on the same scale as the first figure, with all improvements above 7.7 kcal·mol^{-1} given an equal radius. These figures were prepared with MOLSCRIPT [87] and RASTER3D [105].

improvements in binding free energy relative to the wild-type charge distribution, including seven residues with improvements of over 5.0 kcal·mol^{-1}. Of these, all but one have an optimal net charge of $+1e$ — the exception being Glu73, whose optimum is neutral in overall charge. The residue showing the greatest improvement is Asp163, the largest contributing component in the set of peripherally located residues. Other than the neutrally optimizing Glu73, all the remaining largest improvements are seen for a set of aromatic residues. Two of these show improvements of over 3.0 kcal·mol^{-1} regardless of the total charge of the residue, while the other three have an optimal binding free energy worse than wild-type when the residue is constrained to $-1e$ and show improvements of less than 2.5 kcal·mol^{-1} for the neutral optima. Considering how the optimal improvements in binding free energy map on to the structure of the complex (Figure 6-2), it can easily be seen that the largest improvements tend to localize to the region of BLIP around Asp163, on the edge of the binding interface. The residues located directly at the binding interface generally show small optimal improvements, while moderate improvements are seen on the layers located directly behind the first contact layer of the interface.

The above results all use a reference state of the wild-type charge distribution, while another reasonable choice of a reference state is the hydrophobic isostere, as is used in the component analysis. The results of the optimization relative to the hydrophobic reference are displayed in Table 6-5. With this choice of reference state, the two largest optimal improvements are seen for Lys74, which optimizes to the wild-type net charge of $+1e$ to give an improvement of 17.7 kcal·mol^{-1}, and Asp49, which gives an optimal improvement of 11.2 kcal·mol^{-1} at the wild-type charge of -1 e. Both these residues show improvements of over 6.0 kcal·mol^{-1} regardless of the total charge of the residue. The third largest improvement (6.8 kcal·mol^{-1}) is seen for Ser71, which also was seen to be the most significant non-charged component in the component analysis. Ser71 optimizes to a net charge of $+1e$, although an improvement of 3.6 kcal·mol^{-1} is seen for the neutral optimum. A negative charge, however, is excluded at this position, with the optimal charge distribution with a $-1e$ net charge binding 6.0 kcal·mol^{-1} worse than the hydrophobic side chain. The set of aromatic groups which showed large improvement relative to wild type also show large improvements relative to the hydrophobic residue, as does Glu73. Glu163, on the other hand, does not show nearly as large improvements relative to a hydrophobic residue as it does relative to the wild-type charge distribution. The localization of these results on the structure of the complex (Figure 6-2) shows some similarities to the mapping of the wild-type reference results, but also reveals distinct differences. While significant improvements are still seen in the peripheral region near Asp 163, the largest improvements are located directly at the binding interface, and the residues

	Q_{res}	−1	0	+1
Lys	74	−7.04	−13.09	−17.70
Asp	49	−11.15	−6.61	14.10
Ser	71	6.00	−3.58	−6.82
Glu	73	−2.38	−6.14	−4.47
Arg	144	−1.64	−3.84	−5.79
Phe	132	−3.32	−4.84	−5.72
Tyr	143	3.46	−2.02	−5.54
Tyr	137	−3.55	−4.48	−5.26
Trp	112	2.81	−2.12	−4.97
Ser	146	5.11	−1.28	−4.85
Ser	113	3.96	−1.62	−4.67
Trp	162	5.53	−0.96	−4.58
Arg	160	0.62	−2.40	−4.57
Gln	72	2.31	−1.45	−4.10
Phe	142	7.16	−1.43	−4.06
Ser	138	0.20	−2.07	−3.81
Hsd	148	3.86	−1.11	−3.78
Leu	75	−0.60	−2.36	−3.73
Leu	76	−1.74	−2.71	−3.59
Ser	130	4.63	−1.04	−3.54
Phe	9	−1.00	−2.32	−3.49
Lys	70	−1.61	−2.54	−3.43
Leu	164	−0.31	−1.95	−3.36
Gln	161	−0.49	−1.96	−3.30
Ser	128	5.46	−0.40	−3.25
Leu	83	−0.98	−2.14	−3.20
Leu	129	−0.82	−2.03	−3.10
Ser	69	2.05	−1.27	−3.07
Leu	149	0.09	−1.70	−3.05

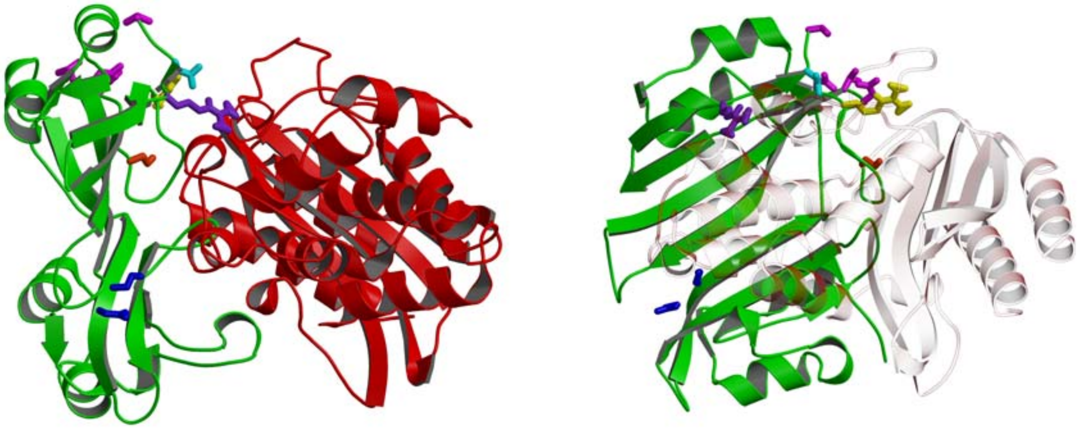
Table 6-5: **Greatest optimal improvements on BLIP side chains for binding TEM1 (relative to hydrophobic).** All BLIP side chains whose optimal improvement in binding free energy relative to a *hydrophobic* reference state is greater than 3.0 kcal·mol^{-1} are tabulated. Results for optimizations constrained to −1, 0, and +1e total residue charge are shown. Highlighted in yellow are those components identified as acting through an “action-at-a-distance” mechanism. All energies are in kcal·mol^{-1}.

			WT Reference			Hϕ Reference		
		Q_{res}	−1	0	+1	−1	0	+1
A	Asp	68	−0.62	−1.50	−2.25	0.60	−0.27	−1.03
	Ser	69	2.12	−1.20	−3.00	2.05	−1.27	−3.07
B	Asp	133	−0.19	−1.50	−2.75	1.24	−0.07	−1.32
	Leu	134	0.16	−0.31	−0.77	0.16	−0.31	−0.77
	Asp	135	−0.53	−1.38	−2.17	0.47	−0.39	−1.17
	Val	165	−0.51	−0.94	−1.69	−0.51	−0.94	−1.69
C	Ser	138	0.06	−2.21	−3.95	0.20	−2.07	−3.81
D	Arg	144	0.53	−1.67	−3.62	−1.64	−3.84	−5.79
E	Arg	160	4.21	1.20	−0.97	0.62	−2.40	−4.57
F	Asp	163	−3.45	−5.98	−7.70	1.60	−0.94	−2.66

Table 6-6: **"Action-at-a-distance" improvements upon optimization.** The results of the optimization (in kcal·mol^{-1}) for all BLIP residues identified as being capable of "action-at-a-distance" interactions are tabulated. Residues located proximally to one another in the structure at grouped together.

in the layers behind the first contact layer show larger improvements than relative to wild type.

Analysis of the structure, in combination with the optimization results, resulted in the identification of ten residues poised to make peripheral interactions with TEM1. The individual optimization of these residues yields improvements of between 0.8 and 7.8 kcal·mol^{-1} over wild type, and between 0.8 and 5.8 kcal·mol^{-1} over hydrophobic isosteres (see Table 6-6). These residues are of all types: hydrophobic, polar, positively charged, and negatively charged. While several of these are located relatively distant from each other, others are rather close, and may interact with each other. The closely positioned residues were grouped together, giving one group of four residues, one of two residues, and four individual residues (see Figure 6-3). The residues within each group were optimized simultaneously under the same constraints as the individual residue optimizations. In both multiple residue groups, all the residues took on a positive charge in the optimum. While not strictly additive, the optimal energy for each group is a significant fraction of the sum of the optimal energies for each



	A	**B**	**C**	**D**	**E**	**F**	**All**
WT Ref.	−4.01	−4.42	−3.95	−3.62	−0.97	−7.70	−15.50
Hϕ Ref.	−2.83	−6.80	−3.81	−5.79	−4.57	−2.66	−18.87

Figure 6-3: **Optimization of peripheral residues on BLIP for binding to TEM1.** All BLIP residues identified as being capable of making “action-at-a-distance” interactions are shown on the structure of the BLIP–TEM1 structure. The two views are related by a 90° rotation, with the TEM1 structure transparent in the head-on view of the interface. Spatially clustered residues are displayed in the same color, with the results of the optimization of residues in each cluster (in kcal·mol^{-1}) also tabulated. The structural figures were prepared with MOLSCRIPT [87] and RASTER3D [105].

residue in the group. The group of Asp68 and Ser69, for example, gives an optimal improvement of 4.0 kcal·mol^{-1} over wild type, compared with improvements of 2.2 and 3.0 kcal·mol^{-1} for residue individually optimized. When the entire set of ten peripheral residues are optimized simultaneously, all residues still take on a net positive charge. The optimal improvement over the wild-type residues for this set is 15.5 kcal·mol^{-1}, and the improvement over all hydrophobic residues at these positions is 18.9 kcal·mol^{-1}.

6.4 Discussion

The largest contributions to the electrostatic portion of the BLIP–TEM1 binding free energy are, naturally, the charged residues in the center of the binding interface for which the large direct interactions across the binding interface more than compensate the desolvation penalty paid for burying a charged group. Polar, but uncharged, groups lining the interface pay a lower desolvation energy, but make up for this with similarly reduced direct interactions, and thus occupy a second tier of significant contributors to the electrostatics of binding. Contributing similar values as the polar residues buried on binding are several charged residues located on the periphery of the binding interface. Some of these residues are located within 4.0 Å of the binding partner, and thus are somewhat desolvated upon binding, whereas the residues further away have almost no desolvation cost associated with binding. The closer residues make stronger interactions, but even residues more than 10.0 Å from the binding partner make significant interactions. In all cases, however, the direct interaction term is significantly larger than the both the desolvation and the indirect interaction terms. Only two of these residues are positively charged, both arginines located on BLIP, and both of these contribute favorably to binding. All the other residues in this class (three on TEM1 and four on BLIP) are negatively charged, and all of these residues contribute unfavorably. Since these residues are not making short-range interactions such as hydrogen-bonds across the binding interface, the interaction term depends primarily on the general properties of the electrostatic potential generated by the binding partner. In the region surrounding all these residues, the potential produced by the other molecule is negative, and thus positively charged residues interact favorably while negatively charged residues make unfavorable interactions. While not forming tight clusters, the residues which make these types of interactions are not evenly distributed around the periphery of the interface. In particular, all but one of the residues on BLIP are located in one area (at the top of the complex in Figure 6-1), and one of the residues on TEM1 is located across from this group. The remaining residue on BLIP and the two other residues on TEM1 are similarly located across a region of solvent from each other. It is important to note that these results are all in the context of the wild-type structure; the negatively charged residues on each side of the interface contribute to the negative potential felt by the residues on the other side. Thus simultaneous mutations made to these residues on both TEM1 and BLIP may not have the same effect as the mutations made on a single molecule.

Surface residues are quite mobile, and thus treatment of these residues by a single conformation with an internal dielectric constant of 4.0 may lead to overestimation of some electrostatic effects. An internal dielectric constant of 20.0 has been suggested as value which reasonably accounts for the increased mobility of the surface of a protein, without requiring the sampling of multiple conformations [1, 2, 129]. While the energetic contributions of more buried residues change significantly with this treatment, the contributions of the residues on the periphery of the interface are much less affected. Since the interactions these residues make are through a region of solvent, the internal dielectric constant has little effect on these energies, and consistent results are seen with different treatments of the system. To further emphasize this, even when the analysis is done in a uniform dielectric of 80.0, similar results are obtained for the peripherally located residues. The interactions these residues make are through solvent, but despite the relatively high screening this causes, they are still able to contribute significantly in an energetic sense.

All the residues identified in the wild-type complex as acting through this “action-at-a-distance” mechanism are charged; electrostatic interactions for neutral polar residues are inherently smaller in magnitude than those of charged residues, and thus where a charged residue is found to have a moderate contribution, a polar residue may be found to have only a small contribution. However, even if the wild-type residue does not make a significant interaction, if the potential in the region of that residue is significant, a mutation to a charged group may enhance (or diminish) the binding affinity. By looking beyond the contribution of the wild-type charge distribution and the hydrophobic reference state, the electrostatic affinity optimization procedure is designed specifically to deal with this.

Quite a large number of residues show optimal improvements relative to wild type of more than 3.0 kcal·mol^{-1}, showing that there is considerable opportunity in this system to improve binding. The largest improvement is seen for Asp163 (7.7 kcal·mol^{-1} better than wild-type), which is also the most unfavorable component in the system, and many of the other residues showing the largest improvements are located in the same area. The majority of the remaining residues which show large improvements are located a layer behind the binding interface. Residues in this region comprise part of the core of the protein, and thus they may not be expected to have evolved to play a large role in binding; mutation of these residues to enhance binding would be likely to destabilize the native state of the inhibitor. The residues which make up the binding interface and which directly contact TEM1 show only small improvements upon binding; the residues in this region are near optimum in their wild-type state. Almost all residues favor a positively charged optimum, and a negatively charged group is the least favorable in all of the top ranking cases. TEM1 produces a negative electrostatic potential over the majority of BLIP, especially in the regions somewhat removed from the binding surface, and thus positive charges are, in general, favored.

The appearance of Asp163 as the residue showing the most room for improvement raises the question of the choice of reference state. This residue is highly unfavorable relative to a hydrophobic isostere, and the optimum must show at least the improvement of the hydrophobic replacement. This is a general property of the optimization, and thus a bias toward unfavorably contributing residues will often be seen in the residues showing the most improvement. In terms of design, this is not a drawback — all that is desired is to improve binding, and thus it makes sense to focus attention on those residues which contribute unfavorably. However, in trying to understand the general electrostatic properties of a complex, and in pinpointing the regions of a binding interface which are most important for binding, it may be more useful to consider the results of the optimization relative to a common hydrophobic reference state. Using this reference, there is little change for the neutral residues which showed large improvements, but large changes are seen for the charged residues. The improvement of Asp163 on optimization drops below 3.0 kcal·mol^{-1} when compared to a hydrophobic isostere, while Lys74 and Asp49, the two most favorably contributing residues, appear as having the greatest possibility of improvement (17.7 kcal·mol^{-1} and 11.2 kcal·mol^{-1} respectively). Both these residues are incredibly important in contributing to the electrostatic free energy of binding, but because the wild-type residues are near optimal, they only appear to be significant when a reference state other than wild type is used. The differences show up clearly on the structural localization of the results, with the highly contributing Asp49 located in the center of the binding interface, deeply buried into a pocket on TEM1. In terms of overall electrostatic interactions, the most important regions are buried at the binding interface, but for the design of improvements, regions somewhat removed from the interface are more worthy of focus, as these areas are the most sub-optimal in the wild-type state.

Given the identification of several charged residues on the periphery of the binding interface which act through solvent to contribute both favorably and unfavorably to binding, the question of how residues in these areas behave in the optimization arises. In addition to the six previously identified residues, four neutral residues (two serines, a leucine, and a valine) in these regions were selected, with optimal improvements between 0.8 and 4.0 kcal·mol^{-1} calculated. Three of these are located very close to one or more of the six charged residues, so these were clustered together in order to account for any cooperativity between these residues. While there is some reduction in the energetic effects of simultaneously optimizing these residues, even when all ten residues are optimized at once, every residue takes on a positive charge in the optimum. The benefit gained by a positive residue interacting with the negative potential produced by TEM1 is greater than any repulsions between the positive groups. With the exception of Arg160, whose wild-type charge distribution is near optimal, all the groups of residues show improvements of more than 3.5 kcal·mol^{-1} relative to wild type, and all groups including Arg160 show improvements over 2.5 kcal·mol^{-1} relative to the hydrophobic reference. While not strictly additive relative to the improvement possible for each group, the simultaneous optimization of these ten residues results in an improvement in binding free energy of 15.5 kcal·mol^{-1}. It is thus clear that large improvements in binding affinity may be gained by optimizing the "action-at-a-distance" interactions made by these surface residues on the periphery of the binding interface.

Previous work by Selzer *et al.* studied the effects on binding of several mutations to BLIP, including several of the peripheral residues discussed here [132]. A single mutant of Asp 163 to alanine improves binding by ten-fold, while a mutation of the same residue to lysine improves binding by 28-fold. Adding a mutation of both Val165 (one of the peripheral residues) and Asn89 to lysine alongside the mutation of Asp163 to lysine results in an improvement in binding of 57-fold. A triple mutant of three peripheral residues (Val134, Asp135, and Asp163) all to lysine binds 170 times better than wild-type, and a quadruple mutant with three peripheral residues (Asp135, Asp163, and Val165) and one other residue (Asn89) all mutated to lysine shows 290-fold improved binding. While two of these mutants include a residue not included in the "action-at-a-distance" set of peripheral residues, the Asn89 to lysine single mutant was found to improve binding by only two-fold, and thus it remains clear that the mutations of the peripheral residues have a significant effect. The electrostatic complementarity of these mutants to TEM1 was studied using residual potentials, and this work is outlined in Appendix A. The mutants with more favorable binding affinities were all found to be more electrostatically complementary both by visual and quantitative analysis of the surface potentials. These results strengthen the conclusion that these peripheral residues are able to project an electrostatic effect through a region of solvent and significantly affect the binding affinity.

6.5 Conclusions

A detailed computational analysis of the electrostatic contributions to the energetics of binding of β-lactamase inhibitor protein (BLIP) to the TEM1 β-lactamase was performed using methods based on continuum electrostatics. While the most electrostatically significant residues on both proteins are located in the center of the binding interface, several charged residues located on the periphery of the binding interface were also identified as making significant contributions, both favorable and unfavorable, to the energetics of binding. All of these residues are somewhat solvent exposed, even in the bound state, and while the closest contacting residues could be involved in direct favorable interactions, residues as far as 10.0 Å from the binding partner still make significant interactions. The energetic importance of these peripheral residues is fairly insensitive to variations in the internal dielectric constant used in the continuum electrostatic calculations — even when fully screened by solvent, using a uniform dielectric of 80.0, contributions of over 1.0 $\text{kcal}\cdot\text{mol}^{-1}$ are seen for the furthest participating residues.

Optimization of the partial atomic charges on the side-chain atoms of each residue of BLIP to give the best electrostatic binding free energy shows that, while relative to hydrophobic reference state the most important region is the center of the binding interface, relative to the wild-type charges the most improvement is to be gained on residues clustered in one region near the periphery of the binding interface. Many of the top improvements are seen for residues which are somewhat buried in the unbound state, and thus mutation of these residues is likely to destabilize the protein. However, two of the closer contacting solvent exposed residues identified through the analysis of the wild-type structure are among those which show the greatest improvements in binding free energy on optimization. By analysis of the structure in parallel with the energetics of optimization, a set of ten residues poised to make these "action-at-a-distance" interactions was identified, four somewhat isolated from one another and the remaining six in two clusters. Simultaneous optimization of all these residues suggests that all residues could make favorable interactions without substantially interfering with the effects of the others.

The results indicate that electrostatic interactions involving solvent exposed residues on the periphery of a protein–protein binding interface can make significant energetic contributions to the binding affinity. As geometric packing considerations for residues which remain on the surface even in the bound state are much less complicated than for interfacial residues, this may provide an alternative design procedure to methods involving a detailed consideration of packing. Furthermore, as these interactions act over moderate distances, the computed effects should be much less sensitive to local imperfections in structural models than are those for short-range interactions. While the study here has only included one system, work is ongoing to extend these results to other protein–protein complexes in order to generalize both the occurrence of these “action-at-a-distance” interactions in natural systems, and to further the application of these interactions in a design protocol.

Chapter 7

Ab initio Charge Determination: Comprehensive Evaluation of Methodologies in Continuum Electrostatic Calculations

Abstract

In order for continuum electrostatic calculations to give accurate results, an appropriate description of molecular charge distributions — most typically partial atomic charges — is necessary. While for some systems, such as biological macromolecules, sets of charges have been parameterized based on experimental data, for many other cases, *ab initio* methods of charge determination may be preferred.

Presented here is a comprehensive evaluation of the performance of numerous methods for the *ab initio* determination of partial atomic charges in continuum electrostatic calculations. Charges were computed using several methods based both on fitting electrostatic potentials and on population analysis, and using various levels of theory ranging from semi-empirical quantum mechanical methods through relatively high level *ab initio* quantum mechanical theories. All charge distributions were evaluated in terms of their ability to reproduce experimental free energies of solvation in the context of a continuum solvation model. Two sets of molecules were used, one derived from the groups seen in proteins, and the other a more diverse set of neutral organic molecules.

The results indicate that there are clearly preferred methods for determining charges, and conversely that there are highly disfavored methods. The agreement with experiment does not increase with increasing levels of theory, although the lowest level methods do perform particularly poorly. None of the methods performed uniformly well across all molecule types, with the top performing methods tending to give charge magnitudes in the middle of the observed range, but both the under- and over-polarized charge distributions performing better for certain systems. The frequently used HF/6-31G* level of quantum mechanics does very well, ranking in the top methods, particularly when coupled with the Merz–Singh–Kollman charge fitting scheme or a restrained fit based on this scheme.

7.1 Introduction

Continuum solvation models have, over the past two decades, been shown to be very useful in gaining important insights into biomolecular processes, as continuum models allow the solvation energetics of biological macromolecules in the aqueous, moderate ionic strength environment which is the milieu for the majority of biology to be calculated relatively quickly [56, 70, 137, 158]. Continuum electrostatic calculations have been used to analyze in detail the role that electrostatic interactions play in the stability of proteins [54, 65, 68, 141, 164], and to further our understanding of the binding energetics of proteins with other proteins, with nucleic acids, and with small molecules [3, 51, 69, 106, 107, 113, 166]. In addition, theoretical and methodological advances have made it possible to use continuum electrostatics as a tool in designing more tightly and specifically associating molecular complexes [80, 92].

An essential requirement for the successful application of continuum electrostatics is an appropriate description of the molecular charge distribution, which is most commonly represented as a set of atom-centered point charges for the molecules of interest. Whereas there are extensive parameter sets including charges for biological macromolecules readily available [11, 29, 73, 100, 140, 152, 159, 160], equally accurate charge models for the small molecules that bind to them is frequently lacking. A great deal of success has been found in fitting chemical parameters to physical data, and most parameter sets for use in molecular mechanics force fields have been determined at least partially in this manner. Where physical data is unavailable, results from *ab initio* quantum mechanical calculations are often used for parameterization. However, it is not obvious that the same set of charges will give the best performance both in molecular mechanics and in continuum electrostatic calculations, and in fact there is significant variation among the charges found in diverse empirical force fields. A recent study also showed that the parameters from several major empirical force fields reproduce experimental hydration free energies quite poorly when used to compute solvation free energies using a continuum electrostatic model [39]. Sitkoff *et al.* were successful in parameterizing the charges on a small set of functional groups found in proteins to give good agreement to experiment in the context of a continuum solvation model [140]. This parameter set, PARSE, is extremely useful for proteins but does not include a sufficient range of functional groups to describe many small molecules, and thus an alternative method for determining partial atomic charges for small molecules is desirable.

Ab initio charge determination methods are of particular interest, since detailed experimental data is not available for many known ligands. In addition, in the context of *de novo* ligand design, the molecule of interest may have no experimental information available at all, and in fact may have never been synthesized. Several *ab initio* methods for the determination of the partial atomic charges of small molecules exist, based both on analysis of the electron density [110] and on fitting point charges to reproduce the electrostatic potential around the molecule [6, 7, 10, 22, 28, 62, 63, 109, 139]. However the best choice of charge determination method, as well as the most appropriate quantum mechanical level of theory and size of basis set, is not clear. While the performance of *ab initio* charge determination methods in molecular mechanics applications has been analyzed [19, 112], and the performance of various parameterized charges in continuum electrostatic calculations has also been considered [39, 140], there has been little consideration of the performance of *ab initio* methods in continuum electrostatic applications. The goal of this work is to analyze in detail the performance of *ab initio* charge determination methods in a continuum solvation model.

7.2 Methods

Small molecule geometries. The structures of all molecules were energy minimized using the quantum chemistry programs JAGUAR [130] or GAUSSIAN98 [50] for all *ab initio* methods, and using the program MOPAC [142] for all semi-empirical methods, starting from an extended conformation with standard bond lengths and angles.

Small molecule partial atomic charges. Partial atomic charges were determined in numerous ways from the wavefunction calculated by a single-point calculation using the quantum chemistry program GAUSSIAN98 [50]. Two levels of theory — HF and B3LYP — and a variety of basis sets — STO-3G, 3-21G, 4-31G, 6-31G, 6-31G(**)(++) and 6-311G(**)(++) — were used. Charges were obtained by Mulliken population analysis [110], as well as by fitting the electrostatic potential (ESP) using the Chelp procedure [22], the ChelpG procedure [10] and the Merz–Singh–Kollman (MK) method [7, 139]. Additionally, an enhanced Merz–Singh–Kollman procedure was performed with an increased size and density of the grid used for the determination of the electrostatic potential. As well as the standard ESP fit, a restrained fit to the potential was performed using the program RESP [6, 28], in 3 ways: (1) a single fit with weak restraints toward zero on all heavy atoms; (2) a two-stage fit with weak restraints on all heavy atoms in the first stage, followed by a second fit with aliphatic carbons more highly restrained and all polar atoms fixed at the values obtained in the first stage; (3) a single fit with aliphatic hydrogens fixed at zero, and all heavy atoms weakly restrained. The restraints used were those suggested by Bayly *et al.* [6]. Charges were fit from the semi-empirical wavefunction calculated by the MOPAC computer program [142] using population analysis, as well as the Merz–Singh–Kollman ESP fitting scheme.

Solvation free energy calculations. Solvation free energies were calculated using a two-component Poisson–Boltzmann/Surface Area (PB/SA) procedure previously described [140]. The electrostatic component was computed by finite-difference solution of the linearized Poisson-Boltzmann equation, using a locally modified version of the computer program DELPHI [55, 57, 134, 136]. A 65 × 65 × 65 grid was used, with focusing boundary conditions in which the longest dimension of the molecule occupies first 23%, then 46%, and finally 92% of one edge of the grid. This results in a final grid spacing of at most 0.33 Å for all molecules. The boundary potentials for each calculation were taken from the previous resolution calculation, and Coulombic potentials were used at the boundary of the lowest resolution box. An internal dielectric constant of 2 was used, and a dielectric of 80 was used for the solvent. The ionic strength was set to zero for consistency with the experimental conditions. The non-polar (cavity and van der Waals) term was calculated from the solvent accessible surface area (calculated using the program MSMS [128]) using the relation $\Delta G = 5.4A + 920$ (ΔG in cal·mol^{-1} and A in Å^2) [140]. A probe radius of 1.4 Å was used for the generation of both the molecular surface (used to define the dielectric boundary) and the solvent accessible surface.

7.3 Results

7.3.1 Molecules representative of protein groups

An initial extensive set of calculations was performed on a set of molecules corresponding to the side chains of the twenty common amino acids with the exception of proline and glycine, as well as a small molecule representation of the peptide backbone. Both charged and neutral states were considered for all ionizable groups. Charges and geometries were obtained for a variety of basis sets and theoretical methods, and these charges were subsequently used in the calculation of solvation free energies using a Poisson–Boltzmann/Surface Area model. The radii from the PARSE parameter set [140] were used for all computations; these radii have the advantage of being quite simple — they are an extension of the Pauling van der Waals radii with the radius of hydrogen atoms set to 1.0 Å rather than 1.2 Å.

The average absolute error in calculated hydration free energies in comparison to experiment was determined for each charge set (see Table 7-1). Considering the results, several observations can be made. The minimal STO-3G basis set, at both HF and B3LYP levels of theory, gives very poor performance in all charge fitting methods, with average errors of over 3.0 kcal$\cdot$mol^{-1} in every case. Mulliken charges also perform very poorly at almost all levels of theory, with everage errors below 2.0 kcal$\cdot$mol^{-1} in only a few cases, and in no case giving an error below 1.5 kcal$\cdot$mol^{-1}. The poor performance of Mulliken charges relative to the other charge determination methods is not particularly surprising, since while the other methods are designed to reproduce the electrostatic potential outside the molecule, which is particularly relevant for solvation free energy calculations, the Mulliken charges are obtained by a partitioning of the electron density within the molecule, with no regard for reproducing the electrostatic potential.

At the B3LYP level of theory, several basis sets perform relatively poorly across all fitting methods based on the electrostatic potential (ESP). These include 3-21G, as well as 6-31G and 6-311G with polarization functions either on heavy atoms or on all atoms but with no diffuse functions. The 6-31G and 6-311G basis sets with neither polarization nor diffuse functions perform somewhat better. The best performance is given by the 4-31G basis set and by the 6-31G and 6-311G basis sets with both polarization and diffuse functions on heavy atoms (and optionally on hydrogens).

The results at the HF level of theory are almost exactly opposite. The worst performance is seen in the 4-31G basis set, and in the 6-31G and 6-311G basis sets

Method	*Basis*	*Mulliken*	*Chelp*	*ChelpG*	*MK ESP*	*MK ESP**	*RESP 1X*	*RESP 2X*	*RESP 2X**	*RESP PH*
B3LYP	STO-3G	5.47	4.30	4.09	3.97	3.97	4.15	4.16	3.98	4.22
B3LYP	3-21G	1.91	1.84	1.55	1.49	1.53	1.56	1.56	1.53	1.66
B3LYP	4-31G	1.79	1.58	1.09	0.95	0.95	0.99	1.00	0.94	1.08
B3LYP	6-31G	1.92	1.48	1.25	1.14	1.13	1.08	1.08	1.12	1.17
B3LYP	6-31G*	1.87	2.10	1.56	1.39	1.37	1.58	1.59	1.37	1.69
B3LYP	6-31G*$^{+}$	3.17	1.63	1.26	1.06	1.03	0.99	0.99	1.02	1.11
B3LYP	6-31G**	3.17	2.18	1.62	1.44	1.43	1.66	1.67	1.43	1.78
B3LYP	6-31G**$^{+}$	1.96	1.72	1.28	1.07	1.04	1.02	1.03	1.04	1.15
B3LYP	6-31G**$^{++}$	3.00	1.76	1.28	1.06	1.04	1.02	1.02	1.03	1.16
B3LYP	6-311G	2.14	1.39	1.30	1.21	1.23	1.12	1.12	1.22	1.15
B3LYP	6-311G*	2.21	1.98	1.48	1.25	1.26	1.36	1.36	1.26	1.47
B3LYP	6-311G*$^{+}$	3.62	1.65	1.20	1.02	1.02	0.93	0.93	1.02	1.05
B3LYP	6-311G**	3.99	2.19	1.62	1.39	1.38	1.59	1.60	1.37	1.71
B3LYP	6-311G**$^{+}$	3.36	1.83	1.26	1.03	1.00	1.08	1.08	1.00	1.20
B3LYP	6-311G**$^{++}$	4.24	1.84	1.27	1.04	1.01	1.09	1.09	1.00	1.22
HF	STO-3G	5.22	3.67	3.52	3.42	3.42	3.62	3.63	3.43	3.67
HF	3-21G	2.95	1.18	1.10	1.26	1.30	1.06	1.06	1.30	1.13
HF	4-31G	2.52	1.66	1.76	1.66	1.68	1.47	1.47	1.67	1.44
HF	6-31G	2.57	1.75	1.87	1.78	1.79	1.58	1.58	1.78	1.54
HF	6-31G*	1.67	1.29	1.14	1.01	1.02	0.86	0.86	1.01	0.99
HF	6-31G*$^{+}$	4.19	1.41	1.47	1.41	1.46	1.17	1.18	1.45	1.32
HF	6-31G**	1.49	1.30	1.17	1.01	1.02	0.88	0.89	1.01	1.01
HF	6-31G**$^{+}$	2.69	1.46	1.48	1.37	1.42	1.16	1.17	1.40	1.31
HF	6-31G**$^{++}$	2.91	1.47	1.48	1.35	1.42	1.17	1.17	1.41	1.30
HF	6-311G	1.80	1.68	1.88	1.80	1.84	1.59	1.59	1.83	1.56
HF	6-311G*	1.83	1.30	1.14	1.13	1.16	0.92	0.92	1.15	1.04
HF	6-311G*$^{+}$	5.11	1.40	1.43	1.46	1.51	1.20	1.21	1.50	1.32
HF	6-311G**	3.11	1.50	1.21	1.03	1.06	0.93	0.93	1.05	1.06
HF	6-311G**$^{+}$	3.30	1.52	1.46	1.32	1.36	1.13	1.14	1.36	1.25
HF	6-311G**$^{++}$	5.54	1.62	1.42	1.28	1.32	1.09	1.09	1.31	1.21
S.E.	AM1	3.18	-	-	2.57	-	-	-	-	-
S.E.	PM3	3.89	-	-	3.04	-	-	-	-	-
S.E.	MNDO	4.50	-	-	2.59	-	-	-	-	-
S.E.	MINDO3	4.54	-	-	3.99	-	-	-	-	-

Table 7-1: **Errors in calculated hydration free energies of molecules in the protein dataset.** Average absolute errors (in kcal·mol^{-1}) in calculated hydration free energies for all charge determination methods are shown computed over a set of molecules representative of protein groups. Methods marked with * used an extended and more dense grid for the calculation of the electrostatic potential.

with no polarization or diffuse functions. The 6-31G and 6-311G basis sets with both polarization and diffuse functions on heavy atoms (or on all atoms) yield moderate performance. The 6-31G and 6-311G basis sets with polarization functions on heavy atoms and optionally on hydrogens, but with no diffuse functions, perform best, closely followed by the 3-21G basis set.

The Chelp method of fitting to the electrostatic potential gave charges which perform somewhat poorly at all levels of theory. The ChelpG method gives slightly better average performance across basis sets, but the Merz–Singh–Kollman method outperforms both in all but one case (HF/3-21G), for which all ESP fitting methods perform quite well, and for which the Chelp and ChelpG methods perform better than with any other theoretical level.

Restrained ESP charge fitting was also carried out with the charges of aliphatic hydrogen atoms constrained to zero, and with these hydrogens assigned a radius of zero. This gives a charge set consistent with a polar hydrogen/united non-polar atom model. In all but one case (HF/4-31G), this polar hydrogen model gives average errors slightly worse than the analogous all-atom set. In the case of HF/4-31G the average errors of both models are virtually identical.

The extension of the Merz–Singh–Kollman grid to include more layers, as well as to sample each layer more finely, does not result in better performance of the fit charges. For unrestrained ESP fit charges, the greatest deviation in average error between the standard grid and the more extensive grid is 0.07 $\text{kcal}\cdot\text{mol}^{-1}$, and is less than 0.05 $\text{kcal}\cdot\text{mol}^{-1}$ for the majority of methods. With charges obtained from restrained ESP fits, the more extensive grid yields charges which perform more poorly on average. This difference is likely a result of the implementation of restraints in the RESP method, which become relatively weaker as the number of points at which the potential is calculated increases [63].

In addition to the geometry at which the single-point calculation was performed, charges were determined from single-point calculations at all levels theory using the

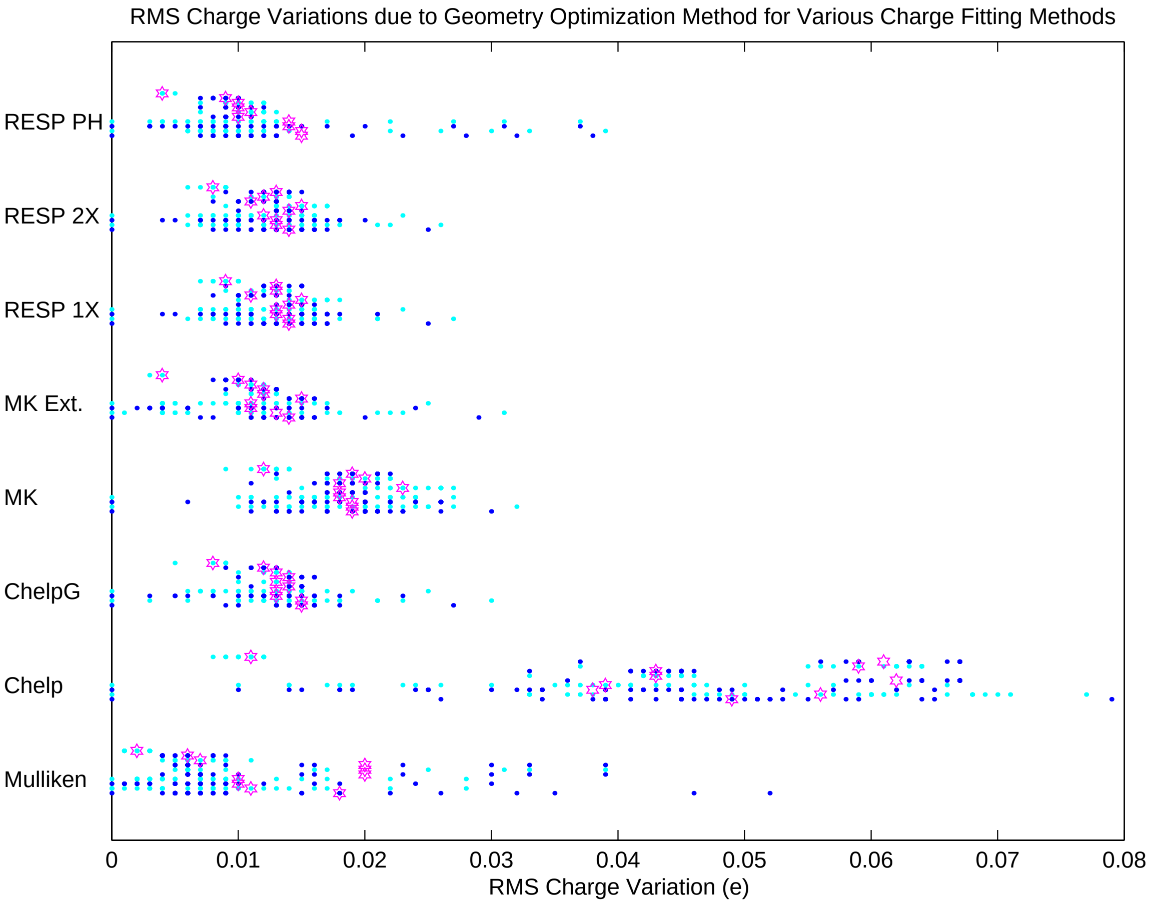


Figure 7-1: **Variation in computed partial atomic charges with geometry.** The variation of partial atomic charges computed from geometries determined at various levels of theory are plotted. For each charge fitting method, the RMS deviation between the charges determined using the geometry at one of several levels of theory with those determined using the geometry determined at another level of theory was computed for every molecule in the protein dataset. By far the greatest variation is seen for the Chelp ESP fitting methodology.

geometries computed at the HF/3-21G, the HF/6-31G*, the B3LYP/6-31G^{*+}, and the B3LYP/6-311G^{**++} levels of theory. These choices cover the range of methods quite well. The charges obtained from the different geometries were compared in detail (Figure 7-1). The charge determination method was seen to play a more significant role in the variation of calculated charges with geometry than did the differences between the basis sets used to calculate the geometry and to calculate the charges. In particular, the Chelp procedure produced the largest variations in charges obtained from different geometries, with RMS deviations of the charges of 0.05e and a maximum deviation of over 1e! The Mulliken procedure also produced deviations in charge of up to 0.6e, and the united non-polar atom charges showed variations of up to 0.5e. The ChelpG procedure, the unrestrained Merz–Singh–Kollman procedure, and the RESP procedure all gave maximum deviations of between 0.1 and 0.2e, and RMS deviations of between 0.01 and 0.02e.

Charges derived from semi-empirical methods perform worse than those from all *ab initio* basis sets with the exception of STO-3G. The semi-empirical population analysis charges all give average errors of about 3.0 kcal·mol^{-1}. While the electrostatic potential fit charges reproduce the experimental hydration free energies better than those computed from population analysis, they still do quite poorly. Only two methods give errors below 3.0 kcal·mol^{-1} — AM1 and MNDO. For semi-empirical methods, only population analysis and Merz–Singh–Kollman ESP fit charges were obtained. However, considering the relatively small variation among ESP fit charges from different procedures in the *ab initio* data, it is unlikely that different ESP fitting schemes would drastically improve the performance of these semi-empirical methods.

7.3.2 Extended set of small organic molecules

A similar evaluation of the performance of the various charge fitting methods was performed using a more extensive set of small molecules. This set included 324 small molecules of diverse functionalities — 228 with a single functionality and 96 polyfunctional molecules [15]. The STO-3G basis set both at the Hartree–Fock and B3LYP levels of theory was excluded due to its very poor performance on the protein data set. Similarly, the Mulliken charge fitting scheme was excluded, again due to poor performance in the first set of calculations. While the Chelp method did not perform nearly as poorly as these in the initial calculations, it did not perform particularly well either. In addition, the Chelp methodology gave such large variations in charges when different geometries were used that is seems a poor choice for a general method. Therefore, the Chelp procedure was also excluded from this set of calculations.

The results of the broad survey of methods (Table 7-2) are qualitatively similar to the results from the protein set. At the Hartree–Fock level, the charges obtained using the 6-31G* and 6-31G** basis sets reproduce the experimental solvation free energies the best, with charges from an unrestrained Merz–Singh–Kollman fit to the electrostatic potential performing slightly better than those obtained from restrained fits with all atoms. The solvation free energies calculated with charges obtained from both the restrained fit with non-polar hydrogens fixed at zero and the ChelpG method reproduced the experimental values more poorly. At the B3LYP level of theory, charges from fitting the ESP obtained with several basis sets all reproduce the experimental solvation free energies quite well. The 6-31G and 6-311G basis sets, either with no diffuse or polarization functions, or with both diffuse and polarization functions on heavy atoms only, as well as both the smaller 3-21G and 4-31G basis sets, all perform well, and again the Merz–Singh–Kollman ESP fitting scheme produces the charges which reproduce the experimental results best, with restrained ESP fitting with all atoms included producing charges which do only slightly worse. Both the ChelpG charges and those from a restrained fit with non-polar hydrogens excluded do significantly worse with all these basis sets. All these top methods give average errors relative to experiment below 1.35 kcal·mol^{-1}, with the lowest average error being 1.24 kcal·mol^{-1} for charges obtained by unrestrained Merz–Singh–Kollman fitting to the B3LYP/6-311G^{*+} electrostatic potential. Several other methods based on the B3LYP/6-311G basis set, with variation of diffuse and polarization functions, produce charges which reproduce experiment with average errors of 1.30 kcal·mol^{-1}, but the restrained fits at these levels of theory do worse.

The monofunctional compounds can, of course, easily be classified into molecular classes, and the performance of the charge determination methods evaluated by class. The number of molecules included for each class ranges from as few as one to as many

Method	*Basis*	*MK ESP*	*RESP 1X*	*RESP 2X*	*RESP Polar H*	*ChelpG*
B3LYP	3-21G	1.25	1.30	1.32	1.48	1.34
B3LYP	4-31G	1.25	1.28	1.29	1.37	1.37
B3LYP	6-31G	1.26	1.28	1.29	1.36	1.38
B3LYP	6-31G*	1.41	1.54	1.55	1.69	1.63
B3LYP	6-31G*+	1.25	1.31	1.32	1.45	1.47
B3LYP	6-31G**	1.46	1.58	1.60	1.74	1.67
B3LYP	6-31G**+	1.27	1.34	1.34	1.47	1.50
B3LYP	6-31G**++	1.30	1.37	1.38	1.48	1.53
B3LYP	6-311G	1.30	1.30	1.31	1.36	1.42
B3LYP	6-311G*	1.30	1.39	1.40	1.53	1.54
B3LYP	6-311G*+	1.24	1.29	1.30	1.44	1.46
B3LYP	6-311G**	1.40	1.50	1.51	1.64	1.65
B3LYP	6-311G**+	1.30	1.37	1.38	1.50	1.54
B3LYP	6-311G**++	1.30	1.37	1.38	1.50	1.54
HF	3-21G	1.61	1.52	1.53	1.54	1.42
HF	4-31G	1.87	1.78	1.78	1.72	1.91
HF	6-31G	1.96	1.86	1.86	1.78	1.98
HF	6-31G*	1.28	1.31	1.31	1.42	1.43
HF	6-31G*+	1.41	1.39	1.39	1.50	1.52
HF	6-31G**	1.29	1.32	1.32	1.44	1.45
HF	6-31G**+	1.41	1.40	1.40	1.51	1.54
HF	6-31G**++	1.42	1.41	1.41	1.51	1.55
HF	6-311G	1.92	1.82	1.82	1.76	1.95
HF	6-311G*	1.39	1.38	1.38	1.50	1.47
HF	6-311G*+	1.48	1.43	1.43	1.54	1.54
HF	6-311G**	1.38	1.39	1.39	1.50	1.53
HF	6-311G**+	1.44	1.42	1.43	1.53	1.57
HF	6-311G**++	1.45	1.43	1.44	1.53	1.58

Table 7-2: **Errors in calculated hydration free energies of diverse small organic molecules.** Average absolute errors (in kcal·mol^{-1}) in calculated hydration free energies relative to experiment for all charge determination methods are shown computed over a large set of small organic molecules.

		Best Method		
Class	N	Quantum Mechanics	ESP Fitting	Error
Monofunctional				
Alkanes	25	RHF / 4-31G	RESP-2X	0.44
Alkenes	22	B3LYP / 6-31G**	ChelpG	0.52
Alkynes	8	B3LYP / 4-31G	ChelpG	0.36
Aromatics	27	RHF / 6-311G	MK	0.39
Alcohols	25	RHF / 6-311G	RESP-PH	0.44
Ethers	12	RHF / 4-31G	MK	0.50
Aldehydes	8	B3LYP / 6-311G	RESP-2X	0.11
Ketones	15	RHF / 6-311G**	ChelpG	0.28
Carboxylic Acids	3	B3LYP / 6-31G*	MK	0.06
Esters	28	B3LYP / 6-31G**	MK	0.31
Amines	20	RHF / 6-31G*+	MK	2.87
Pyridines	15	RHF / 6-311G	MK	0.45
Nitriles	3	B3LYP / 6-311G*	RESP-2X	0.10
Amides	1	B3LYP / 6-311G**++	ChelpG	0.01
Nitro	3	B3LYP / 6-31G**	RESP-2X	0.09
Fluorocarbons	1	B3LYP / 6-311G*+	MK	0.00
Chlorocarbons	8	RHF / 6-31G**++	RESP-1X	0.15
Thiols	2	RHF / 4-31G	RESP-2X	0.05
Thioethers	2	RHF / 6-311G	MK	0.59
Overall	228	B3LYP / 6-311G*+	MK	1.15
Polyfunctional				
Aliphatic	52	RHF / 6-31G**	MK	1.50
Unsaturated	14	B3LYP / 4-31G	ChelpG	0.84
Aromatic	30	B3LYP / 3-21G	RESP-1X	0.96
Overall	96	B3LYP / 3-21G	RESP-1X	1.32
All molecules	324	B3LYP / 6-311G*+	MK	1.24

Errors are average absolute errors in kcal·mol^{-1}.

Table 7-3: **Best performing charge determination methods by molecule class.** The charge determination method producing the smallest average error in calculated solvation free energies is listed for each molecular class in the extended set of small organic molecules. Average absolute errors are given in kcal·mol^{-1}.

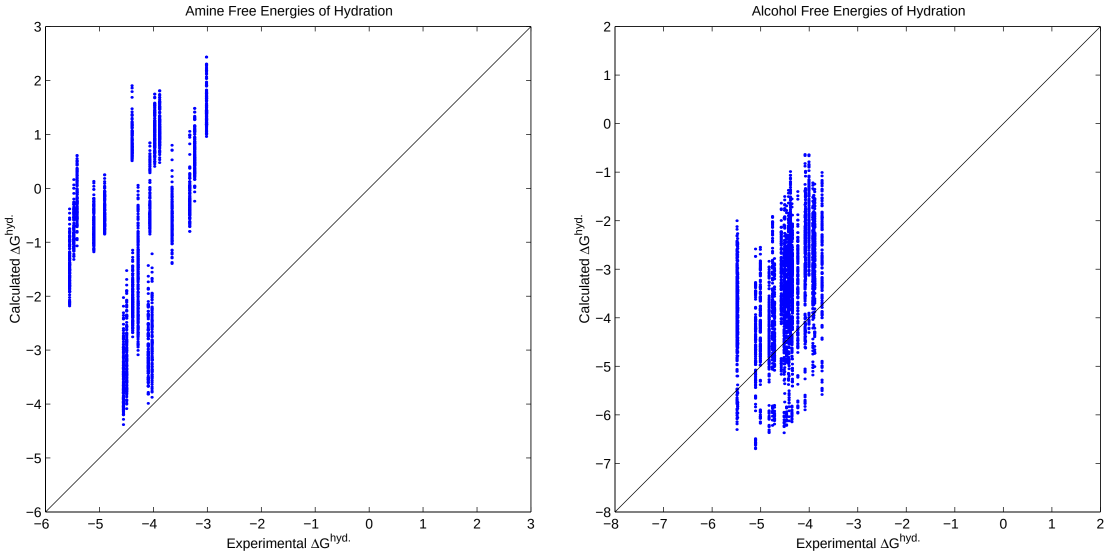


Figure 7-2: **Variation in computed hydration free energies for amines and alcohols.** For both the set of all amines and of all alcohols, the calculated hydration free energy (in kcal·mol^{-1}) is plotted relative to the experimental value. While a similar range is seen in the computed values for both sets, the results for alcohols span the line of $y = x$, while the computed energies for amines are uniformly higher than the experimental values.

as twenty-eight. With the exception of the amines, the best performing method for every class gives an average error of less than 0.60 kcal·mol^{-1}. However, the method for charge determination which is best for each class varies significantly, both in the potential fitting method and in the quantum mechanical level of theory used to generate the ESP. Over all monofunctional compounds, the best method is the same as for the full set (B3LYP/6-311G^{*+} electrostatic potentials with Merz-Singh-Kollman charge fitting), with an average error of 1.15 kcal·mol^{-1}. However, the average error taking the best method for each class is 0.41 kcal·mol^{-1} when each molecule is weighted equally, and 0.59 kcal·mol^{-1} when each class is weighted equally.

The set of amines do very poorly, with the best method giving an average error of 2.89 kcal·mol^{-1}. Looking at how the methods perform as a whole (Figure 7-2), it is clear that all methods underestimate the favorable free energy of solvation; the calculated free energy of hydration is greater than the experimental value for every method and for every molecule. In addition, the range of performance of each molecule differs significantly, with some molecules having some methods which reproduce experiment well, while others do poorly with all methods. Comparatively, the alcohol set (whose best performing method gave an average error of 0.44 kcal·mol^{-1}) behaves in a qualitatively different manner. For this set, every molecule has some methods which do well, and both positive and negative errors are seen for every molecule as well.

The polyfunctional molecules, which all include at least two functional groups, but possibly two of the same type, can loosely be grouped into aliphatic, unsaturated, and aromatic compounds, with any molecule containing an aromatic group considered aromatic (even if it also contains other carbon types as well) and a molecule containing any number of non-aromatic double or triple bonds being classified as unsaturated. The best performing methods for the three groups of polyfunctional molecules reproduce experimental values with average errors of 1.50 kcal·mol^{-1} for the aliphatic group, 0.84 kcal·mol^{-1} for the unsaturated group, and 0.96 kcal·mol^{-1} for the aromatic group, with the methods producing the best results for each set again being quite different. The best performing method for the full set of polyfunctional compounds is single-stage restrained fitting to the B3LYP/3-21G potential, one of the highly performing methods on the full set of molecules, with an average error of 1.32 kcal·mol^{-1} relative to experimental values. The average error using the best method for each class is 1.24 kcal·mol^{-1} with molecule based weighting, and 1.10 kcal·mol^{-1} with group based weighting.

The amount of data to consider in looking at the performance of all methods for every molecule class is too large to be feasible. However, the performance of a select set of methods over all classes is displayed in Table 7-4. The methods all use two-stage RESP fit charges and are based on the Hartree–Fock potentials with a 3-21G, 6-31G*, or 6-31G^{*+} basis set, or the B3LYP potentials with a 4-31G, 6-31G*, or 6-31G^{*+} basis set. These include methods which both perform well and which perform poorly at both the HF and the B3LYP levels, and include both small

			RHF			B3LYP		
Class	N	Best	3-21G	6-31G*	6-31G*+	4-31G	6-31G*	6-31G*+
Monofunctional								
Fluorocarbons	1	0.00	0.48	0.15	0.22	0.18	0.61	0.04
Thiols	2	0.05	0.26	0.60	0.47	0.32	0.72	0.68
Alkanes	25	0.44	0.44	0.54	0.54	0.54	0.50	0.54
Nitriles	3	0.10	0.24	0.78	1.30	0.15	0.15	0.71
Chlorocarbons	8	0.15	1.40	0.25	0.26	0.16	0.64	0.42
Ketones	15	0.28	0.45	0.47	0.51	0.73	1.68	0.54
Aldehydes	8	0.11	0.16	0.20	1.06	0.39	1.30	0.16
Amides	1	0.01	0.03	0.05	1.52	0.54	1.80	0.09
Alkenes	22	0.52	1.11	0.88	1.14	0.67	0.70	0.89
Alcohols	25	0.44	0.64	1.33	0.61	0.70	2.23	1.08
Aromatics	27	0.39	1.20	0.64	0.57	1.49	1.54	1.16
Nitro	3	0.09	2.70	1.97	2.61	1.21	0.09	1.16
Esters	28	0.31	2.54	1.55	2.28	1.50	0.32	1.23
Carboxylic Acids	3	0.06	3.15	1.72	2.66	1.88	0.36	1.12
Alkynes	8	0.36	1.67	1.67	2.03	0.83	1.15	1.53
Thioethers	2	0.59	0.99	1.53	1.44	1.57	1.71	1.71
Ethers	12	0.50	1.18	2.09	1.88	1.43	2.61	2.07
Pyridines	15	0.45	1.35	2.16	1.64	2.46	3.16	2.40
Amines	20	2.87	3.18	3.37	3.04	3.47	3.72	3.26
Overall	228	1.15	1.36	1.29	1.32	1.26	1.52	1.27
Polyfunctional								
Aliphatic	52	1.50	1.90	1.57	1.57	1.64	1.85	1.63
Unsaturated	14	0.84	1.82	0.97	1.50	0.91	0.99	1.01
Aromatic	30	0.96	2.04	1.23	1.56	1.09	1.51	1.25
Overall	96	1.32	1.93	1.38	1.56	1.36	1.62	1.42
All molecules	324	1.24	1.53	1.31	1.39	1.29	1.55	1.32

Errors are average absolute errors in kcal·mol^{-1}.
All numbers based on RESP-2X charge fitting method.

Table 7-4: **Performance of select charge determination methods by molecule class.** The average error in calculated free energies of hydration (in kcal·mol^{-1}) are displayed for each molecular class for a select set of charge determination methods. The results are roughly grouped according to the average performance across the set of methods. Where only a few methods gave poor performance, these are indicated in red, and where only a few methods performed well, these are indicated in green.

and moderately sized basis sets. For several molecule classes (alkanes, fluorocarbons, and thiols) all the methods give average errors below 1.0 kcal·mol^{-1}, although the fluorocarbon and thiol set contain only one and two molecules respectively. For three additional classes (ketones, nitriles, and chlorocarbons), only one method gives an average error of above 1.0 kcal·mol^{-1}, with B3LYP/6-31G* performing badly on ketones, H/6-31G^{*+} performing badly on nitriles, and HF/3-21G performing badly on chlorocarbons. Charges from both the 3-21G and 6-31G^{*+} basis sets at the Hartree–Fock level perform poorly on alkenes, while charges from both HF/6-31G^{*+} and B3LYP/6-31G* potentials perform poorly on both aldehydes and the single amide in the set. Charges from the Hartree–Fock level of theory with both the 3-21G and the 6-31G^{*+} basis sets, as well as from the B3LYP level with the 4-31G basis set, perform well on alcohols, while charges from B3LYP with the 6-31G* basis set do particularly poorly. For aromatic molecules, only two methods, Hartree–Fock with either the 6-31G* or the 6-31G^{*+} basis set, give average errors below 1.0 kcal·mol^{-1}, and five additional molecule classes have only one of the methods which performs well. For esters, nitro compounds, and carboxylic acids, B3LYP/6-31G* charges are the only ones which do well, while for alkynes, the only method which produces charges which do well is that using B3LYP with the 4-31G basis set. Only charges from HF/3-21G potentials do reasonably well for the two thioethers, and even this method gives errors of 0.99 kcal·mol^{-1}. None of the selected methods produces charges which do well for three sets of molecules, ethers, pyridines, and amines, although for amines, none of the entire set of methods do well, while for both ethers and pyridines, some of the other methods give errors as low as 0.5 kcal·mol^{-1}. Overall, with averages taken either over each molecule or over each class of molecules, three of the methods are seen to do well with the protein set, HF/6-31G*, B3LYP/4-31G, and B3LYP/6-31G^{*+} outperform the other methods, with B3LYP/6-31G* charges being particularly disfavored. For the polyfunctional molecules, HF/3-21G charges do poorly for all classes, while again HF/6-31G*, B3LYP/4-21G, and B3LYP/6-31G^{*+} do relatively well across all sets. HF/6-31G^{*+} charges do not perform any worse than others on the aliphatic set, but do less well on the unsaturated and aromatic sets. The B3LYP/6-31G* charges, on the other hand, do poorly with both the aliphatic set and the aromatic set, but do reasonably well on the unsaturated molecules. These results strengthen the overall observations from the monofunctional groups that the HF/6-31G*, B3LYP/4-31G, and B3LYP/6-31G^{*+} charges give the best overall performance of these select methods, all giving average errors within 0.10 kcal·mol^{-1} of the overall best performing method.

Looking at the charges on several functional groups (see Figure 7-3), several observations can be readily made. First of all, the use of electrostatic potentials produced by Hartree–Fock quantum mechanics leads to higher magnitude partial atomic charges than when potentials from B3LYP quantum mechanics are used. While there are overlapping regions, where certain HF and B3LYP based methods yield similar charges, even when there is substantial overlap, the bias toward charges of higher magnitude for the HF method is clear. Secondly, in all cases considered there is some relationship between the value of the partial atomic charges and the quality of the reproduction of experimental free energies of hydration. For aldehydes, all methods which result in an oxygen charge of approximately $-0.5e$ give computed solvation free energies close to the experimental value, while all methods which result in an oxygen charge which deviates from $-0.5e$ give computed solvation free energies which agree with experiment more poorly. Similarly, for nitro compounds, computed solvation energies are close to experimental values for all methods resulting in oxygen charges of just below $-0.4e$, while the agreement with experiment becomes worse as the oxygen charge deviates from this value. A trend toward an alcohol hydrogen charge of about $+0.45e$ is also seen, although in this case the variation around the general trend is greater. In all three cases, there are similar trends for the more buried atoms of the functional group (aldehyde C, nitro N, and alcohol O), although in all these cases there are much larger deviations seen.

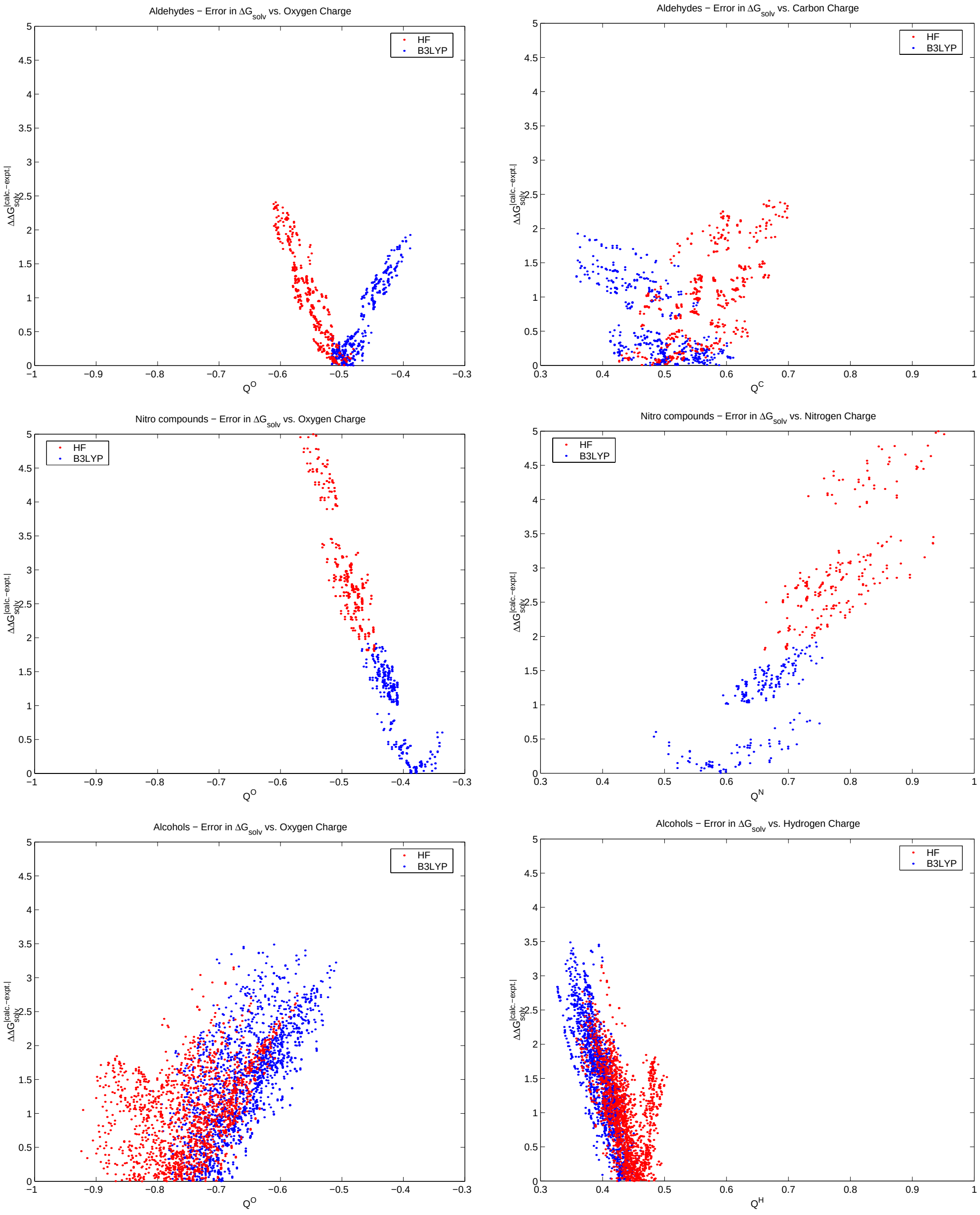


Figure 7-3: **Relation of computed hydration free energies and partial atomic charges on select functional groups.** The errors in computed free energies of hydration (in kcal·mol^{-1}) are plotted in relation to the partial atomic charges on the functional group for aldehydes, nitro-compounds and alcohols. **Red:** Methods using HF quantum mechanical results generally produced more highly polar functional groups. **Blue:** Methods based on B3LYP wavefunctions generally yield smaller magnitude charges. In all cases some relation between the charge and the error in computed energies is seen, although it is more pronounced in certain instances.

7.4 Discussion

Partial atomic charges are not a quantum mechanical observable. Thus, the computation of partial atomic charges is not entirely straightforward, and several different procedures have been developed. One class of methods involves partitioning the electron density between atoms, and combining the assigned electron density with the nuclear charge to give a partial charge for each atom. A second class of methods involves computing the electrostatic potential (ESP) for the molecular wavefunction (which is an observable), and then fitting a set of partial charges to best reproduce this potential. Both procedures can be done in numerous ways, with the most commonly used partitioning scheme being Mulliken population analysis [110], and several methods, including Merz–Singh–Kollman [7, 139], RESP [6, 28], Chelp [22], and ChelpG [10], regularly used for fitting charges to the electrostatic potential.

Since there is no single rigorous definition of partial atomic charge, the best choice of charges for any procedure which treats molecular electrostatics with a point charge model is somewhat ambiguous. In this case, the argument can be made that the best choice of charges is the set which best reproduces experimental results for the quantities of interest. With continuum electrostatics, one of the most important values to compute accurately is the free energy of solvation, since interactions in solvent can be reduced to differences in solvation free energy combined with Coulomb's Law in vacuum. The ability to reproduce solvation free energies with a continuum model has been used in the parameterization of the PARSE charge and radii set for proteins [140], and to evaluate the accuracy of various other parameter sets in continuum electrostatic calculations [39].

The free energy of solvation is not purely electrostatic in nature, and thus a continuum electrostatic model alone can not be expected to reasonably reproduce solvation free energies. In particular, the hydrophobic effect, related to the unfavorable free energy of solvation of non-polar molecules, requires a separate treatment. For a series of hydrocarbons, good agreement with experiment is attained using a linear relation to the solvent accessible surface area of the molecule [25, 140]. This is reasonable, considering that the larger the exposed surface area of a molecule, the larger the number of water molecules that will be involved in restructuring around the molecule, and this entropically unfavorable solvent reordering is a primary determinant of the positive solvation free energy [24, 26, 135]. In the model used here, this hydrophobic term is equated with the cost of forming a cavity of given surface area in the solvent, and is applied equally to all molecules, non-polar and polar. The exact form of the relation is obtained by fitting to the solvation free energies of a series of hydrocarbons, given a set of radii. While this could be done for every set of charges used, since the electrostatic contribution to the solvation free energy for hydrocarbons is small for all charge determination methods, the relation determined in the PARSE parameter development [140] (using completely hydrophobic hydrocarbons) was used in all cases. An alternative approach would be to fit the surface area term to the complete set of molecules for each charge determination method, which would improve the overall performance of all methods. This procedure, however, would add considerable complexity, and would make the comparison of different methods more difficult. In addition, it is unlikely that adding this variation would make a large difference in the results.

Initial calculations were run on a set of molecules based on amino acid functionalities using an extensive set of charge determination methods. This set contains a reasonable number of functionalities, as well as both positively and negatively charged molecules. The range of results is significant, with average errors as low as 0.86 kcal·mol^{-1} and as high as 5.47 kcal·mol^{-1}, indicating that choosing an appropriate method for determining charges is essential for accurate continuum electrostatic calculations.

Charges determined by Mulliken population analysis perform uniformly worse than those determined by fitting to the electrostatic potential. This is as may be expected since the solvation free energies are directly related to the electrostatic potential projected by the molecule into the solvent, and thus matching the potential well should lead to reasonable reproduction of solvation free energies. Of course, there remains the question of which method produces an electrostatic potential most compatible with the continuum model. Methods based on population analysis, however, make no attempts to accurately reproduce the electrostatic potential, and as a result, any accurate representation of quantities dependent on the potential would purely be a result of chance. The four semi-empirical methods, as well as the minimal STO-3G *ab initio* basis set, also perform uniformly more poorly than all other levels of theory, with any of ESP fit charges. Charges fit to the electrostatic potential by any procedure with potentials computed with any *ab initio* method above the STO-3G level do reasonably well, with average errors below 2.0 kcal·mol^{-1} for all but two methods. Thus it seems that the *ab initio* quantum mechanical electrostatic potential, as long as some minimum level of theoretical completeness is reached, is quite realistic, adequately reproducing the solvation energies.

The charges fit to the ESP using the Chelp procedure show a much greater dependence on geometry than do any of the other ESP fitting methods, with RMS variations in charge more than double that seen for the other procedures when different quantum methods are used for the geometry minimization step. This sensitivity to geometry was noted by Breneman and Wiberg [10] as being a result of the method by with the points at which to compute the ESP is determined. The ChelpG procedure, specifically designed to overcome this drawback, produces charges which vary much less with differences in geometry. For this method, and for the Merz–Singh–Kollman based procedures, RMS differences in charge are typically only $0.02e$ for any choice of geometry optimization procedure.

Due to the poor behavior of semi-empirical methods and the STO-3G *ab initio* basis set, as well as the Mulliken population analysis and Chelp ESP fitting procedures, these methods were not considered for the larger set of organic molecules. In general, the results from the larger set of molecules match those from the initial set, with the same methods identified as performing the best. One notable exception, however, is the charges fit from potentials computed at the HF/3-21G level of theory. While these methods gave good performance on the initial set of protein functionalities, they do much more poorly on the larger set. Looking at the results broken down by molecular class, it becomes readily apparent why this so. The charges from HF/3-21G potentials do particularly well on those functionalities over-represented in the protein set — amides, alcohols, and alkanes make up 38% of the protein set, but only 22% of the monofunctional compounds in the larger set — and do worst on more underpresented, or completely absent, functionalities — no nitro-compounds or esters are present in the protein set, and two carboxylic acids comprise only 8% of the protein set, but these functionalities make up 15% of the larger set of monofunctional compounds. For the other methods, the differences in performance between the molecule types are less biased toward those found in the protein set, and thus the performance of these methods between the two sets of molecules show similar trends. In most cases, however, the performance is worse for the larger set of molecules. This is not surprising, due to the necessity to balance the performance of each method over a much larger set of functionalities. In addition, several molecule types included only in the larger set, such as ethers and pyridines, show relatively poor performance in all top methods, and this contributes to the increased average error given by all these methods.

With the 6-31G and 6-311G basis sets at the HF level of theory, the best performance is obtained with one or two polarization functions, and the addition of diffuse functions reduces the agreement with experiment. Conversely, at the B3LYP level of theory, the same basis sets perform better with both polarization and diffuse functions than with polarization functions alone. Looking at the charges obtained by the different methods, it is seen that the B3LYP based methods generally give lower magnitude charges than the HF based methods. In a similar fashion, the addition of diffuse functions tends to give larger magnitude charges. As a result, some of the largest magnitude charges are found for HF methods including diffuse functions, while some of the lowest magnitude charges are obtained by using B3LYP methods without diffuse functions. In between are the charges obtained from HF methods with no diffuse functions and B3LYP methods with them. While for some systems (such as nitro-compounds, esters and carboxylic acids) the under-polarized charges give the best results, for other systems (such as ketones, alcohols and aromatics), the under-polarized charges do particularly poorly. Similar results are seen for the over-polarized charges, with very good performance seen for some molecules, and very bad performance seen for others. The more intermediate charges perform optimally for some systems, aldehydes being a particularly clear example, but rarely are seen to perform at the extreme end of poor reproduction of experimental values; when under-polarized charges are optimal, it is the over-polarized charges which do worst, and vice versa. Thus, when the results are taken as a whole, it is the methods which produce the intermediate polarity charge distributions which do the best.

For many levels of theory, and for all of the top performing levels, charges fitt to the electrostatic potential by the ChelpG procedure generally perform worse than those fitted by the Merz–Singh–Kollman scheme. The largest differences in the charges derived by these two methods is in the hydrocarbon charge distributions. For all non-aliphatic hydrocarbons, ChelpG yields smaller magnitude charges on the CH dipole, and this difference is greater than that seen by varying the level of theory at which the potentials are generated. In the case of aromatic residues, these smaller charges result in a much poorer performance for ChelpG relative to MK based schema, whereas a slight benefit is seen for ChelpG in the performance on alkenes, and a larger benefit is seen in the performance on alkynes. The poor performance on aromatic residues, coupled with the relatively large number of aromatic molecules in the dataset, leads to a slightly poorer performance by ChelpG overall. For alkanes, there are significant differences in charges between the ChelpG and MK methods, with ChelpG again producing smaller charges (although not as small as are obtained by restrained fits),

	MK ESP	RESP 1X	RESP 2X	RESP Polar H	ChelpG
Alkanes					
HF/6-31G*	0.60	0.54	0.54	0.45	0.56
HF/6-31G*+	0.60	0.54	0.54	0.45	0.57
B3LYP/6-31G*	0.56	0.51	0.50	0.44	0.52
B3LYP/6-31G*+	0.58	0.54	0.54	0.45	0.56
Aromatics					
HF/6-31G*	0.50	0.63	0.64	0.62	1.80
HF/6-31G*+	0.54	0.56	0.57	0.56	1.89
B3LYP/6-31G*	1.27	1.53	1.54	1.53	2.57
B3LYP/6-31G*+	0.86	1.15	1.16	1.14	2.67
Alkenes					
HF/6-31G*	1.00	0.88	0.88	0.89	0.67
HF/6-31G*+	1.26	1.14	1.14	1.16	0.72
B3LYP/6-31G*	0.76	0.70	0.70	0.72	0.56
B3LYP/6-31G*+	1.01	0.89	0.89	0.91	0.58
Alkynes					
HF/6-31G*	1.80	1.67	1.67	1.67	1.05
HF/6-31G*+	2.17	2.03	2.03	2.03	1.01
B3LYP/6-31G*	1.28	1.16	1.15	1.15	0.61
B3LYP/6-31G*+	1.66	1.53	1.53	1.54	0.54

Table 7-5: **Performance of select charge determination methods on hydrocarbons.** Even on pure hydrocarbons, there is significant variation in the performance of different methods. ChelpG ESP fitting does uniformly worse on aromatic molecules, but uniformly better on alkenes and alkynes, than do the MK based methods. Average absolute errors are given in kcal·mol^{-1}.

but these changes have little effect on the energetics of solvation, as can be seen by the similar performance of the restrained and unrestrained MK based methods.

That such large variations in charge are seen between different electrostatic potential fitting procedures suggests that the weak electrostatic potential produced by hydrocarbons, even in the slightly polar unsaturated systems, poorly defines a point charge distribution. In the polar functionalities on the other hand, the potential is much stronger, and thus clearly defines the fit partial atomic charges. For these groups, the ChelpG and MK fitting procedure give nearly identical charges, with much smaller variations seen between the different fitting methods than is seen between different levels of theory for the calculation of the electrostatic potential.

In general, the charges determined by restrained fitting perform similarly to those obtained by an unrestrained fit to the same potential. In some cases slightly better results are seen with unrestrained charges, while in other cases the reverse is true. Again this tends simply to highlight the point that a relatively broad range of hydrocarbon charges give equivalent electrostatic potential fields — the small magnitude of the potential requires large changes in charge to make significant changes in the energetics. However, the polar hydrogen model charges, with aliphatic hydrogens fixed to have no charge, in general do more poorly, except on pure hydrocarbons. When polar atoms are present, the higher electronegativity of the heteroatoms can lead to substantial charges on aliphatic groups. In a united-atom model for aliphatic groups, this forces the entirety of the charge onto the carbon, whereas in an all-atom model the charge can be distributed across hydrogens as well. This leads to a much better fit for the all-atom models, as a single point charge can not adequately describe a polarized aliphatic group. For pure hydrocarbons, however, no aliphatic group is particularly polarized, and thus the united-atom model performs well.

It should be noted that all the electrostatic potential fitting procedures have several parameters which may be varied. For all methods, the density and expansiveness of the ESP grid can be changed, and the RESP method could be applied to the Chelp and ChelpG grids, or vice versa, the Chelp/ChelpG method applied to the Merz–Singh–Kollman grid. In addition the Chelp/ChelpG methods have a variable parameter of the SVD cutoff value, and the restraint forces in the RESP method can similarly be varied. It is entirely possible that variation of these parameters could improve the performance of the charges obtained through these procedures.

Another consideration is that the choice of radii and charges for continuum electrostatic calculations are not independent. Similar solvation energies may be computed if the radii of the atoms in a molecule increase along with the polarity. All the calculations here used a fixed set of radii — identical to those found in the PARSE parameter set for C, N, O, S, and H — based on Pauling van der Waals radii. However a different choice of radii may favor the performance of different methods. Larger radii would likely make methods which result in more polarized charge give better agreement to experiment, as would smaller radii for methods which yield less polarized charges. Pauling radii, however, give a very simple set of radii, based only on atom type and hybridization, and thus are easily extendable to any molecule. In addition, a simple set of radii, based not on parameterization but on detailed computation or experimental observation (the Pauling radii are derived from crystal packing data) [116], is more consistent with the approach we have taken to evaluate the performance of existing charge determination methods, as opposed to fitting a new set of parameters.

7.5 Conclusions

A detailed analysis of the performance of charges determined by a large number of *ab initio* methods in continuum solvation calculations was performed using two sets of molecules, one based on amino acid side chains, and a second representative of a broad range of organic functionalities. The results clearly demonstrate that particular basis sets and levels of quantum mechanical theory yield charges which give much closer agreement to experiment than others. Rather than larger basis sets and higher levels of theory giving better results, the best results for the data set based on protein groups are obtained with the modestly sized 6-31G* basis set at the Hartree–Fock level of theory. In addition, on the same data set, the charges determined at the HF/3-21G and B3LYP/4-21G levels of theory perform surprisingly well, surpassing the performance of many higher levels of theory. With a more extensive set of molecules, these theoretical methods, with the exception of HF/3-21G, continue to produce the charges which most accurately reproduce experimental values. Although the best method for the larger set is based on potentials from a relatively costly B3LYP/6-311G^{*+} quantum mechanical calculation, the charges from certain lower levels of theory perform with almost identical accuracy.

Semi-empirical methods, and the minimal *ab initio* basis set STO-3G, were both found to produce charges which did a very poor job of reproducing experimental solvation free energies, as was the case for charges at any theoretical level from Mulliken population analysis. Interestingly, the Merz–Singh–Kollman electrostatic potential fitting method, and the associated RESP restrained ESP fitting procedure in general produce charges which are more suited for the solvation calculations done here than are the charges produced by the Chelp or ChelpG ESP fitting methods, although this may be a result of uneven sampling of certain chemical functionalities.

No method does well on all types of molecules, with some methods producing partial charges too large in magnitude, and others producing charges which are too small. The top performing methods attain a good balance for many molecules, but still err on both sides for some functionalities. For molecules containing amines, none of the methods produces charges which can adequately reproduce experiment, due primarily to the inadequacies of a four-point charge model to describe the electrostatic field produced by an amine. An extended model, with a “dummy” atom in the position of the nitrogen lone-pair may alleviate some of these problems. In general, the same charges on functional group atoms give the best results in all molecules. This suggests that a functional group-based parameter set may give the best results for a large number of molecules. While parameter-set based methods always have the drawback of being not directly extensible to new functionalities, if the development was done using a clear protocol of combining charges derived from quantum mechanical electrostatic potentials and any available experimental data, this may be applicable to a large subset of the molecules of interest. The development of such a parameter set is beyond the scope of this work, but could significantly enhance the accuracy of continuum solvation calculations on small molecules.

Chapter 8

General Conclusions

The electrostatic contributions to binding in several protein–ligand complexes were analyzed, both in terms of the behavior of the wild-type system, and in the context of designing more favorable electrostatic interactions. The results show that electrostatic interactions have been optimized to some degree in certain natural systems, but also that electrostatic optimization can be used in the design of complexes with improved affinity. Each design took a slightly different approach, demonstrating the versatility of the electrostatic optimization procedure.

In consideration of the glutaminyl-tRNA synthetase from *Eschericia coli*, the charge distributions of the natural ligands were found to be remarkably close to optimal, despite significant differences in binding energy in many cases. This agreement between the natural and optimal charges suggests that the optimization of electrostatic interactions played an important role in the evolution of this system. As a high degree of specificity in aminoacyl-tRNA synthetases is essential for faithful translation of the genetic code, and as electrostatics are recognized as being key determinants of specificity, it is not surprising that many electrostatic interactions are made, although the degree of optimization could not have been predicted *a priori*.

Two systems were studied with the aim of designing improved inhibitors of HIV-1 viral–cell membrane fusion. For the case of a small D-peptide which binds to the

HIV-1 gp41 N-terminal coiled coil, two tryptophans at the interface were identified as the best targets for modification. A novel computational screening algorithm, based on the theory of electrostatic optimization, was developed in order to evaluate the effects on binding of a large number of tryptophan derivatives at these two sites. In another case, the binding of a protein construct to an isolated helical peptide from the C-terminal region of HIV-1 gp41 was analyzed. The electrostatic optimization procedure pinpointed several locations where mutations might improve the binding free energy. Four mutations at three positions were modeled, and the computed binding free energies were more favorable than the original system in every case, with one triple mutant showing a five hundred-fold improvement in calculated K_d.

In the complex of β-lactamase inhibitor protein with TEM1 β-lactamase, a novel type of electrostatic interaction was identified, involving through-solvent interactions of charged residues on the periphery of the binding interface. Both favorable and unfavorable interactions of this type were identified in the natural system at distances as high as 10.0 Å from the interface. In addition, a set of ten residues with the potential for making favorable interactions of this type were identified through electrostatic optimization. The set included positively charged groups which make favorable interactions in the wild-type complex and negatively charged groups which contribute unfavorably to the stability of the natural complex, as well as several uncharged residues which made little energetic contribution to binding in the natural system, but which could contribute significantly if mutated to a charged group.

Finally, numerous methods for the calculation of partial atomic charges on small organic molecules were evaluated in terms of their performance in continuum solvation calculations. As the choice of charges for small molecule ligands of proteins is often unclear, this has an important role to play in the analysis of the binding energetics of small molecules. Semi-empirical based methods gave very poor results, and a significant range of performance was seen for different methods based on *ab initio* quantum mechanics. The commonly used method of fitting partial atomic charges to the electrostatic potential computed from the HF/6-31G* wavefunction ranks among the top methods. The top performing methods tend to give similar charges on polar functional groups, suggesting that an appropriately derived parameter set based on functional groups may be useful.

The results described herein have demonstrated the utility of electrostatic optimization as a tool both in the analysis and design of protein complexes, opening up many possibilities for further application of the procedure. In natural systems, applications to additional enzyme systems could provide insight into the generality of electrostatic optimization during the evolution of natural systems. Some preliminary work has also been done on considering the role of optimized electrostatic interactions in enzyme catalysis, and further work in that direction is ongoing. The usefulness of electrostatic optimization in the design of high affinity complexes can not be missed, and further applications in this field are a natural extension of much of the work described here.

Appendix A

Residual Potentials and Affinity: Analysis of "Action-at-a-Distance" Mutants of β-Lactamase Inhibitor Protein[1]

Abstract

The effect of a series β-lactamase inhibitor protein (BLIP) mutants on the binding to TEM1 β-lactamase was considered by analysis of the electrostatic complementarity in each system. The correlation coefficient of the BLIP desolvation potential and the TEM1 interaction potential on the surface of BLIP is found to be strongly correlated to the experimental binding free energies. In many cases this increased correlation can be seen visually as a reduced residual potential — defined as the sum of these two potentials — which has been shown to be a good measure of electrostatic complementarity. An additional mutation of Asp133 to lysinex is proposed, which calculations suggest would enhance binding affinity both alone and in concert with previously identified mutations. The mutations are seen to act somewhat locally, in that they act on patches of the surface relatively close to the site of mutation, reducing the residual potential. However, the interactions are not specific, with two of the three most effective mutations located more than 7.5 Å from TEM1. In addition, even for residues in relative proximity to the interface, the effect of the mutation is

---

[1] This work was done as a collaborative effort with Brian A. Joughin

computed to have a similar effect even in very different conformations. The results suggest that the mutations act in an intriguing manner, with relatively long-range electrostatic effects projected through a region of solvent to improve the electrostatic complementarity of the ligand to the receptor.

A.1 Introduction

The field of protein design has made substantial advances over the last twenty years, based largely on phrasing the appropriate inverse problem and developing methods capable of addressing inverse design [40, 114]. Much current protein design work involves the construction of stabilizing protein side-chain arrangements by methods such as dead-end elimination [33, 37, 58, 88, 90, 96], self-consistent mean-field theory [84–86], simulated annealing [61, 91], and genetic algorithms [35, 72]. That is, successful design is achieved by consideration of detailed atomic interactions and their effect on packing geometry and energetics. The design of protein binding interfaces can be achieved by a similar overall approach, although the additional requirement to treat solvation and electrostatic interactions adds a further layer of complexity [17, 93].

An alternative strategy that does not demand the same detailed packing together of side chains into an exquisite three-dimensional jigsaw puzzle may be desirable in many cases. One such method involves the enhancement of affinity through relatively long-range electrostatic effects by the mutation of surface residues located somewhat outside of the protein–protein binding interface. When mutations are not located directly at the binding interface, a detailed consideration of the packing of residues may be unnecessary. Moreover, because the effects of such mutations act over a relatively long range, such a strategy should be more tolerant of local imperfections in structural models. Less apparent, however, is how effective these types of mutations can be (since much of the interaction may be screened by solvent), whether the sites where such mutations will be most effective are localized on the structure, and, if so, how these locations might be determined. An important consideration in this type of design is the counterplay of favorable direct electrostatic interactions and unfavorable desolvation effects, which has been shown to be incredibly important in understanding the energetics of electrostatic interactions in biological systems [68]. The lessons learned from detailed analyses of short-range electrostatic interactions such as salt-bridges and hydrogen bond networks may or may not prove to be extendable in a straightforward manner to electrostatic interactions acting over a longer range.

We have begun to address these issues by analyzing a set of previously identified single and multiple mutants of the β-lactamase inhibitor protein (BLIP) which affect the affinity of this protein for TEM1 β-lactamase (TEM1) [132]. Using methods based on continuum electrostatics, we were able to consider in detail the electrostatic contribution to the energetics of binding for the wild-type and mutant structures. The degree of electrostatic complementarity is seen to correlate well with the experimentally determined binding affinities, suggesting that these tools may be particularly useful both in understanding and in designing these types of mutations.

Our laboratory has previously described a measure of electrostatic complementarity between two binding partners, which we denote the residual potential [77, 94]. The consideration of electrostatics in binding involves balancing favorable interactions between the members of the complex in the bound state with the loss of favorable interactions each component makes with solvent on the transition from the unbound to the bound state. It can be shown rigorously that in a perfectly complementary complex, this balance is met by having the interaction potential of the receptor opposite in sign and equal in magnitude the desolvation potential of the ligand everywhere within and on the ligand surface [77]. Thus the residual potential is defined as:

$$\phi^{resid} = \phi_{rec}^{inter} + \phi_{lig}^{desolv} \tag{A.1}$$

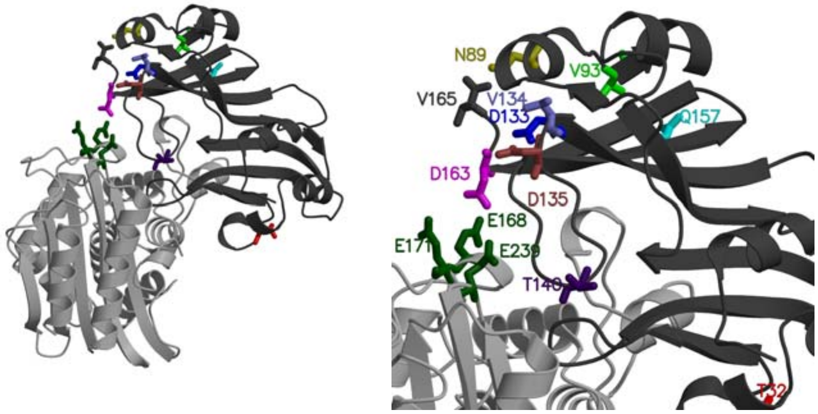
The residual potential is near zero in regions of high complementarity and is larger in magnitude in regions which are uncomplementary. It is important to note that the definition of the residual potential is fundamentally asymmetric, describing the complementarity of one component, defined as the ligand, for binding to the other component, defined as the receptor. A complex in which one ligand which is perfectly complementary to its receptor is generally not perfectly complementary when the roles of the components are reversed; the receptor is not perfectly complementary to the ligand. A numerical measure of the complementarity of a ligand for its receptor can be attained from the correlation of ϕ_{rec}^{inter} and ϕ_{lig}^{desolv} over all points of interest, typically the ligand surface:

$$r = \frac{\sum[(\phi_{rec}^{inter} - <\phi_{rec}^{inter}>)(\phi_{lig}^{desolv} - <\phi_{lig}^{desolv}>)]}{[\sum(\phi_{rec}^{inter} - <\phi_{rec}^{inter}>)^2 \sum(\phi_{lig}^{desolv} - <\phi_{lig}^{desolv}>)^2]^{1/2}} \tag{A.2}$$

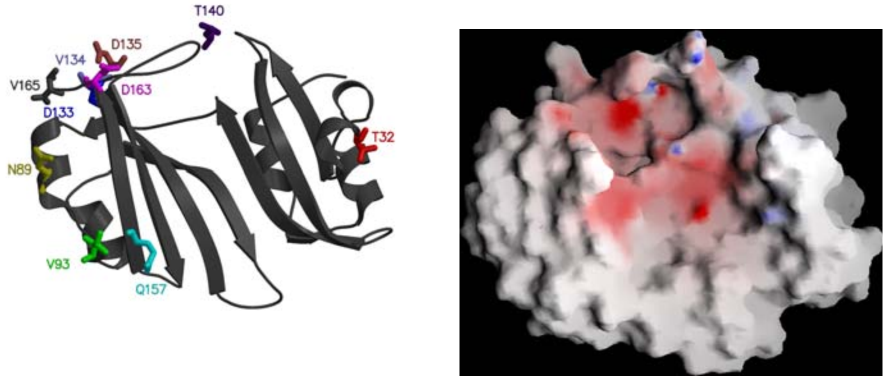
In a perfectly complementary system the correlation coefficient will be -1.0. Negative values smaller in magnitude indicate imperfect complementarity, while positive values indicate anti-complementarity, with the sign of the desolvation potential of the ligand matching that of the interaction potential of the receptor in an overall sense.

A.2 Results and Discussion

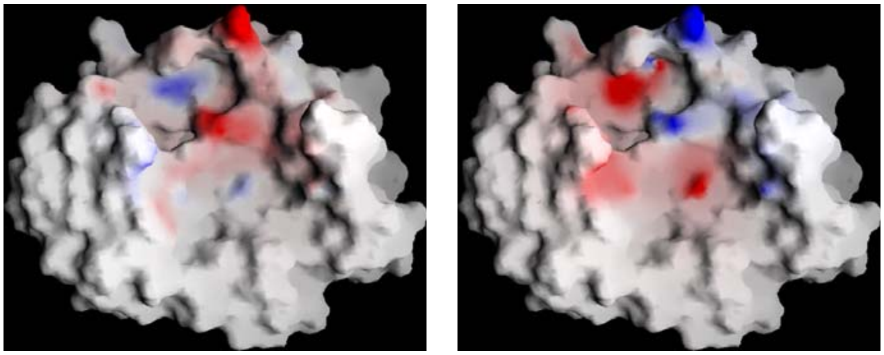
Wild-type BLIP binds to TEM1 with a K_d of 1.25 nM [132], burying 2978 Å^2 of solvent exposed surface, and forming sixteen hydrogen bonds and four salt-bridges across the binding interface, making it a fairly typical enzyme–inhibitor complex [27]. Desolvation, interaction and residual potentials on the surface of BLIP were computed and are displayed in Figure A-1 along with an overview of the structure. The desolvation potential of BLIP is quite complementary to the interaction potential of TEM1 projected onto the BLIP surface; most regions of positive desolvation potential are well matched by regions of negative interaction potential, and vice versa. However, examination of the residual potential makes it clear that BLIP is not perfectly complementary to TEM1. Specifically, there is a negative residual potential over a large area of the binding surface resulting from an excess negative interaction poten-



A: BLIP–TEM1 complex **B:** Notable residues



C: BLIP binding interface **D:** Desolvation potential



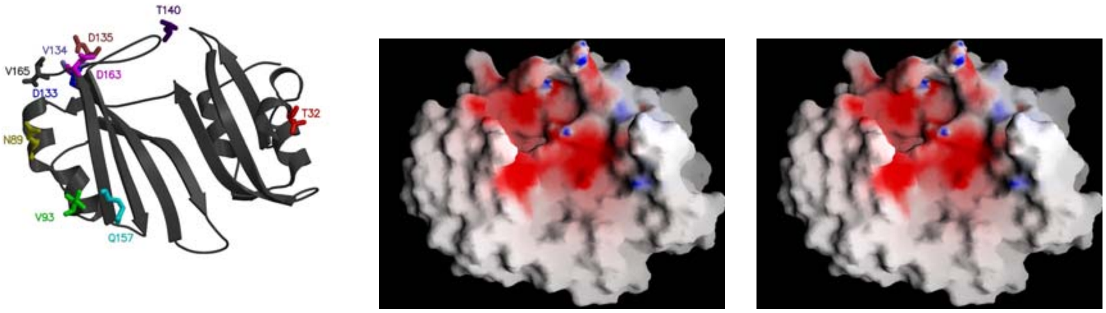
E: Interaction potential **F:** Residual potential

Figure A-1: **The BLIP–TEM1 complex.** **A–B:** The structure of the complex between BLIP and TEM1 is shown with all mutated side chains included. **C:** A view of the BLIP binding interface in same orientation as panels D-F. **D–F:** BLIP desolvation potential (D), TEM1 interaction potential (E) and residual potential (F) displayed on the surface of BLIP. Panels A-C were made with MOLSCRIPT [87] and RASTER3D [105]. Panels D-F were made with GRASP [111].

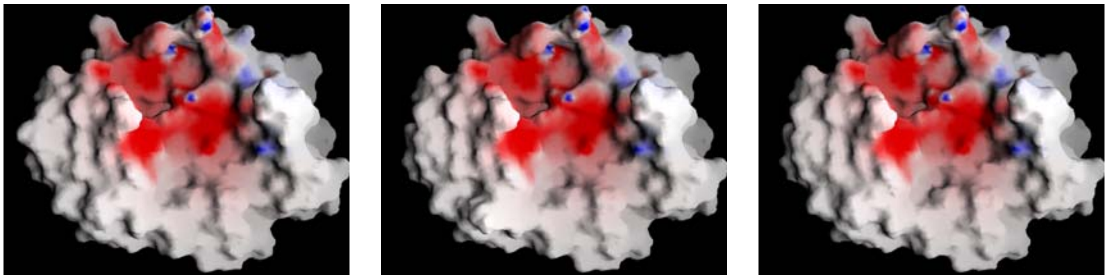
tial from TEM1. To increase the complementarity would require either a reduction in the negative interaction potential of TEM1 or an increase in the positive desolvation potential of BLIP. This suggests that the binding affinity of BLIP for TEM1 would likely be improved by mutations which increase the positive charge on the inhibitor, that is, mutations of acidic residues to neutral or basic residues and mutations of neutral residues to basic residues.

The observation that increased positive charge on BLIP should promote binding was noted by Schreiber and co-workers, and was used as a guide in the design of a set of BLIP mutants [132]. All mutated residues were located on the periphery of the binding interface (Figure A-1), satisfying the "action-at-a-distance" design specification. Several of these mutants had little effect on binding affinity, while others had significant effects, suggesting, as noted by the original authors, that there are specific regions where the mutations are most effective, as opposed to all the mutations acting through a gross delocalized electrostatic effect dependent only on the total charge of the ligand. This is easily understood in the context of the residual potential; there are clearly regions in which the residual potential is much larger, and thus much less complementary, than others.

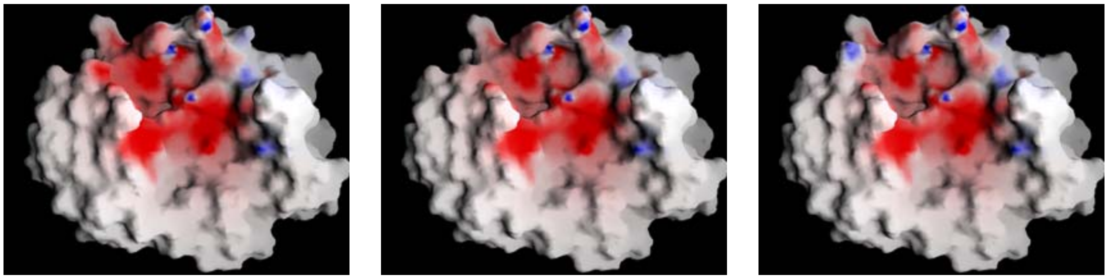
A model of each mutant structure was built, and the structures and electrostatic potentials were analyzed. Residual potentials of the mutants are displayed in Figure A-2, and the correlation coefficients of the desolvation potential of BLIP with the interaction potential of TEM1 (evaluated over the surface of BLIP) for each mutant are detailed in Table A-1, along with the computed electrostatic component of the binding free energy. The correlation of the experimental binding free energies to the correlation coefficient is quite strong (Figure A-3), suggesting that electrostatic effects, and in particular the improvement of the overall electrostatic complementarity of BLIP for TEM1, is a primary means by which these mutants act. Thus the residual potential and its quantitative analysis shows significant promise as a tool in understanding, and potentially designing, these types of surface mutations which act,



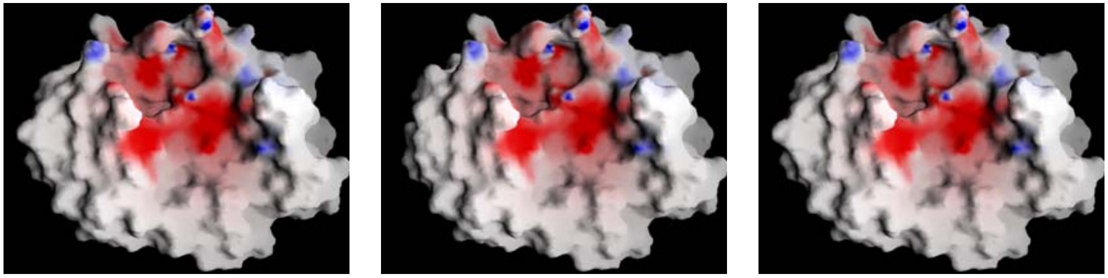
A: Structural overview **B:** Wild type **C:** 32→K



D: 89→K **E:** 93→K **F:** 140→K



G: (140,157)→K **H:** 163→A **I:** 163→K



J: (89,163,165)→K **K:** (134,135,163)→K **L:** (89,135,163,165)→K

Figure A-2: **Residual potentials on the surface of BLIP for a series of BLIP mutants.** The residual potential for each mutant is shown on the surface of BLIP in the orientation displayed in panel A. Panel A was made with MOLSCRIPT [87] and RASTER3D [105]. Panels B–L were made with GRASP [111].

BLIP Mutations	ΔQ^{tot}	$\Delta\Delta$SASA	$\Delta\Delta G^{es}$	Corr.	$\frac{K_d^{WT}}{K_d^{Mut}}$ [a]
Wild type	0	0	0.0	−0.60	1.0
T32K	+1	0	−0.1	−0.60	0.7
N89K	+1	0	−0.2	−0.60	2.2
V93K	+1	0	−0.3	−0.60	2.2
T140K	+1	24	0.1	−0.60	1.0
T140K, Q157K	+2	24	−0.1	−0.60	2.0
D163A	+1	-47	−5.1	−0.65	9.6
D163K	+2	95	−4.2	−0.69	28.4
V165K, D163K, N89K	+4	94	−4.2	−0.68	56.8
D163K, D135K, V134K	+5	95	−4.7	−0.69	173
D163K, V165K, D135K, N89K	+6	94	−4.9	−0.70	291
D133K	+2	0	−2.0	−0.64	N/A
D133K, D163K	+4	95	−5.3	−0.71	N/A
D133K, D135K, D163K	+6	95	−5.8	−0.72	N/A

[a] [132]

Table A-1: **Energetic details of mutations to BLIP.** For all BLIP mutants studied, the change in total charge, the difference in solvent-accessible surface area buried on binding, and the difference in computed electrostatic binding free energy (in kcal·mol^{-1}), both relative to wild type, are tabulated. Also shown are the correlation of the BLIP desolvation and TEM1 interaction potentials on the surface of BLIP, and the ratio of the wild type and mutant dissociation constants as determined experimentally by Selzer *et al.* [132].

via through-solvent interactions, to promote binding.

Four single mutants and one double mutant were found to have no change in the correlation between the desolvation and interaction potentials, and visually had no change in the residual potential. In addition, these mutations had calculated changes in electrostatic binding free energy of less than 0.5 kcal·mol^{-1} in magnitude. These mutations all have differences in K_d relative to wild type of less than three-fold, and include all the low-activity mutants.

Only two of the single mutants result in significant changes to the correlation of the interaction and desolvation potentials, and have visible differences in the residual potential, compared to wild type. Both these mutants involve Asp163, with the

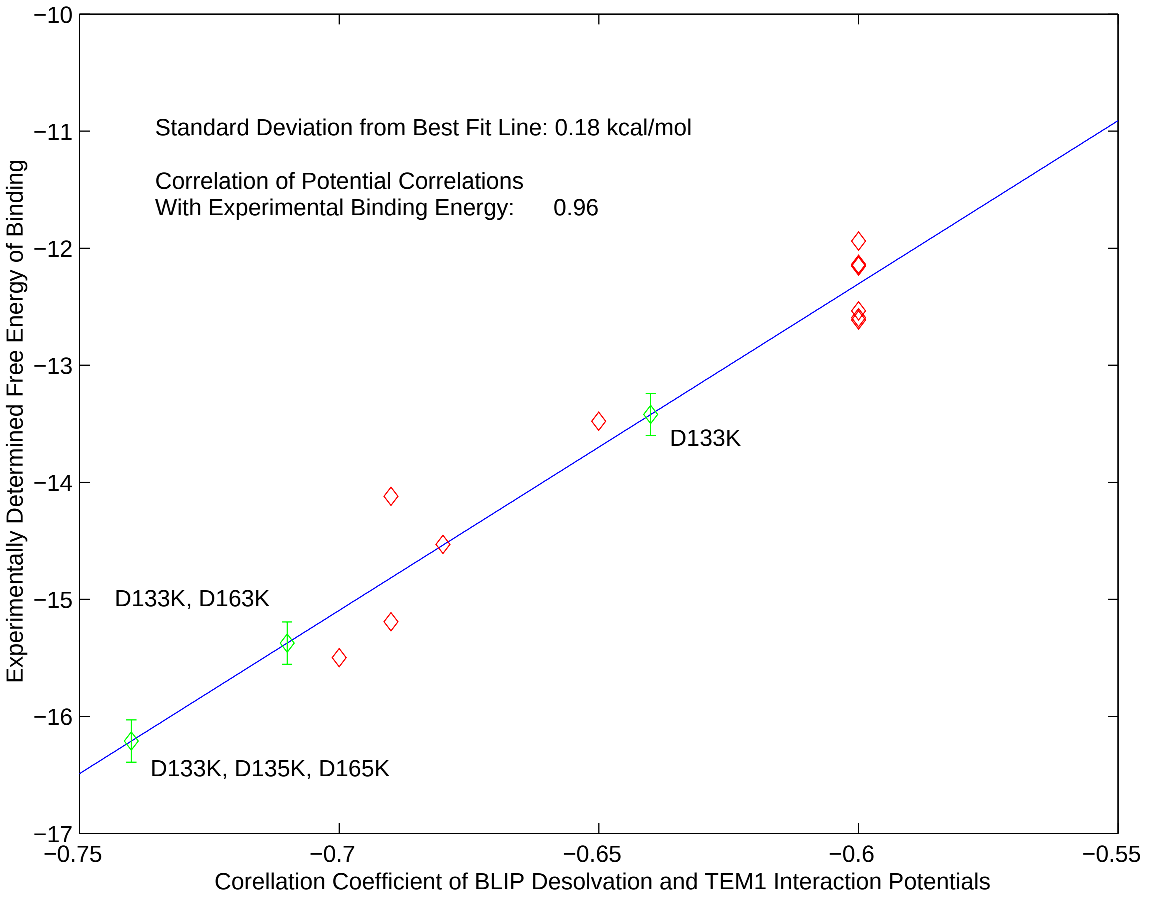
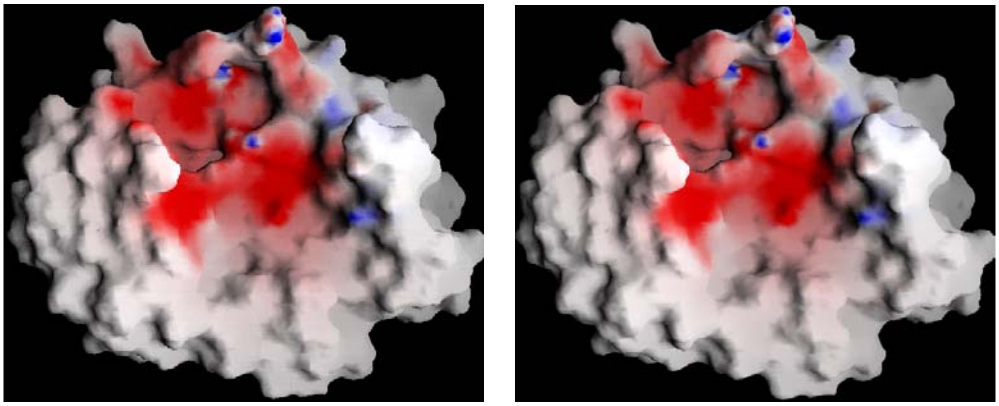


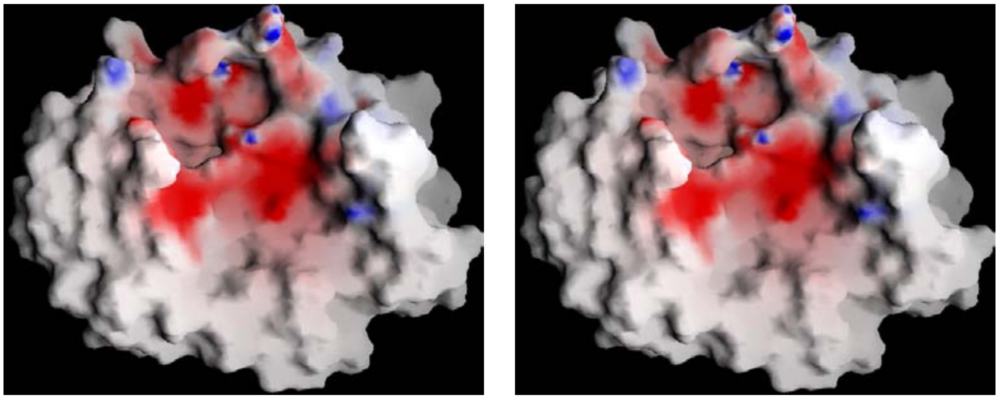
Figure A-3: **Correlation of experimental binding free energies with quantitative analysis of the residual potential.** The variation of experimental binding free energies (in kcal·mol^{-1}) with the correlation coefficient between the BLIP desolvation potential and the TEM1 interaction potential, calculated on the surface of BLIP, is shown. In red are the experimentally verified mutations and in green are predictions including the D133K mutation.

mutation to alanine showing qualitatively similar, but smaller in magnitude, effects as the mutation to lysine. The excess negative residual potential is visibly reduced in both mutants, and the correlation between the interaction and desolvation potentials becomes more negative. The computed change in electrostatic binding free energy relative to wild type is similar for both mutants (-5.1 kcal·mol^{-1} for D163A and -4.2 kcal·mol^{-1} for D163K). Experimentally, the D163A mutant binds ten-fold better than wild type, while the D163K mutant binds 28-fold better. The minimum energy structure of lysine at position 163 brings it into fairly close contact (although not hydrogen-bonding or salt-bridging distance) to several acidic residues on TEM1. To determine whether this geometry is required for the computed effect, a model was constructed with the lysine in an extended conformation, not making these interactions. The same observations are made for this structure (data not shown) and thus the details of the structure do not seem to be overly important. The three multiple mutants which contain the D163K mutation have similar differences with respect to wild type. Two of these contain the additional mutation of Asp135 to lysine, and have calculated electrostatic binding free energies relative to wild type of -4.7 kcal·mol^{-1} and -4.9 kcal·mol^{-1}, as compared to -4.2 kcal·mol^{-1} for both the single D163K mutant and the multiple mutant containing D163K but not D135K. The correlation of the interaction and desolvation potentials is only slightly improved for these mutants, but a slight decrease in the excess negative residual potential can be seen relative to the single D163K mutant. These D163K and D135K containing mutants have experimental improvements in binding affinity of 170-fold and 290-fold over wild type, whereas the multiple mutant containing only D163K has a binding affinity of only 56-fold better than wild type, and only two-fold better than the single D163K mutant.

Our ability to correlate experimental changes in binding affinity with the degree of electrostatic complementarity in this system suggests that a similar approach may be useful in design. In particular, mutations to surface residues which improve the electrostatic complementarity of the complex are likely to lead to tighter binding as well. A detailed analysis of the contribution of each residue of BLIP to the electrostatic binding free energy suggested that Asp133 is a likely candidate (see Chapter 6). We generated models of three mutants: a single mutant of D133K alone, a double mutant of D133K and D163K, and a triple mutant of D133K, D135K and D163K. The multiple mutants contain the suggested D133K mutation along with the best



A: Wild type **B:** 133→K



C: (133,163)→K **D:** (133,135,163)→K

Figure A-4: **Residual potentials of newly designed BLIP mutants.** The residual potentials are displayed on the surface of BLIP for a series of newly designed BLIP mutants. All figures are in the same orientation as in Figure A-2. These figures were made with GRASP [111].

one or two of the experimentally verified mutations. The single mutation visually decreases the excess residual potential as compared to wild type, and the correlation of the desolvation and interaction potentials is significantly improved. The calculated improvement in electrostatic binding free energy of the single mutant relative to wild type is 2.0 kcal·mol^{-1}. The two multiple mutants have better correlation between the desolvation and interaction potentials than any of the initial mutants studied, and the residual potential is slightly visually improved over the best of the initial mutants as well. The calculated electrostatic binding free energies of the two multiple mutants were 5.3 kcal·mol^{-1} and 5.8 kcal·mol^{-1} improved over wild type, more favorable than any of the previous mutants.

A.3 Conclusion

We have examined the electrostatic complementarity of a series of mutants of the β-lactamase inhibitor protein (BLIP) and analyzed the results with comparison to experimental binding free energies to TEM1 β-lactamase. We find that the correlation coefficient of the BLIP desolvation potential and the TEM1 interaction potential on the surface of BLIP is strongly correlated to the experimental binding free energies. In addition, this increased correlation can be seen visually as a reduced residual potential in many cases. An additional mutation of Asp133 to lysine is proposed, which calculations suggest would enhance binding affinity both alone and in concert with previously identified mutations. The effects of these mutations are localized to the extent that they act on patches of the surface, somewhat locally improving the residual potential. However, the interactions are not specific; two of the three most effective mutations (D133K and D135K) are more than 7.5 Å from TEM1, and the D163K mutation has similar computed effects even in two very different conformations. This helps to confirm the overall mechanism by which these mutations act — relatively long-range electrostatic interactions act through a region of solvent to improve the overall electrostatic complementarity of the ligand for binding to its target receptor. Further work investigating the design of surface mutations which permute the residual potentials toward increased complementarity is ongoing.

A.4 Methods

All calculations were done using the X-ray crystal structure of the BLIP–TEM1 complex solved by James and co-workers as an initial model (Protein Data Bank [125]

ID 1jtg) [144]. Hydrogen atoms were added using the HBUILD facility [14] within the CHARMM computer program [11] using the PARAM22 all-atom parameter set [100]. Visual analysis of structure suggested no reason for the ionizable residues to be in their non-standard protonation states, and thus all histidines were left in their neutral state, and all acidic residues were left charged. This results in a net charge of $-2e$ for BLIP and $-7e$ for TEM1. Binding was considered in the rigid-body docking approximation.

Continuum electrostatic calculations were performed by numerical solution of the linearized Poisson–Boltzmann equation, using a locally modified version of the computer program DELPHI [55, 57, 134, 136]. A grid of dimension $129 \times 129 \times 129$ was used with focusing boundary conditions, in which the largest dimension of the molecule occupies first 23% then 92% of one edge of the grid, resulting in a final grid spacing of 0.59 Å. An internal dielectric constant of 4 and an external dielectric constant of 80 was used, along with an ionic strength of 0.145 M and a 2.0 Å ion exclusion layer [9]. The molecular surface (used to define the dielectric boundary) was generated with a probe radius of 1.4 Å. Protein partial atomic charges and radii were taken from the PARSE parameter set [140] with a few minor changes. Charges on the bridging ring carbons of tryptophan were assigned to $0e$, charges for proline and for disulfide bridged cysteine residues were taken from the PARAM19 parameter set [11], and the charges from glutamate and lysine side chains were used for charged C and N termini respectively. Surface potentials were displayed with the program GRASP [111], and were numerically analyzed with locally written software.

Model structures were generated by holding the positions of all backbone atoms and all those of non-mutated side chains constant, while allowing mutated residues to minimize from a sampling of cardinal torsions using the adapted-basis Newton–Rhapson (ABNR) algorithm in the computer program CHARMM [11] with the PARAM22 all-atom parameter set [100]. When multiple mutations were made to the same molecule, the positions of mutated residues located somewhat distantly from each other were minimized singly and combined, while the positions of mutated residues located within two residues of each other in sequence were minimized simultaneously.

Appendix B

ICE User's Manual[1]

B.1 Introduction

ICE (INTEGRATED CONTINUUM ELECTROSTATICS) is a suite of software for the analysis and optimization of electrostatics in biomolecular systems. ICE is particularly designed for the analysis of the association, and to a lesser degree the stability, of biological macromolecules. Included are an interface to the DELPHI Poisson–Boltzmann solver, software for performing component analyses, and software for performing electrostatic optimizations.

B.1.1 Molecular binding free energy

The free energy of binding of two molecules can be broken up into several components. There are direct interactions between the molecules in the bound state, including van der Waals and electrostatic interactions, as well as components resulting from the different interactions the molecules make with solvent in the bound and unbound states. The solvation terms include the loss of van der Waals interactions with solvent upon binding, the reduction of favorable induced electrostatic interactions between solvent and both molecules on binding, and the entropy dominated hydrophobic effect

---

[1] Portions of ICE are derived from software written by Erik Kangas and Zachary Hendsch.

resulting from differences in water structure at the boundaries of each free molecule and of the complex. Other contributing terms include entropic costs from the loss of translational and rotational degrees of freedom on binding, and from the change in the conformational mobility of the molecules on binding. Thus, a general breakdown of the free energy of binding is as follows:

$$\Delta \mathrm{G}^{bind} = \Delta \mathrm{G}^{inter.}_{VDW} + \Delta \mathrm{G}^{inter.}_{ES} + \Delta \mathrm{G}^{solv.}_{ES} + \Delta \mathrm{G}^{solv.}_{h\phi} + \Delta \mathrm{G}^{conf.} + \Delta \mathrm{G}^{ent.} \tag{B.1}$$

with two direct interaction terms (van der Waals and electrostatic), two solvation terms (electrostatic and hydrophobic), a term for differences in relative conformational energies, and a final term dealing with entropic terms not encompassed by other terms, such as the loss of conformational entropy on binding and the change of three translational and three rotational degrees of freedom into vibrational modes.

B.1.2 Electrostatic binding free energies

The electrostatic binding free energy can be written by eliminating all non-electrostatic terms from the general expression:

$$\Delta \mathrm{G}^{bind}_{ES} = \Delta \mathrm{G}^{inter.}_{ES} + \Delta \mathrm{G}^{solv.}_{ES} (+\Delta \mathrm{G}^{conf.}_{ES}) \tag{B.2}$$

where the last term — the electrostatic portion of the energetics of conformational change — is present only when the rigid body binding approximation is not used. Ignoring this last term, this can reformulated in the following way:

$$\Delta \mathrm{G}^{bind}_{ES} = \Delta \mathrm{G}^{Coulomb.}_{A-B} + \Delta \mathrm{G}^{Screening}_{A-B} + \Delta \mathrm{G}^{Desolv.}_{A} + \Delta \mathrm{G}^{Desolv.}_{B} \tag{B.3}$$

That is, the electrostatic binding free energy consists of the direct Coulombic interactions between the binding partners, the solvent screening of this interaction, and the desolvation penalty of each binding partner. Typically the first two components are

γ (cal·mol^{-1}·Å^{-2})	b (cal·mol^{-1})	Source
5.4	920	Fit to hydrocarbon solvation energies with PARSE parameter set [140].
25	0	From solubilities of hydrocarbons [24].
75	0	Macroscopic surface tension of water.

Table B-1: Source of parameters for surface-area based solvation energy term.

computed as a single value, the solvent screened Coulombic interaction, along with a value for each desolvation penalty.

B.1.3 Surface area dependent binding free energies

In addition to the electrostatic component of solvation, there is an energetic cost to form a cavity in water. This is largely an entropic term, and varies approximately linearly with surface area of the cavity. Thus, when two molecules bind, the difference in surface area between the unbound molecules and the complex will contribute to the binding free energy. This is generally a favorable contribution to binding, since the surface area of the complex is almost always smaller than that of the two free ligands. This hydrophobic solvation free energy is generally computed from the solvent-accessible surface area (A) using the relation:

$$\Delta \mathrm{G}_{h\phi}^{solv.} = \gamma A + b \tag{B.4}$$

The choice of γ and b is open to some debate, with values of γ ranging from 5 to 75 cal·mol^{-1}·Å^{-2} and some question over the importance of a non-zero value for b. Some of the most commonly values for γ and b are summarized in Table B-1 (note that the energy units are *calories* not kilocalories).

B.1.4 Calculation of relative free energies in solution

Very often the relative free energies of a set of molecules in solution are desired. This is required for evaluating the relative free energies of various conformations of a molecule, for determining the preferred titration state(s) of a molecule, and for computing the relative binding free energies of a set of mutants, as a few commonly encountered examples. A relatively fast procedure for these types of computations involves the following thermodynamic cycle:

$$\begin{tikzcd}
\mathrm{A}_{solution} \arrow[r, "\Delta\mathrm{G}^{\mathrm{A}\rightarrow\mathrm{B}}_{solution}"] \arrow[d, "-\Delta\mathrm{G}^{hydration}_{\mathrm{A}}"'] & \mathrm{B}_{solution} \\
\mathrm{A}_{vacuum} \arrow[r, "\Delta\mathrm{G}^{\mathrm{A}\rightarrow\mathrm{B}}_{vacuum}"'] & \mathrm{B}_{vacuum} \arrow[u, "\Delta\mathrm{G}^{hydration}_{\mathrm{B}}"']
\end{tikzcd} \tag{B.5}$$

This cycle yields an equation for $\Delta\mathrm{G}^{\mathrm{A}\rightarrow\mathrm{B}}_{solution}$ of:

$$\Delta\mathrm{G}^{\mathrm{A}\rightarrow\mathrm{B}}_{solution} = -\Delta\mathrm{G}^{hydration}_{\mathrm{A}} + \Delta\mathrm{G}^{\mathrm{A}\rightarrow\mathrm{B}}_{vacuum} + \Delta\mathrm{G}^{hydration}_{\mathrm{B}} \tag{B.6}$$

Each of state A and B may consist of a single molecule, as when conformational energies are being computed, or of multiple molecules, as in the case of binding, where state A may consist of the binding partners infinitely separated and state B may consist of the complex itself. The energies obtained by this procedure are not dependent on the choice of reference state, and thus can be directly compared for a set of related transformations. However, it should be noted that for the comparison to be physically meaningful, and least suspect of procedural artifacts, all systems should be treated as identically as possible.

The hydration free energies of both states can be quickly computed in the continuum solvation model using a Poisson–Boltzmann/Surface Area treatment:

$$\Delta \mathrm{G}^{hydration} = \Delta \mathrm{G}^{PB} + \Delta \mathrm{G}^{SA} \tag{B.7}$$

The Poisson–Boltzmann contribution can be computed by any of numerous solvers, such as DELPHI. The surface area component is generally computed from the solvent-accessible surface area (A), which is computable by several software packages including MSMS and CHARMM, using the relation:

$$\Delta \mathrm{G}^{SA} = \gamma A + b \tag{B.8}$$

Values of γ ranging from 5 to 75 cal/mol·Å^2 have been suggested, and the importance of a non-zero value for b is also open to debate.

The relative free energies in vacuum can be calculated using the empirical force-field of choice, or a semi-empirical or quantum mechanical method. For biological macromolecular systems, one of the standard biomolecular empirical force-field packages (such as the CHARMM-based PARAM19 or PARAM22) are most typically used.

There are several points which should be emphasized. First, the "vacuum" state is used primarily as a reference state so as to enable solvation energies to computed easily in a continuum based method. As a result, this state does not need to conform to a physically realistic state, but rather should be chosen for the ease of computation of the relative free energies *in vacuo*. For example, an internal dielectric constant of greater than 1 is often used for continuum solvation calculations, with $\epsilon_{int} = 2$ and $\epsilon_{int} = 4$ both commonly used. In these cases, the "vacuum" state should be chosen as a uniform dielectric of the ϵ_{int} used in the continuum electrostatic calculations. With this choice, the "vacuum" electrostatic energy can be computed simply using Coulomb's Law in the appropriate dielectric. A second important point involves the choice of parameters for each computation. For the most accurate final free energy, each leg of the thermodynamic cycle should be computed as accurately as possible. For solvation free energies, this may involve using a parameter set (such as PARSE) designed for use in continuum electrostatic calculations, whereas for *in vacuo* energies, a parameter set designed with these types of calculations in mind may be more appropriate. Since it is a *difference* in free energies that is being computed for each leg of the cycle, the reference states assumed for each type of calculation and each parameter set have been eliminated in each of the computed ΔGs. Thus, it is entirely valid to use a different set of parameters for each computation, and in many cases this may be the preferred treatment. It is similarly valid to use any desired variations in the force-field used for the computation of each leg — for example, instituting non-bond cutoffs and exclusion or scaling of 1–2, 1–3, and 1–4 electrostatic interactions in the *in vacuo* free energy calculation is valid, even if these same exclusions are not present in the computation of the solvation free energies.

B.1.5 Component analysis of electrostatic interactions

In order to fully understand the role of electrostatic interactions in a complex, it is useful to be able to break down the electrostatic binding free energy into contributions from the various portions of each molecule. For proteins, we divide every residue into a side chain, backbone carbonyl and backbone amino group, and for nucleic acids we define a base, ribose and phosphate group for every nucleotide. For small molecules, the definition of groups will vary dependent on the molecular structure. Once the groups are defined, we determine the desolvation energy of every individual group. We also determine, for every pair of groups in each molecule, the *difference* in the solvent screened Coulombic interaction between the pair in the bound and unbound states (we refer to these as *indirect* interactions). In addition, for every pairing of groups between the molecules, the solvent screened Coulombic interaction in the bound state is computed — these are the *direct* interactions. For a given group, the sum of its desolvation and all indirect and direct interactions gives the mutation free energy, the energetic cost, or gain, of mutating that group to a hydrophobic isostere We also define a contribution energy, which is the sum of a group's desolvation and *half* of its indirect and direct interactions. Summing the contribution energy of all groups yields the net electrostatic binding free energy of the complex.

The component analysis framework is easily applied to considerations of protein stability as well as affinity. The groups can be defined in the same way, but some model of the unfolded state must be used as the reference state. For protein side chains, the side chain alone in solution is often used, although many other choices are valid. For the protein backbone, on the other hand, a reasonable model is somewhat less clear — a short section of protein backbone is one possibility. In many cases, the direct contribution to stability of the backbone is ignored, and only the interactions that backbone groups make with side chains are considered.

B.1.6 Electrostatic complementarity: The residual potential

The consideration of electrostatics in binding involves balancing favorable interactions between the members of the complex in the bound state with the loss of favorable interactions each component makes with solvent on the transition from the unbound to the bound state. It arises from electrostatic optimization theory that in a perfectly complementary complex, this balance is met by having the interaction potential of the receptor opposite in sign and equal in magnitude the desolvation potential of the ligand. Thus the residual potential is defined as:

$$\phi^{resid} = \phi^{inter}_{rec} + \phi^{desolv}_{lig} \tag{B.9}$$

The residual potential is near zero in regions of high complementarity and is larger in magnitude in regions which are uncomplementary. It is important to note that the definition of the residual potential is fundamentally asymmetric, describing the complementarity of one component defined as the ligand, for binding the other component defined as the receptor. A complex in which one ligand which is perfectly complementary to its receptor is generally not perfectly complementary when the roles of the components are reversed; the receptor is not perfectly complementary to the ligand. A numerical measure of the complementarity of a ligand for its receptor can be attained from the correlation of ϕ_{rec}^{inter} and ϕ_{lig}^{desolv} over all points of interest, typically the ligand surface:

$$r = \frac{\sum[(\phi_{rec}^{inter} - \overline{\phi}_{rec}^{inter})(\phi_{lig}^{desolv} - \overline{\phi}_{lig}^{desolv})]}{[\sum(\phi_{rec}^{inter} - \overline{\phi}_{rec}^{inter})^2 \sum(\phi_{lig}^{desolv} - \overline{\phi}_{lig}^{desolv})^2]^{1/2}} \tag{B.10}$$

In a perfectly complementary system the correlation coefficient will be -1. Negative values smaller in magnitude indicate imperfect complementarity, while positive values indicate anti-complementarity, with the sign of the desolvation potential of the ligand matching that of the interaction potential of the receptor in an overall sense.

B.1.7 Optimization of electrostatic binding free energy

The electrostatic contribution to the binding free energy includes the ligand and receptor desolvation penalties and the bound-state screened Coulombic interaction between the ligand and the receptor. This can be written in matrix notation as:

$$\Delta \mathrm{G}^{es} = \vec{Q_l}^{\dagger}\mathbf{L}\vec{Q_l} + \vec{Q_r}^{\dagger}\mathbf{C}\vec{Q_l} + \vec{Q_r}^{\dagger}\mathbf{R}\vec{Q_r} \tag{B.11}$$

where $\vec{Q_l}$ and $\vec{Q_r}$ are the ligand and receptor charge distributions, $\mathbf{L}$ is the ligand desolvation matrix, $\mathbf{R}$ is the receptor desolvation matrix, and $\mathbf{C}$ is the solvent screened interaction matrix.

For a given receptor, $\vec{Q_r}$ is fixed, allowing a variational binding free energy to be defined:

$$\Delta \mathrm{G}^{var} = \vec{Q_l}^{\dagger}\mathbf{L}\vec{Q_l} + \vec{C}_{Q_r}^{\dagger}\vec{Q_l} \tag{B.12}$$

in which the only variable is the ligand charge distribution, $\vec{Q_l}$. Since the ligand desolvation penalty must be unfavorable for any physically meaningful geometry, the matrix $\mathbf{L}$ is positive definite, and thus $\Delta \mathrm{G}^{var}$ forms a concave-up paraboloid in ligand charge space. As a result, a single minimum variational binding free energy can be found by setting the derivative of $\Delta \mathrm{G}^{var}$ to zero. The resulting optimal ligand charge distribution and its variational binding free energy is given by:

$$\vec{Q_l}^{opt} = -\tfrac{1}{2}\mathbf{L^{-1}}\vec{C}_{Q_r} \tag{B.13}$$

$$\Delta \mathrm{G}^{opt} = -\tfrac{1}{4}\vec{C}^{\dagger}_{Q_r}\mathbf{L^{-1}}\vec{C}_{Q_r} \tag{B.14}$$

This simple optimization procedure can be extended by applying various constraints. The total charge on the system can be fixed to a given value, or can be required to be an integer. Subsets of charges can also be optimized, with the remaining charges fixed either at wild-type values or at some other reference value. Charges can also be constrained in a proportional manner, which is particularly useful for optimizing chemically equivalent groups, and for optimizing in a more "chemical" ligand-charge space. Additional manipulations of the ligand charge distribution can be used to optimize specificity of binding, either in general or against a given decoy, and this optimization can be combined with the affinity optimization.

B.2 Theory

B.2.1 The continuum electrostatic model

To treat solvent explicitly in a computation, by placing the system of interest within a large region of individually considered solvent molecules, is very costly. A significantly less computationally intensive approach is to employ a continuum model, considering the effects of solvent as a bulk entity, rather than as a microscopically distinct ensemble of molecules. For hydrophobic interactions, this treatment most often leads to a surface area dependent energy term, whereas for electrostatic interactions, a dielectric continuum model is frequently used. In the continuum electrostatic approach, molecules are generally described as a set of point charges located at atomic centers embedded in a region of low dielectric constant described by the molecular surface, with the solvent treated as a region of higher dielectric constant with, possibly, some concentration of mobile ions. The electrostatic potential produced by such a system can be obtained by solution of the Poisson–Boltzmann equation:

$$\vec{\nabla} \cdot [\epsilon(\vec{\mathrm{r}})\vec{\nabla}\phi(\vec{\mathrm{r}})] - \epsilon(\vec{\mathrm{r}})\kappa^2(\vec{\mathrm{r}})\mathrm{sinh}[\phi(\vec{\mathrm{r}})] = -4\pi\rho(\vec{\mathrm{r}}) \tag{B.15}$$

where $\kappa^2 = \frac{8\pi z^2 I}{ekT}$ describes the effect of mobile ions using a Debye–Hückel model. From the electrostatic potential, the electrostatic free energy of the system is then given by $\mathrm{G} = \frac{1}{2}\sum_i \phi_i q_i$, with the sum taken over all charges.

When the electrostatic potential in solvent is relatively small, as is the case for many systems of biological interest, the Poisson–Boltzmann equation can be linearized, replacing the hyperbolic sine dependence of the salt term with the first term in the series expansion ($\sinh(x) = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots$) and yielding:

$$\vec{\nabla} \cdot [\epsilon(\vec{\mathrm{r}})\vec{\nabla}\phi(\vec{\mathrm{r}})] - \epsilon(\vec{\mathrm{r}})\kappa^2(\vec{\mathrm{r}})\phi(\vec{\mathrm{r}}) = -4\pi\rho(\vec{\mathrm{r}}) \tag{B.16}$$

Within this linearized Poisson–Boltzmann model, all charges act independently, and thus the contributions to the electrostatic free energy from various parts of the system are separable; the contribution from any subset of the system can be considered independently, with the total energy being a simple sum of the various parts. As a result, the binding free energy can easily be partitioned into the contributions from each molecule, each functional group, or even each atom.

B.2.2 Electrostatic contributions to binding

Separating the contributions of various chemical groups provides a logical separation of contributions to the energy into three terms for every group. These are the desolvation energy of the individual group, the solvent-screened interactions between the group and all groups on the binding partner in the bound state, and the difference in solvent screening of the interactions between the group and other groups on the same molecule in the bound and unbound states. These are termed the desolvation, the direct interactions, and the indirect interactions, respectively, and we can reconstitute the full electrostatic binding energy by:

$$\Delta \mathrm{G}^{es} = \sum_i \Delta \mathrm{G}_i^{solv.} + \sum_i \sum_j \Delta \mathrm{G}_{ij}^{dir.} + \frac{1}{2} \sum_i \sum_j \Delta \mathrm{G}_{ij}^{indir.} \tag{B.17}$$

with the indirect terms halved to avoid double counting.

In addition to the individual group solvation energies and the pair-wise interaction energies, two measures of the overall contribution of a group to the binding free energy can readily be defined. The first, denoted the mutation energy, corresponds the difference in binding free energy of the natural system and that of a hypothetical system in which the group in question (and only that group) is replaced by a hydrophobic isostere. That is, the mutation energy is the energy gained by "turning on" the charges on the group of interest in the context of the natural charges on the rest of the system. When the charges on a group are eliminated, all the interactions made by that group are lost along with the desolvation of the group, and thus the mutation energy is defined as:

$$\Delta \mathrm{G}_i^{mut.} = \Delta \mathrm{G}_i^{solv.} + \sum_j \Delta \mathrm{G}_{ij}^{dir.} + \sum_j \Delta \mathrm{G}_{ij}^{indir.} \tag{B.18}$$

While the mutation energy is particularly useful in that it corresponds exactly to a physical transformation, it suffers one drawback — the sum of the mutation energies of every group does not equal the binding free energy, as all interactions are counted twice. As it is useful for understanding a system to be able to partition the energy between groups, the contribution energy is defined as:

$$\Delta\mathrm{G}_i^{contrib.} = \Delta\mathrm{G}_i^{solv.} + \frac{1}{2}\sum_j \Delta\mathrm{G}_{ij}^{dir.} + \frac{1}{2}\sum_j \Delta\mathrm{G}_{ij}^{indir.} \tag{B.19}$$

such that the sum of all the contribution energies is the total electrostatic binding free energy. While useful for partitioning the energy between groups in a meaningful way, the contribution energy does not correspond to any physical transformation. Thus neither the contribution nor the mutation energy is a perfect measure, but both are complementary, and used together can give significant insight into how various groups contribute to the overall energetics of binding.

The overall electrostatic contribution to the free energy of a ligand (l) binding to a receptor (r) can be written as:

$$\Delta\mathrm{G}^{es} = \Delta\mathrm{G}_{r,l}^{int.} + \Delta\mathrm{G}_l^{hyd.} + \Delta\mathrm{G}_r^{hyd.} \tag{B.20}$$

where $\Delta\mathrm{G}_{r,l}^{int.}$ is the total solvent-screened interaction free energy between the receptor and ligand in the bound state given by:

$$\Delta\mathrm{G}_{r,l}^{int.} = \sum_{i\in r}\sum_{j\in l} \Delta\mathrm{G}_{ij}^{dir.} \tag{B.21}$$

$\Delta\mathrm{G}_l^{hyd.}$ is the change is the ligand hydration free energy on binding given by:

$$\Delta\mathrm{G}_l^{hyd.} = \sum_{i\in l} \Delta\mathrm{G}_i^{solv.} + \frac{1}{2}\sum_{i\in l}\sum_{j\in l} \Delta\mathrm{G}_{ij}^{indir.} \tag{B.22}$$

and $\Delta\mathrm{G}_r^{hyd}$ is the equivalent term for the receptor.

B.2.3 Optimization of electrostatic interactions

Breaking down the electrostatic binding free energy further, and considering every atom in the system as its own group, leads to an interesting result. When each group is an atom, each group solvation free energy can be written as $\Delta \mathrm{G}_i^{solv.} = \frac{1}{2}(\phi_{ii}^{bound} - \phi_{ii}^{unbound})q_i$, where ϕ_{ii} is the potential produced by charge i at position i. However, due to the linear response of the linearized Poisson–Boltzmann model, the potential produced by any charge at position i can be related to the potential produced by a single unit charge at the same position (Φ_i) by $\phi_i = q_i \Phi_i$. This leads to an expression of the group solvation energy in terms of the group charge and the bound and unbound potentials of a unit charge at the atomic center:

$$\Delta \mathrm{G}_i^{solv.} = \frac{1}{2} q_i (\Phi_{ii}^{bound} - \Phi_{ii}^{unbound}) q_i \tag{B.23}$$

Similarly, with single atom groups the pairwise indirect interactions can be written in terms of the potential generated by charge i at position j as $\Delta \mathrm{G}_{ij}^{indir.} = (\phi_{ij}^{bound} - \phi_{ij}^{unbound})q_j$, into which the substitution of $\phi_{ij} = q_i \Phi_{ij}$ gives:

$$\Delta \mathrm{G}_{ij}^{indir.} = q_i (\Phi_{ij}^{bound} - \Phi_{ij}^{unbound}) q_j \tag{B.24}$$

Using the same procedure for the direct interactions yields:

$$\Delta \mathrm{G}_{ij}^{dir.} = q_i (\Phi_{ij}^{bound}) q_j \tag{B.25}$$

with only the bound state potentials contributing. For both the direct and indirect interaction, $\Delta \mathrm{G}_{ij} = \Delta \mathrm{G}_{ji}$, by the reciprocity implicit in the continuum model.

Substituting Equation B.25 into Equation B.21 gives:

$$\Delta \mathrm{G}_{r,l}^{int.} = \sum_{i \in r} \sum_{j \in l} q_i (\Phi_{ij}^{bound}) q_j \tag{B.26}$$

which can be written in matrix form as $\vec{Q}_r^\dagger \mathbf{C} \vec{Q}_l$, where $\vec{Q}_r$ is a vector of the charges on the receptor, $\vec{Q}_l$ is a vector of the charges on the ligand, and the elements of the matrix $\mathbf{C}$ are given by Φ_{ij}^{bound}. In a similar fashion, substituting Equations B.23 and B.24 into Equation B.22, gives:

$$\Delta \mathrm{G}_l^{hyd.} = \frac{1}{2} \sum_{i \in l} q_i (\Phi_{ii}^{bound} - \Phi_{ii}^{unbound}) q_i + \frac{1}{2} \sum_{i \in l} \sum_{j \in l} q_i (\Phi_{ij}^{bound} - \Phi_{ij}^{unbound}) q_j \tag{B.27}$$

This too can be written in matrix form as $\vec{Q}_l^\dagger \mathbf{L} \vec{Q}_l$, where the diagonal elements of of the matrix $\mathbf{L}$ are given by $\frac{1}{2}(\Phi_{ii}^{bound} - \Phi_{ii}^{unbound})$, and the off-diagonal elements are given by $\frac{1}{2}(\Phi_{ij}^{bound} - \Phi_{ij}^{unbound})$. Naturally, the change in receptor hydration free energy on binding, $\Delta \mathrm{G}_r^{hyd.}$, can be written in the same fashion as $\vec{Q}_r^\dagger \mathbf{R} \vec{Q}_r$, with the receptor desolvation matrix, $\mathbf{R}$, analogous to the ligand desolvation matrix, $\mathbf{L}$. Combining these terms gives an expression for the overall electrostatic binding free energy in matrix form:

$$\Delta \mathrm{G}^{es} = \vec{Q}_l^\dagger \mathbf{L} \vec{Q}_l + \vec{Q}_r^\dagger \mathbf{C} \vec{Q}_l + \vec{Q}_r^\dagger \mathbf{R} \vec{Q}_r \tag{B.28}$$

B.2.4 Type-I (Affinity) Optimum

The Type-I optimum (also referred to as the affinity optimum), is the ligand charge distribution whose binding affinity to a given receptor is better than that of all other ligands of the same geometry.

For a given receptor, $\vec{Q}_r$ is fixed, allowing a variational binding free energy to be defined:

$$\Delta \mathrm{G}^{\mathrm{var.}} = \vec{Q}_l^\dagger \mathbf{L} \vec{Q}_l + \vec{Q}_r^\dagger \mathbf{C} \vec{Q}_l \tag{B.29}$$

in which the only variable is the ligand charge distribution, $\vec{Q}_l$. Since the ligand desolvation penalty must be unfavorable for any physically meaningful geometry, the matrix $\mathbf{L}$ is positive definite [77], and thus ΔG^{var} forms a concave up paraboloid in ligand charge space. As a result, a single minimum variational binding free energy can be found by setting the derivative of $\Delta\mathrm{G}^{var.}$ to zero:

$$\frac{\partial \Delta\mathrm{G}^{var.}}{\partial \vec{Q}_l} = 2\vec{Q}_l^{\dagger}\mathbf{L} + \vec{Q}_r^{\dagger}\mathbf{C} = 0 \tag{B.30}$$

The resulting optimal ligand charge distribution and its variational binding free energy are given by:

$$\vec{Q_l}^{\text{Type-I}} = -\tfrac{1}{2}\mathbf{L^{-1}}\mathbf{C}^{\dagger}\vec{Q}_r \tag{B.31}$$

$$\Delta\mathrm{G}^{\text{Type-I}} = -\tfrac{1}{4}\vec{Q}_r^{\dagger}\mathbf{C}\mathbf{L^{-1}}\mathbf{C}^{\dagger}\vec{Q}_r \tag{B.32}$$

B.2.5 Type-II (Specificity) Optimum

The Type-II optimum (also referred to as the specificity optimum), is the ligand charge distribution which binds better to a given target receptor than to any other receptor with the same geometry.

Varying the electrostatic binding free energy with respect to the receptor charges gives:

$$\frac{\partial \Delta G^{es}}{\partial \vec{Q}_r} = \mathbf{C}\vec{Q}_l + 2\mathbf{R}\vec{Q}_r = 0 \tag{B.33}$$

For a given target receptor charge distribution, $\vec{Q}_r$, the ligand charges which preferentially bind the target is then:

$$\vec{Q_l}^{\text{Type-II}} = -2\mathbf{C^{-1}}\mathbf{R}\vec{Q}_r \tag{B.34}$$

B.2.6 Type-III (Best Hapten) Optimum

The Type-III optimum (also referred to as the best hapten optimum) is somewhat more complicated. This is a ligand which, when a *receptor* is **affinity** optimized for binding to the ligand, the resulting receptor will be **specificity** optimized for binding to some other target ligand. That is, if receptor R is the Type-II optimal ligand for binding ligand X, and receptor R is also the Type-I optimal ligand for binding ligand L, then ligand L is the Type-III optimum for the target ligand X. These definitions result in a series of equations, leading to the Type-III definition. Receptor R is Type-I optimized against L:

$$\vec{Q}_R = -\frac{1}{2}\mathbf{R}^{\mathbf{-1}}\mathbf{C}^{\dagger}_{\mathbf{LR}}\vec{Q}_L \tag{B.35}$$

Receptor R is Type-II optimized against X:

$$\vec{Q}_R = -2\mathbf{C}^{\mathbf{-1}}_{\mathbf{XR}}\mathbf{X}\vec{Q}_X \tag{B.36}$$

Combining these equations and solving for the ligand L charge distribution $\vec{Q}_L$ gives:

$$\vec{Q}_L^{\text{Type-III}} = 4\mathbf{C}^{\dagger\mathbf{-1}}_{\mathbf{LR}}\mathbf{R}\mathbf{C}^{\mathbf{-1}}_{\mathbf{XR}}\mathbf{X}\vec{Q}_X \tag{B.37}$$

Since ligands L and X share the same geometry, and differ only in their charge distributions, $\mathbf{L} \equiv \mathbf{X}$ and $\mathbf{C_{LR}} \equiv \mathbf{C_{XR}}$, and thus we can write:

$$\vec{Q}_L^{\text{Type-III}} = 4\mathbf{C}^{\mathbf{-1}}\mathbf{R}\mathbf{C}^{\dagger\mathbf{-1}}\mathbf{L}\vec{Q}_X \tag{B.38}$$

using the standard notation for ligand–receptor binding (note the transposition of $\mathbf{C}$ to account for the standard ordering of the interaction term $\Delta G^{inter} = \vec{Q}_R^{\dagger}\mathbf{C}\vec{Q}_L$).

B.3 Optimization in real-world problems

B.3.1 Fixed charges in optimization

For realistic systems, it would be useful to be able to consider only a subset of charges on both the ligand and the receptor. In order to do this, we rewrite Equation B.20

as:

$$\begin{aligned}\Delta G^{es} &= (\vec{Q}^{\dagger}_{l,v}\mathbf{L_{vv}}\vec{Q}_{l,v} + \vec{Q}^{\dagger}_{l,v}\mathbf{L_{vf}}\vec{Q}_{l,f} + \vec{Q}^{\dagger}_{l,f}\mathbf{L_{ff}}\vec{Q}_{l,f}) \\ &+ (\vec{Q}^{\dagger}_{r,v}\mathbf{C_{vv}}\vec{Q}_{l,v} + \vec{Q}^{\dagger}_{r,v}\mathbf{C_{vf}}\vec{Q}_{l,f} + \vec{Q}^{\dagger}_{r,f}\mathbf{C_{fv}}\vec{Q}_{l,v} + \vec{Q}^{\dagger}_{r,f}\mathbf{C_{ff}}\vec{Q}_{l,f}) \\ &+ (\vec{Q}^{\dagger}_{r,v}\mathbf{R_{vv}}\vec{Q}_{r,v} + \vec{Q}^{\dagger}_{r,v}\mathbf{R_{vf}}\vec{Q}_{r,f} + \vec{Q}^{\dagger}_{r,v}\mathbf{R_{ff}}\vec{Q}_{r,f})\end{aligned} \tag{B.39}$$

where all the terms have been split up into contributions from variable (v) and fixed (f) regions, the first line describing the ligand desolvation, the second line describing the ligand–receptor interaction, and the last line describing the receptor desolvation.

With this description, the equations for the Type-I optimum become:

$$\frac{\partial \Delta G}{\partial \vec{Q}_{l,v}} = 2\mathbf{L_{vv}}\vec{Q}_{l,v} + \mathbf{L_{vf}}\vec{Q}_{l,f} + \mathbf{C}^{\dagger}_{\mathbf{vv}}\vec{Q}_{r,v} + \mathbf{C}^{\dagger}_{\mathbf{fv}}\vec{Q}_{r,f} \tag{B.40}$$

$$\vec{Q}^{\text{Type-I}}_{l,v} = -\frac{1}{2}\mathbf{L}^{\mathbf{-1}}_{\mathbf{vv}}(\mathbf{L_{vf}}\vec{Q}_{l,f} + \mathbf{C}^{\dagger}_{\mathbf{vv}}\vec{Q}_{r,v} + \mathbf{C}^{\dagger}_{\mathbf{fv}}\vec{Q}_{r,f}) \tag{B.41}$$

Similarly, the equations for the Type-II optimum become:

$$\frac{\partial \Delta G}{\partial \vec{Q}_{r,v}} = \mathbf{C_{vv}}\vec{Q}_{l,v} + \mathbf{C_{vf}}\vec{Q}_{l,f} + 2\mathbf{R_{vv}}\vec{Q}_{r,v} + \mathbf{R_{vf}}\vec{Q}_{r,f} \tag{B.42}$$

$$\vec{Q}^{\text{Type-II}}_{l,v} = -\mathbf{C}^{\mathbf{-1}}_{\mathbf{vv}}(\mathbf{C_{vf}}\vec{Q}_{l,f} + 2\mathbf{R_{vv}}\vec{Q}_{r,v} + \mathbf{R_{vf}}\vec{Q}_{r,f}) \tag{B.43}$$

The Type-III optimum is quite cumbersome in this description, but is included for completeness:

$$\begin{aligned}\vec{Q}^{\text{Type-III}}_{l,v} &= \mathbf{C}^{\mathbf{-1}}_{\mathbf{vv}}[2\mathbf{R_{vv}}\mathbf{C}^{\dagger\mathbf{-1}}_{\mathbf{vv}}(\mathbf{C}^{\dagger}_{\mathbf{fv}}\vec{Q}_{R,f} + 2\mathbf{L_{vv}}\vec{Q}_{X,v} + \mathbf{L_{vf}}\vec{Q}_{X,f}) \\ &- (\mathbf{R_{vf}}\vec{Q}_{R,f} + \mathbf{C_{vf}}\vec{Q}_{L,f})]\end{aligned} \tag{B.44}$$

B.3.2 Simultaneous optimization of multiple ligands

Many naturally occurring systems involve multiple ligands binding to a single receptor. This can either be due to the existence of multiple binding sites for a single molecular species, or to the binding of multiple molecular species, to either overlapping or separated binding sites. Extending the optimization formulation to these systems is relatively straightforward. Equation B.20 holds for the binding reaction:

$$L + R \rightleftharpoons C \tag{B.45}$$

so, for the more general binding reaction:

$$L_1 + L_2 + \cdots + L_m + R_1 + R_2 + \cdots + R_n \rightleftharpoons C \tag{B.46}$$

we can extend the expression for the electrostatic binding free energy to:

$$\begin{aligned} \Delta G^{es} &= \textstyle\sum_{i \in \mathrm{Lig.}} \vec{Q}_{l_i}^{\dagger} \mathbf{L_i} \vec{Q}_{l_i} + \frac{1}{2} \sum_{i \in \mathrm{Lig.}} \sum_{j \in \mathrm{Lig.}} \vec{Q}_{l_i}^{\dagger} \mathbf{C_{ij}^{LL}} \vec{Q}_{l_j} \\ &+ \textstyle\sum_{i \in \mathrm{Rec.}} \sum_{i \in \mathrm{Lig.}} \vec{Q}_{r_i}^{\dagger} \mathbf{C_{ij}^{RL}} \vec{Q}_{l_j} \\ &+ \textstyle\frac{1}{2} \sum_{i \in \mathrm{Rec.}} \sum_{i \in \mathrm{Rec.}} \vec{Q}_{r_i}^{\dagger} \mathbf{C_{ij}^{RR}} \vec{Q}_{r_j} + \sum_{i \in \mathrm{Rec.}} \vec{Q}_{r_i}^{\dagger} \mathbf{R_i} \vec{Q}_{r_i} \end{aligned} \tag{B.47}$$

where the first and last terms describe the desolvation of each ligand and receptor, and the middle three terms describe inter-ligand, ligand–receptor, and inter-receptor interactions. This can be re-written in a block-matrix form as:

$$\begin{aligned}
\Delta G^{es} \; = \; & \begin{bmatrix} \vec{Q}^{\dagger}_{l_1} & \vec{Q}^{\dagger}_{l_2} & \cdots & \vec{Q}^{\dagger}_{l_m} \end{bmatrix}
\begin{bmatrix} \mathbf{L_1} & \frac{1}{2}\mathbf{C}^{\mathbf{LL}}_{\mathbf{1,2}} & \cdots & \frac{1}{2}\mathbf{C}^{\mathbf{LL}}_{\mathbf{1,m}} \\ \frac{1}{2}\mathbf{C}^{\mathbf{LL}}_{\mathbf{2,1}} & \mathbf{L_2} & \cdots & \frac{1}{2}\mathbf{C}^{\mathbf{LL}}_{\mathbf{2,m}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}\mathbf{C}^{\mathbf{LL}}_{\mathbf{m,1}} & \frac{1}{2}\mathbf{C}^{\mathbf{LL}}_{\mathbf{m,2}} & \cdots & \mathbf{L_m} \end{bmatrix}
\begin{bmatrix} \vec{Q}_{l_1} \\ \vec{Q}_{l_2} \\ \vdots \\ \vec{Q}_{l_m} \end{bmatrix} \\
+ & \begin{bmatrix} \vec{Q}^{\dagger}_{r_1} & \vec{Q}^{\dagger}_{r_2} & \cdots & \vec{Q}^{\dagger}_{r_n} \end{bmatrix}
\begin{bmatrix} \mathbf{C}^{\mathbf{RL}}_{\mathbf{1,1}} & \mathbf{C}^{\mathbf{RL}}_{\mathbf{1,2}} & \cdots & \mathbf{C}^{\mathbf{RL}}_{\mathbf{1,m}} \\ \mathbf{C}^{\mathbf{RL}}_{\mathbf{2,1}} & \mathbf{C}^{\mathbf{RL}}_{\mathbf{2,2}} & \cdots & \mathbf{C}^{\mathbf{RL}}_{\mathbf{2,m}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}^{\mathbf{RL}}_{\mathbf{n,1}} & \mathbf{C}^{\mathbf{RL}}_{\mathbf{n,2}} & \cdots & \mathbf{C}^{\mathbf{RL}}_{\mathbf{n,m}} \end{bmatrix}
\begin{bmatrix} \vec{Q}_{l_1} \\ \vec{Q}_{l_2} \\ \vdots \\ \vec{Q}_{l_m} \end{bmatrix} \\
+ & \begin{bmatrix} \vec{Q}^{\dagger}_{r_1} & \vec{Q}^{\dagger}_{r_2} & \cdots & \vec{Q}^{\dagger}_{r_n} \end{bmatrix}
\begin{bmatrix} \mathbf{R_1} & \frac{1}{2}\mathbf{C}^{\mathbf{RR}}_{\mathbf{1,2}} & \cdots & \frac{1}{2}\mathbf{C}^{\mathbf{RR}}_{\mathbf{1,n}} \\ \frac{1}{2}\mathbf{C}^{\mathbf{RR}}_{\mathbf{2,1}} & \mathbf{R_2} & \cdots & \frac{1}{2}\mathbf{C}^{\mathbf{RR}}_{\mathbf{2,n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}\mathbf{C}^{\mathbf{RR}}_{\mathbf{n,1}} & \frac{1}{2}\mathbf{C}^{\mathbf{RR}}_{\mathbf{n,2}} & \cdots & \mathbf{R_n} \end{bmatrix}
\begin{bmatrix} \vec{Q}_{r_1} \\ \vec{Q}_{r_2} \\ \vdots \\ \vec{Q}_{r_n} \end{bmatrix}
\end{aligned} \tag{B.48}$$

with the first term describing the ligand desolvation and the inter-ligand interactions, the second term describing the ligand–receptor interactions, and the third term describing the receptor desolvation and the inter-receptor interactions. Written more simply, this is:

$$\Delta G^{es} = \vec{Q}^{\dagger}_{L}\mathbf{L}\vec{Q}_{L} + \vec{Q}^{\dagger}_{R}\mathbf{C}\vec{Q}_{L} + \vec{Q}^{\dagger}_{R}\mathbf{R}\vec{Q}_{R} \tag{B.49}$$

exactly the same form as Equation B.20. Several points deserves special attention. First, both the **L** and **R** block matrices are square, whereas the $\mathbf{C}^{\mathbf{RL}}$ block matrix is not. Secondly, the designation of "ligands" and "receptors" is completely arbitrary, particularly in the case of multiple receptors. In this case, it may be more appropriate to speak of a set of ligands forming a complex, and having the complete binding free energy described by a single block matrix term. However, in order to keep the same formulation, and to be able to extend the optimization forms in a straightforward manner, it is useful to maintain a perspective of one molecule being the receptor. In addition, for many systems it is easy to identify a single binding target, the receptor, even when multiple ligands bind. Finally, the **L** and **R** block matrices are **not** guaranteed to be positive definite when they consists of multiple sub-matrices; each diagonal sub-matrix will be positive definite, but with non-zero off-diagonal submatrices, the full matrix may not be.

The possible non-positive definiteness of **L** and **R** has important ramifications for optimization. First, the direct forms of obtaining the optimum are only valid for positive definite matrices. Otherwise, the direct form does not produce the optimal ligand, but rather the ligand at a *saddle point* on the binding free energy surface. Secondly, in the general case constraints *must* be applied during optimization, since if the matrices contain negative eigenvalues, there is no minimum on the binding free energy surface; the binding free energy approaches negative infinity as the eigenvectors corresponding to the negative eigenvalues become more highly populated. One more word on negative eigenvalues — the corresponding eigenvectors describe charges which interact cooperatively on binding, and thus populating these eigenvectors in a designed ligand should lead to favorable cooperative effects between the ligands.

B.3.3 Optimization over multiple conformations

Ligand and receptor conformations frequently change upon binding. In addition, in both the bound and unbound states, neither the ligand nor the receptor is truly in one conformation, but rather in an ensemble of states, perhaps all of a similar average conformation, or perhaps in several distinct conformations with variation about each one. It would be useful to incorporate these variations into the optimization framework.

For the simple case of a change in conformation on binding, with a single conformation for the unbound receptor, the unbound ligand, and the complex, the formalism of Equation B.20 remains the same, but the matrices must be redefined. The interaction matrix, **C**, involves only the bound state and thus remains unchanged.

The ligand and receptor desolvation matrices, **L** and **R**, on the other hand involve both the unbound and bound states, and thus must incorporate the conformational change. In the rigid-body binding approximation, the diagonal elements of of the desolvation matrices are given by $\frac{1}{2}(\Phi_{ii}^{bound} - \Phi_{ii}^{unbound})$, and the off-diagonal elements given by $\frac{1}{2}(\Phi_{ij}^{bound} - \Phi_{ij}^{unbound})$. While the same description is formally valid for non-rigid binding, operationally a different definition is useful. In the frequently used finite-difference solution of the Poisson–Boltzmann equation, a difference of two states placed identically on the finite-difference grid is required in order to cancel an artifactual “grid-energy” which is a result of the method. When the bound and unbound geometries are different, this requires the potential to be broken up into contributions from the reaction field and from Coulomb's Law. Thus we rewrite $(\Phi^{bound} - \Phi^{unbound})$ as:

$$(\Phi^{bound} - \Phi^{unbound}) = [(\Delta\Phi_{solv.}^{bound} + \Phi_{Coul.}^{bound}) - (\Delta\Phi_{solv.}^{unbound} + \Phi_{Coul.}^{unbound})] \tag{B.50}$$

where $\Delta\Phi_{solv.}$ is the change in potential upon moving the molecule from a uniform dielectric medium to a solvated state, and $\Phi_{Coul.}$ is the Coulombic potential in the same uniform dielectric. Defining two new matrices for both the ligand and receptor: $\mathbf{L_{solv.}}$ and $\mathbf{R_{solv.}}$ as the matrices of solvation potentials and $\mathbf{L_{Coul.}}$ and $\mathbf{R_{Coul.}}$ as the Coulombic potential matrices, allows for the simple substitution of:

$$\mathbf{L} = \mathbf{L_{solv.}^{bound}} + \mathbf{L_{Coul.}^{bound}} - \mathbf{L_{solv.}^{unbound}} - \mathbf{L_{Coul.}^{unbound}} \tag{B.51}$$

and the analogous expression for **R**. It should be noted that in this definition, the **L** and **R** matrices are *not* guaranteed to be positive definite.

For the more complex case, where multiple states exist for the ligand, receptor, or the complex, several definitions are possible, depending on the approximations made. These approximations, which contribute substantially to the resulting complexity of the model, involve the relative energies of the conformations in each ensemble, and the effect of optimization of these relative energies.

Simple averaging. The simplest case makes two key simplifications: **(1)** All conformations are assumed to be energetically degenerate, and thus are equally populated; **(2)** Varying the charges on each ligand does not affect the relative populations of the different conformations. These approximations allow the contribution of each state to the energetics to be combined with a simple average, yielding the following expression for the free energy of binding:

$$\begin{aligned}
\Delta \mathrm{G}^{es} &= \vec{Q}_l^\dagger \left[\frac{\sum_{i=1}^{N^C} (\mathbf{L}_i^{solv.} + \mathbf{L}_i^{Coul.})}{N^C} - \frac{\sum_{j=1}^{N^L} (\mathbf{L}_j^{solv.} + \mathbf{L}_j^{Coul.})}{N^L} \right] \vec{Q}_l \\
&+ \vec{Q}_r^\dagger \left[\frac{\sum_{i=1}^{N^C} \mathbf{C}_i}{N^C} \right] \vec{Q}_l \\
&+ \vec{Q}_r^\dagger \left[\frac{\sum_{i=1}^{N^C} (\mathbf{R}_i^{solv.} + \mathbf{R}_i^{Coul.})}{N^C} - \frac{\sum_{j=1}^{N^R} (\mathbf{R}_j^{solv.} + \mathbf{R}_j^{Coul.})}{N^R} \right] \vec{Q}_r
\end{aligned} \tag{B.52}$$

The central terms are all reducible to simple matrices, and thus this has the basic form of Equation B.20. Once again however, the guarantee of positive definite desolvation matrices is lost.

Pre-Boltzmann weighting. A slightly more complex model relaxes the first simplification, describing the relative populations of the various conformations by a Boltzmann weighting of the wild-type energies. However the second assumption, that these populations are not affected by varying the ligand charges, remains. Under these assumptions, the averages taken must be Boltzmann weighted, but as the charges do not affect the weighting, the averaging can once again be done directly on the matrices, yielding:

$$\begin{aligned}
\Delta \mathrm{G}^{es} &= \vec{Q}_l^\dagger \left[\frac{\sum_{i=1}^{N^C} e^{-\beta \varepsilon_i^C} (\mathbf{L}_i^{solv.} + \mathbf{L}_i^{Coul.})}{\sum_{i=1}^{N^C} e^{-\beta \varepsilon_i^C}} - \frac{\sum_{j=1}^{N^L} e^{-\beta \varepsilon_j^L} (\mathbf{L}_j^{solv.} + \mathbf{L}_j^{Coul.})}{\sum_{j=1}^{N^L} e^{-\beta \varepsilon_j^L}} \right] \vec{Q}_l \\
&+ \vec{Q}_r^\dagger \left[\frac{\sum_{i=1}^{N^C} e^{-\beta \varepsilon_i^C} \mathbf{C}_i}{\sum_{i=1}^{N^C} e^{-\beta \varepsilon_i^C}} \right] \vec{Q}_l \\
&+ \vec{Q}_r^\dagger \left[\frac{\sum_{i=1}^{N^C} e^{-\beta \varepsilon_i^C} (\mathbf{R}_i^{solv.} + \mathbf{R}_i^{Coul.})}{\sum_{i=1}^{N^C} e^{-\beta \varepsilon_i^C}} - \frac{\sum_{j=1}^{N^R} e^{-\beta \varepsilon_j^R} (\mathbf{R}_j^{solv.} + \mathbf{R}_j^{Coul.})}{\sum_{j=1}^{N^R} e^{-\beta \varepsilon_j^R}} \right] \vec{Q}_r
\end{aligned} \tag{B.53}$$

in which ε_i is the energy of conformation i, including any energy terms desired, including non-electrostatic contributions. Once again, this an equation of the same form as Equation B.20.

Full Boltzmann weighting. In the final, most complicated, model the second assumption is also relaxed. The populations of the various conformations are dictated by a Boltzmann distribution of the conformational energies, and the variation of ligand charges may affect the distribution by perturbing the energies. In this case, the Boltzmann averaging can not be done on the matrices, but rather must be recalculated individually for every charge distribution. The energy of a state can be written as $\mathrm{G}_i = \varepsilon_i^{np} + \vec{Q}_{l,i}^{\dagger}\mathbf{L}_i\vec{Q}_{l,i}$, with the first term encompassing all non-electrostatic terms, and the second describing the electrostatic contribution ($\mathbf{L}_i \equiv \mathbf{L}_i^{solv.} + \mathbf{L}_i^{Coul.}$). Separating out the non-electrostatic energy terms, indicated by $\xi_i = e^{-\beta\varepsilon_i^{np}}$, allows the energy of the unbound ligand ensemble to be written as:

$$\mathrm{G}^{L,es} = \frac{\sum_{j=1}^{N^L} \xi_j^L e^{-\beta(\vec{Q}_{l,j}^{\dagger}\mathbf{L}_j\vec{Q}_{l,j})}(\vec{Q}_{l,j}^{\dagger}\mathbf{L}_j\vec{Q}_{l,j})}{\sum_{j=1}^{N^L} \xi_j^L e^{-\beta(\vec{Q}_{l,j}^{\dagger}\mathbf{L}_j\vec{Q}_{l,j})}} \tag{B.54}$$

Similarly, for the unbound receptor ensemble, we have:

$$\mathrm{G}^{R,es} = \frac{\sum_{j=1}^{N^R} \xi_j^R e^{-\beta(\vec{Q}_{r,j}^{\dagger}\mathbf{R}_j\vec{Q}_{r,j})}(\vec{Q}_{r,j}^{\dagger}\mathbf{R}_j\vec{Q}_{r,j})}{\sum_{j=1}^{N^R} \xi_j^R e^{-\beta(\vec{Q}_{r,j}^{\dagger}\mathbf{R}_j\vec{Q}_{r,j})}} \tag{B.55}$$

For the ensemble of complexes, we have a similar expression, but the electrostatic energy of a single complex conformation must be written in terms of the ligand and receptor charge distributions. This expression is:

$$\mathrm{G}_i^{C,es} = \vec{Q}_{l,i}^{\dagger}\mathbf{L}_i^C\vec{Q}_{l,i} + \vec{Q}_{r,i}^{\dagger}\mathbf{C}_i\vec{Q}_{l,i} + \vec{Q}_{r,i}^{\dagger}\mathbf{R}_i^C\vec{Q}_{r,i} \tag{B.56}$$

where $\mathbf{L}_i^C \equiv \mathbf{L}_i^{C,solv.} + \mathbf{L}_i^{C,Coul.}$ gives the self-energy of the ligand in the context of the complex conformation i, and $\mathbf{R}_i^C \equiv \mathbf{R}_i^{C,solv.} + \mathbf{R}_i^{C,Coul.}$ gives the self-energy of the receptor in the same context. As in all cases, $\mathbf{C}_i$ describes the solvent-screened interaction between the ligand and receptor in the bound state. This gives the expression for the bound complex ensemble as:

$$\mathrm{G}^{C,es} = \frac{\sum_{i=1}^{NC} \xi_i^C e^{-\beta \mathrm{G}_i^{C,es}} \mathrm{G}_i^{C,es}}{\sum_{i=1}^{NC} \xi_i^C e^{-\beta \mathrm{G}_i^{C,es}}} \tag{B.57}$$

The electrostatic binding free energy is then given by $\Delta\mathrm{G}^{es} = \mathrm{G}^{C,es} - (\mathrm{G}^{R,es} + \mathrm{G}^{L,es})$. This does **not** conform to the standard of Equation B.20, and thus the standard methods of solution do not apply. In particular, this formulation may have multiple minima, which makes the optimization procedure much more complicated.

B.3.4 Poor behavior of optimization matrices

The ligand and receptor desolvation matrices are required by physics to be positive definite. However, due to numerical approximations in the methods used to compute the elements of these matrices, the actual computed matrices may not be, and small negative eigenvalues may be observed in some cases. In addition, the matrix inversion procedures used to obtain the direct solutions for the optimal charge distributions can be poorly behaved when very small positive eigenvalues are present, as a result of the need to take the inverse of these values. As a result, pre-conditioning of the matrices for use in the optimization procedure is required. In particular, singular value decomposition is used the remove all negative and small positive eigenvalues from the matrices. The eigenvectors corresponding to these eigenvalues form the null-space, and are generally excluded from the optimization. However, in certain cases when constraints are applied during optimization, the incompleteness of the basis set used in the optimization makes the satisfaction of the imposed constraints impossible. In these circumstances, it becomes necessary to allow eigenvectors from the null-space to be used, but only to satisfy the constraints. This is achieved by placing an artificial harmonic penalty on the vectors of the null-space, with no linear component — any deviation from a non-zero population of the null-space will result in an energetic penalty during optimization, and thus only when absolutely necessary will these vectors be used.

B.3.5 Application of constraints

All the results to this point have focused on describing an analytical solution to the globally optimal ligand. However, in many cases, this global optimum may be unphysical, with partial atomic charges unrealizable in a chemical system, with non-integral net charges, or with various other pathogenic behaviors. In these instances, the global solution to the optimization problem may not be desired, but rather the optimal solution which satisfies input constraints of the total charge of the system and on maximal partial atomic charges may be the preferred target. This can easily be accomplished by minimizing the target function described for each optimum (for the Type-I optimum this is simply Equation B.11) subject to the applied constraints. Due to the quadratic nature of the free energy surface, linear and quadratic programming methods can make this optimization very efficient, and the LOQO optimization package [133, 154, 155] is used for these constrained optimizations. Constraints can also be used to focus on particular regions of a molecule during optimization, or to enforce proportionality of particular partial atomic charges.

B.4 Overview of Program

The ICE software suite is built around two sets of libraries — one for C^{++} and one for PERL — which implement the bulk of the functionality. Interfacing with these libraries are several pieces of software, some of which are meant to be directly run by the end user, and others which are generally called by these programs and not directly executed by the end user.

The first key end user program is the script *delphi.prl*. Written in PERL, this provides a comprehensive interface to the finite-difference Poisson–Boltzmann solver DELPHI [55, 57, 134, 136], although extensions to other solvers are planned. All continuum electrostatic calculations are run through this script.

For component analyses, two scripts are used. The first, *comp_anal.prl*, sets up and executes all the continuum electrostatic calculations needed for component analysis. This script also compiles the results of the calculations into a format suitable for analysis. The results are analyzed using the script *rank_comp.prl*, which allows sorting through the results based on numerous values.

To perform electrostatic optimizations, again two pieces of software are used. The first is the script *matrix_elements.prl*, which generates and runs the individual continuum electrostatic jobs required for optimization. This script also compiles the results of the calculations into the binary format required by *optimize*. During this procedure, two additional programs are called indirectly: *binpot*, a program which compiles the information from the DELPHI output FRC files into a single binary file, and *getmatrix*, a program which reads in the potential.bin files (created by *binpot*) for each calculation and compiles the results into a single binary file. The electrostatic optimization protocol itself is implemented in *optimize*. Simple optimizations are performed using routines derived from *Numerical Recipes in C* [119], while constraints are implemented through an invisible interface to the LOQO computer program [133, 154, 155].

Optimization of protein ligands is automated in the *protein_scan.prl* script. This script will perform an optimization on every residue in a protein, fixing the residue charge to every integer within a specified range (such as -1 to $+1$ e), and output the results both numerically tabulated and visually mapped onto the protein structure.

Finally, the ligand scanning methodology is implemented in *ligand_scan.prl*. This script implements multiple stages of ranking a set of ligand derivatives, with the ability to estimate charges by specified rules, to setup and run all required quantum mechanical calculations for the computation of partial atomic charges, and to compute binding free energies using the correct shape and charge distribution.

B.4.1 Continuum electrostatic calculations

Calculations of simple transformations can easily be performed directly through *delphi.prl*. The most commonly used examples are a computation of the solvation free energy of a molecule, and the computation of the electrostatic binding free energy of a system. For every calculation, one group of charges is defined for the computation of the potential, and the interaction of this group with any number of other groups of charges can be obtained from this potential. For solvation energies to be meaningful, it is essential that two calculations are performed, with an identical placement of the charged group on the finite-difference grid in both calculations.

1. Compute electrostatic free energies for a system.
    - **Script:** delphi.prl
    - **Syntax:** `delphi.prl`
    - **Summary:** Sets up and runs a series of DELPHI jobs for the evaluation of the electrostatic energy of a system. The parameters of the calculation are described in the "param.file" and the description of the calculation in the "run.file".

B.4.2 Component analysis

The calculations required for component analysis are two calculations (bound and unbound states, or folded and unfolded, states) for each group in the system. In general, for either proteins or nucleic acids, this results in three jobs for every residue. The potentials of the two states are computed for every group, and the interactions of this group with every other group is obtained from these potentials.

1. Compute components.

- **Script:** comp_anal.prl
- **Syntax:** `comp_anal.prl crdfile [run]`
- **Summary:** Sets up and runs all DELPHI jobs required for component analysis, reading the configuration from "component.cfg". If the "run" option specified, checks for completion of jobs and re-submits incomplete calculations.

2. Compile component computation output into a suitable format for analysis.

   - **Script:** comp_anal.prl
   - **Syntax:** `comp_anal.prl crdfile`
   - **Summary:** Reads output files from DELPHI calculations and compiles them into text matrix files with all data required for component analysis. Checks for normal completion of DELPHI jobs, analyzing only completed components.

3. Analyze results.

   - **Script:** rank_comp.prl
   - **Syntax:** `rank_comp.prl [options]`
   - **Summary:** Performs any of a variety of analyses on the results of a component analysis. Components can be ranked by any energy value, and details of individual interactions can be listed.

B.4.3 Optimization

For optimization, a number of calculations must be done to obtain the elements of the matrices used to define the binding free energy. For every variable ligand or receptor charge, one set of bound/unbound calculations must be done. These can be used to give all the required matrix elements. However, if there are any fixed ligand or receptor charges, an additional calculation (bound/unbound) of each of these sets of charges must be done. If this is excluded, while the optimization may proceed, the total electrostatic energies will be incomplete.

1. Compute all matrix elements.
    - **Script:** matrix_elements.prl
    - **Syntax:** `matrix_elements.prl matrix.cfg`
    - **Summary:** Sets up and runs all DELPHI jobs required for optimization of selected basis points, as described in the "matrix.cfg" configuration file. Also sets up and runs additional jobs require to compute the entire electrostatic binding free energy. Checks for completion of jobs and that "residual" ligand calculation applies to the current selection of basis points.
2. Compile matrix element computation output into appropriate format for *optimize*.
    - **Script:** matrix_elements.prl
    - **Syntax:** `matrix_elements.prl matrix.cfg -compile`
    - **Summary:** Reads output files from DELPHI calculations and compiles the results into a single binary file with all data required for optimizations and binding free energy calculations. Checks for normal completion of DELPHI jobs.
3. Perform optimizations.
    - **Program:** optimize
    - **Syntax:** `optimize optimize.cfg [options]`
    - **Summary:** Performs various optimizations and binding free energy calculations. The graphical user interface enables interactive modification

of constraints and optimization parameters, while the extensive command line options are well suited for scripting.

B.4.4 Optimization of protein ligands

When the ligand of an optimization procedure is a protein, it is generally beneficial to perform the optimizations on a per residue basis, and to evaluate the effects of different charge constraints on each residue. No additional continuum electrostatic calculations are required once the optimization matrices have been computed, but at least three optimizations per residue are generally performed.

1. Calculate matrices as outlined above.
2. Perform residue-by-residue optimizations.
    - **Script:** protein_scan.prl
    - **Syntax:** `protein_scan.prl crdfile [options]`
    - **Summary:** Performs a series of optimizations for each residue in the ligand, one optimization for every integral charge within a range. Output formats include text, LaTeX, and a MOLSCRIPT structural image.

B.4.5 Ligand scanning

The ligand scanning procedure requires all charge optimization matrices to be precomputed. Once this is done, initial charges are estimated on all ligands, and this charges are rapidly ranked. In the second stage, several calculations must be done: (1) quantum mechanical geometry optimization of the modified ligand; (2) calculation of the electrostatic potential of the ligand; (3) computation of ligand partial atomic charges, and subsequent use in re-ranking the list. The third stage also involves two steps: (1) fitting the quantum mechanically derived ligand geometry into the structure; (2) performing a single binding free energy calculation using the exact charges and geometry for the ligand, followed by another re-ranking of the ligand list.

1. Calculate matrices as outlined above.
2. Rank ligands based on estimated (rule-based) charges.
    - **Script:** ligand_scan.prl
    - **Syntax:** `ligand_scan.prl -l1`
    - **Summary:** Generates ligands with estimated charges, based of rules detailed in the "ligandscan.cfg" configuration file, and calculates binding free energies based on these. Outputs a ranked list of ligands.
3. Compute actual charges and rank ligands based on these charges.
    - **Script:** ligand_scan.prl
    - **Syntax:** `ligand_scan.prl -l2`
    - **Summary:** Computes partial charges on top ranking ligands (as scored in the first stage), performing all necessary quantum mechanical calculations, again as described in "ligandscan.cfg". If the charges have been computed, uses these charges to re-rank the ligand list.
4. Compute binding free energies based on actual charge and shape
    - **Script:** ligand_scan.prl
    - **Syntax:** `ligand_scan.prl -l3`
    - **Summary:** Sets up and runs a continuum electrostatic calculation for each top ranking ligand, using both the correct charge and shape as computed in the second stage. If these computations are complete, uses these energies in the ranking of the ligand list.

B.5 Manual Pages

B.5.1 binpot

NAME

binpot – Extract the electrostatic potentials at atom centers into a single binary file.

SYNOPSIS

binpot [options] crdfile

DESCRIPTION

binpot extracts the electrostatic potentials at atom centers calculated from a set of Poisson-Boltzmann calculations, and stores the results in a single binary file. Takes a required argument of **crdfile** which is the CHARMM coordinate file on which the calculations were run. The output potentials are in units of kcal/mol, NOT kT.

The output is a binary file containing DataMatrix: Each row is the difference potential for one offset.

The difference energy is determined by taking the difference potential times the charges from the frc files at these points and dividing by 2 and converting to kcal/mol.

binpot assumes that the directories are named 'base'n where n is an integer starting from 0 and increasing until there are no more existing directories. It assumes that the final file names are given by "final_base_namelow.frc", "final_base_namehigh.frc", and "final_base_namenamemidx.frc", with similarly named reference files.

OPTIONS

Numerous options control the general operation of the program, including naming of input and output files.

-b base [offset]
Base name of offset directories.

–final_name=final_base_name [bound_final0_0]
Base name to use for final state jobs. Will be ignored unless -noauto is also specified.

-fx n [0]
Use n for the job number of final state. Ignored if -noauto is also set.

-h, –help
Print a brief help message and exit. The **crdfile** argument is not required in this case only.

-noauto focus_levels
Do not attempt to automagically determine the final and reference state base names. Requires specification of the number of focus levels present.

-o out_file [potential.bin]
Set name of the output binary file to out_file.

-p in_file out_file [potential.bin potential.txt]
Read in binary file in_file and output a text summary as out_file.

–reference_name=reference_base_name [unboundreference1_0]
Base name to use for reference state jobs. Will be ignored unless -noauto is also specified.

-rx n [1]
Use n for the job number of reference state. Ignored if -noauto is also set.

-t out_file [potential.txt]
Output a summary of the results as text to out_file.

-v, -vv, –verbosity=n
Set the verbosity level to n. Possible levels are currently 0 to 2, with increasing levels resulting in increased output. -v is equivalent to –verbosity=1 and -vv is equivalent to –verbosity=2.

ADDITIONAL OPTIONS

The following options control the details of how the difference potentials are calculated. These options should be used with care, and only when the user is sure of what they want.

–final_only
Output only the final state potentials, rather than the difference (final - reference) potentials. While there maybe some uses for this, are you sure that this is what you want? Any grid potentials will NOT be eliminated.

–overfocus_mode=[overfocus_box standard_error] [overfocus_box]
Select the mode by which elements on the edge of an overfocussing box are determined. Possibilities are overfocus_box, which uses a file output by the PBE-solver PERL script, and standard_error, which uses an analysis of the standard error of the grid points. The first method is preferred, as it is unambiguous, independent of parameters, and can be applied even with only a single offset. The standard_error option exists primarily for backwards compatability as this was the method implemented in the initial versions of the software.

–overlap_only
Output the difference potentials only for atoms present in both the final and reference states. This could be useful if all you are interested is, for example, the desolvation potentials, and not the interaction potentials.

–reference_only
Output only the reference state potentials, rather than the difference (final - reference) potentials. Are you sure that this is what you want? This option exists mostly as a complement to the –final_only option and is likely even less usefull. Any grid potentials will NOT be eliminated.

AUTHOR

David F. Green <dfgreen@lms.mit.edu> and Erik Kangas.

BUGS

Please report bugs to the author at <dfgreen@lms.mit.edu>.

COPYRIGHT

Copyright Massachusetts Institute of Technology.

B.5.2 comp_anal.prl

NAME

comp_anal.prl – PERL script to set up and run calculations for component analysis

SYNOPSIS

comp_anal.prl crdfile [options]

DESCRIPTION

comp_anal.prl sets up, executes, and processes the output of all the continuum electrostatic calculations required for component. One calculation is done for every component in the system – typically three for every protein residue (side chain, amino, and carbonyl) and for every nucleic acid (base, ribose, phosphate). Some simple error checking is done, and jobs will be resubmitted for any calculation determined not to have finished. No checking is done for currently running jobs, so be careful about this.

The matrix element calculations each have the charges on the group in question charged to their wild-type values, and the difference in bound and unbound (or folded and unfolded) state potentials are computed. The bound and unbound shapes can be defined, as can the folded shape, but the unfolded shape is assumed to the the amino acid side chain free in solution.

OPTIONS

By default **comp_anal.prl** runs in job submission mode, if no previous component analysis is detected. The configuration file is read in, parsed, and the appropriate computations are configuread and submitted. A large number of jobs can be generated, so it is preferrable to have a batch queuing system in place to handle the multiple jobs. **comp_anal.prl** currently supports both NQS and PBS as the queuing protocol. If a previous component analysis is detected, **comp_anal.prl** will run by default in data compilation mode, reading the results of each component run and processing the results into text data files for analysis by **rank_comp.prl** The operation mode can be changed by the specification of the following flag.

run

Forces execution in job submission mode, checking for incomplete jobs and resubmitting these.

CONFIGURATION FILE

All configuration other than parameters for the continuum calculations, are done through a single configuration file, component.cfg. Parameters from the continuum calculations are set in a **delphi.prl** style param.file.

AUTHOR

David F. Green <dfgreen@lms.mit.edu> and Zachary S. Hendsch.

BUGS

Please report bugs to the author at <dfgreen@lms.mit.edu>.

COPYRIGHT

Copyright Massachusetts Institute of Technology.

B.5.3 delphi.prl

NAME

delphi.prl – PERL script to run DelPhi continuum electrostatic calculations.

SYNOPSIS

delphi.prl [options]

DESCRIPTION

delphi.prl sets up, executes, and processes the output of a set of continuum electrostatic calculations, using the DelPhi program.

OPTIONS

All parameter file options can be specified on the command line as keyword=value pairs. In addition, the following options may be specified on the command line:

-v
Run verbose mode.

-s
Run in silent mode.

paramfile=file_name [param.file]
Use file_name for the parameter file.

CONFIGURATION FILES

delphi.prl reads in several files, some required and some optional. The molecular structure and charges are read from a CRD file which is required. The atomic radii are similarly read in from a required radius file. The configuration of the calculation is done either through the command line, or through a parameter file which is almost always used, but not strictly required. The description of the calculations to do is detailed in a required run file. An optional definitions file provides a mechanism for aliases to be used in the run file.

Coordinate File (complex.crd)

This is a standard CHARMM format CRD file, with charges occupying the last column (the WMAIN array). A few things should be noted. First, all data in this file will be written to a PDB format file, so field sizes should be consistent with both. In particular, while chain identifiers of greater than a single character are valid in the CRD format, these will be truncated to a single character in the PDB format. So, in short, don't use chain identifiers longer than a single character. Another key point about chain identifiers is that "X" has a special significance as the dummy chain, and thus is always omitted from charge and shape selections. Again, in short, never use "X" for a chain identifier.

Radius File (radii.siz)

This is a DelPhi format radius file. The basic format of each line is: "aaaaaannnrrrrrrrr", where "aaaaaa" denotes a six character atom name field, "nnn" denotes a three character residue name field, and "rrrrrrrr" denotes an eight character radius field. The text string "aaaaaannnrrrrrrrr" often heads the file. Atoms match radii entries as follows: **(1)** If both the atom and residue names match a radius file entry, that radius is assigned; **(2)** If (1) failed to match, if the atom name matches a radius file entry with a blank residue name, that radius is assigned; **(3)** If (1) and (2) fail to match, if the first character of the atom name matches a radius file entry with a single character atom name and a blank residue name, that radius is assigned; **(4)** If (1), (2) and (3) all fail to match, a zero radius is assigned, and an error message will appear in the PB solver log, if the atom was not a hydrogen.

Parameter File (param.file)

The parameter file is used to specify global options for the calculations. Options required for the basic setup of the jobs and default options for all calculations are set here. All options are specified in keyword=value pairs, one per line. The allowed keywords are listed below, classified by the type of parameter that is set. All text from a "#" to the end of that line is ignored as a comment.

Calculation mode

calc = (all | setup | run | output) [all]

Calculation type: all means to do everything (setup, run, output); setup will just generate initial files (pdb, crg, param, radius); run will do no setup, just run jobs; output will just calculate output tables from previously run jobs.

Input files

crdfile = file_name [complex.crd]
Name of CRD file to use.

runfile = file_name [run.file]
Name of run file to use.

definitions = file_name [definitions.dat]
Name of definitions file to use.

rad_file = file_name [radii.size]
Name of radius file to use.

Continuum Configuration

innerdiel = x [4.0]
Internal dielectric constant.

outerdiel = x [80.0]
External (solvent) dielectric constant. If outerdiel is set to other than 80, you must also set the "temperature" variable to properly deal with salt. This is true both when the dielectric is being changed to account for water at different temperatures, and when the dielectric is being changed to that of a non-aqueous solvent.

salt = x [0.145]
Ionic strength (in M).

radprb = x [1.4]
Radius of solvent probe molecule for determining the molecular surface (in Angstroms).

stern = x [2.0]
Radius of ionic probe molecule for determining the ion exclusion layer (in Angstroms). If this is set to any number below 1e-6, it will be reset to 1e-6 to account for a DelPhi oddity which resets the Stern layer to a default value of 2.0 if this is set to 0.

temperature = (-1 | x) [-1]
Set temperature to x for the salt term of PB equation. The special option -1 uses the default value in the solver.

surface = (delphi | alternate | smooth | modsmooth | exact) [delphi]
How to compute molecular surface: delphi invokes the internal surface

generator within DelPhi; alternate specifies to use an externally computed surface, using the surface_gen keyword to provide futher details. smooth, modsmooth and exact are options only valid for the DelPhi v3.0 internal surfacer.

surface_gen = (chump) [chump]

External surface generation scheme to use. Currently the ChuMP surface is the only external surface supported.

Finite Difference Configuration

grid = n [65]

Number of grid lines on each side of cubic grid. This number must be odd, so that (0,0,0) falls on a grid point.

offset = x0,y0,z0 = x1,y1,z1 = ... [0.0,0.0,0.0]

Vectors by which the molecule will be offset relative to the grid. Multiple offsets can be specified, separated by "=". All calculations specified in the run file will be repeated for each offset, and the output energies will be averaged over all offsets.

focus = x0 x1 ... [23. 92.]

Set fill percentage for focussing calculations. These correspond to what percentage of one grid edge will the maximum dimension (x, y, or z) of the molecule occupy. For all calculations other than the lowest percent fill, boundary potentials are obtained from the next lowest percent fill calculation. For the lowest percent fill, the boundary keyword determines how boundary potentials are computed. Values of greater that 100 are valid, in which case an "over-focussing" procedure is used. However, at least one calculation at less that 100% fill is always required.

focus_split = (0 | 1) [0]

Flag to allow charged atoms to fall outside of an overfocussing box. Setting to 0 (false) causes the program to exit with a warning under these circumstances. Setting to 1 (true) causes the program to continue, either using sequentially lower focussing calculations for charges falling outside the box, or by generating extra calculations (see focus_cons).

focus_cons = (0 | 1) [0]

Flag for how to deal with charged atoms falling outside the box in overfocussed calculations. Setting to 0 (false) causes the previous focussing level to by used. This is done iteratively is multiple overfocussed levels are used. Setting to 1 (true) uses a more conservative method, doing a separate set of calculations for the portion of the molecule falling outside the box. This option has not been recently tested. This option is meaningless if focus_split is set to 0.

boundary = (1 | 2 | 4 | 5) [4]
Select type of boundary conditions to use for lowest focussing calculation. 1 selects zero potential boundary conditions. 2 selects the Coulomb dipole approximation. 4 selects Coulombic potentials. 5 selects a uniform electric field of 1 kt / e * grid unit in the "x" direction.

max_rad = x [3.0]
Number of Angstroms to add (subtract) from the maximum (minimum) x, y, and z coordinates to define dummy atom positions. This is used to consistently place the molecule on the grid for all calculations. Should be larger than the largest radius of any atom.

lit = ('a' | n) [a]
Number of linear iterations to perform for each finite difference calculation. An automatic convergence procedure is activated by the value "a".

nlit = n [0]
Number of non-linear iterations to perform for each finite difference calculation.

de = (-1 | x) [-1]
Change in total energy of finite difference grid at which convergence is considered to be reached. Depending on value of conab, this may be an absolute or a relative energy. The special option -1 uses the default value in the solver. Requires executable to contain enhanced convergence evaluation by LPL.

inter = (-1 | n) [-1]
Number of iteractions between convergence checks. The special option -1 uses the default value in the solver. Requires executable to contain enhanced convergence evaluation by LPL.

conab = (-1 | 0 | 1) [-1]
Type of convergence method. The special option -1 uses the default value in the solver. If set to 0 (false) uses relative energy based convergence. If set to 1 (true) switches on absolute energy based convergence. Requires executable to contain enhanced convergence evaluation by LPL.

energy = (G | S | C | AS | AG) [G]

Which energy terms to compute within the finite diffence solver. Multiple terms can be entered as comma separated values (with no white space). The terms are G (total grid energy), S (solvation energy), C (coulombic energy), AS (analytic surface solvation energy), and AG (analytic grid energy). For most applications, only the total grid energy is necessary, and other options are not thoroughly tested and may cause the solver to crash.

Input/Output Options

version = (delphi3.0 | delphi96) [delphi96]

Program version for automatically generated configuration files.

loadbd = (yes | no) [no]

Load an externally generated surface into the PB solver. This is automatically set to yes if ”surface=alternate” is specified, but may be used with ”surface=delphi” to load a previously computed surface (watch out of appopriate naming of the file if this is the case).

compression = (none | compress | gzip | bzip) [gzip]

Determines type of compression to use for data files.

keepphimap = (0 | 1) [0]

Flag to keep or remove potential maps after each run. If set to 0 (false) potential maps are deleted. If set to 1 (true) potential maps are saved.

adjoint = (0 | 1) [0]

Flag to output additional information required for the adjoint approximation software by AA. If set to 0 (false) the additional files are not output. If set to 1 (true) all required files are output.

Executable Options

delphi_exec = file_name [/programs/i386/bin/delphi]

Name of DelPhi executable to use.

delphi_exec_flags = executable_flags []

Flags to pass to DelPhi executable.

surface_gen_exec = file_name [/programs/i386/bin/chump]

Name of surface generation executable.

surface_gen_flags = executable_flags [-ignorefilewarning]

Flags to pass to surface generation program.

Directory and Naming Scheme

setup_dir = dir_name [**setup_files**]

Name of directory for input files to be stored in.

partial_inter_dir = dir_name [**partial_inter**]

Name of directory for energies from each calculation to be stored in.

output_dir = dir_name_root [**offset**]

Root of directory names for job output to be stored in. Full directory name is:

<root><offset>

param_root = file_name_root [][

Root of name for parameter files written to setup directory. Full name is:

<root><state><job>_<offset><focus>.prm

crg_file = file_name [**delphi.crg**]

Name of charge file created for input in setup directory.

all_pdb = file_name [**delphi.pdb**]

Name for all atom PDB file created for input in setup directory.

spec_pdb = file_name_root []

Root of name for charged and shaped atoms PDB file created for input in setup directory. Full name is:

<root><state><job>_<each>_<13/15>.pdb

Backwards Compatability

all_columns = (0 | 1) [**1**]

Flag for treatment of hydrophobic final and reference states (and thus the output table column contains only "null" and "0.000 (0.000)" results). If set to 0 (false) these columns are removed from the output tables (this was the behaviour in initial versions of the software). If set to 1 (true) these columns are still output.

all_rows = (0 | 1) [**1**]

Flag for treatment of hydrophobic output states (and thus the output table row contains only "null" and "0.000 (0.000)" results). If set to 0 (false) these rows are removed from the output tables (this was the behaviour in initial versions of the software). If set to 1 (true) these rows are still output.

perfill_low = x

perfill_mida = x

perfill_midb = x

perfill_high = x

Set fill percent for the focussing calculations using an older protocol. The mida options is only used with a three- or four-step focussing procedure, and the midb option only with a four-step focussing procedure. Using these options are exactly equivalent to giving the same fill percentages to the focus keyword. These options remain for backwards compatability only, the focus keyword should be used in place of these.

Run File (run.file)

The run file is used to specify the particulars of the calculations you want run. Atoms determining the shape, charges to be used, and groups to calculated final energies on are set here. Calculation specific options, such as changing the dielectric constant or ionic strength, can also be set here (all parameter file options are accepted). The run file is split into sections by ”mark=mark_name” keywords, with the keywords ”start”, ”output”, ”final”, ”reference”, and ”end” recognized. Multiple ”final” and ”reference” sections may be defined. All text from a ”#” to the end of that line is ignored as a comment. There are two primary types of entries specific to the run file, an atom selection and a selection name. These entries are interpreted as follows:

atom_selection

This should be a PERL syntax logical string, with all atoms for which the string evaluates as TRUE added to the group. The keywords recognized are ”atomno”, ”resno”, ”resid”, ”resname”, ”atomtype”, ”segid”, ”xcoor”, ”ycoor”, ”zcoor”, and ”charge”.

selection_name

This should be a string describing the name of the corresponding atom selection. The keywords ”atomno”, ”resno”, ”resid”, ”resname”, ”atomtype”, and ”segid” will be expanded, although the behaviour if the various atoms in the atom selection do not give the same expansion may not be that desired.

mark=start

This defines the beginning of the run file. All commands before this point will be ignored, with a warning given.

mark=output

This defines groups of atoms at which to compute output energies. Each output group is multiplied by the potential of all "final" and "reference" calculations for which the output group is present in the shape definition. Any number of groups may be defined here, in paired lines of "atoms_charged" and "name". If no output groups are defined, a single group of all atoms (named DEFAULT) will be used.

atoms_charged = atom_selection

Define the group of atoms in the output group. One additional keyword is allowed in this section, the "each" specification. The "each" keyword must be the first entry in the selection, followed by "atom", "residue", or "chain". This "each type" combination will be expanded, replacing the single output group with a group for every atom, every residue, or every chain, with the rest of the selection left unchanged. When using the "each" keyword, remember to make sure you include "atomno", "resno" or "segid" in your "name" entry (depending on the each mode), so that each output group gets a unique name.

name = selection_name

The name of the output group.

mark = (final | reference)

This defines the selection of atoms (charged and shape) for a "final" or "reference" state computation. Any number of final and reference states may be entered, with each new section defined by a "mark=...". Also any variations in parameters desired for the calculation are entered in this section, using the same syntax as in the parameter file. The only difference between "final" and "reference" calculations is in the final processing of the energies. Each "final" state will have energies output to "final.table", and each "reference" state will have energies output to "reference.table". In addition, a "difference.table" will be output, with the energies of each "final" state after subtraction of the energies of all "reference" states with any overlap of charged atoms with the "final" state in question. This is done for the energy of each output group. NOTE: There is no checking done for double counting of energies in the "final - reference" calculation. The onus is on the user to ensure that the appropriate states are defined.

name_group = selection_name
Name of the group. Keyword expansion done based of ”atoms_shape” atom_selection.

atoms_charged = atom_selection
Define atoms to be charged in calculation of potential.

atoms_shape = atom_selection
Define atoms used in determination of the internal dielectric region.

atoms_center = atom_selection
Define atoms used to center the overfocussing box. This option is ignored for all calculations below 100% fill.

mark=end
This defines the end of the run file. All commands after this point will be ignored, with a warning given.

Definitions File (definitions.dat)

The definitions file provides a mechanism for defining aliases for use in the run file. The format is ”alias=definition”, one per line. All occurrences of the alias (as a bare word, separated on both sides by white space) will be substituted. Alias definitions may contain other aliases, and will be expanded appropriately.

AUTHOR

David F. Green <dfgreen@lms.mit.edu> and Zachary S. Hendsch.

BUGS

Please report bugs to the author at <dfgreen@lms.mit.edu>.

COPYRIGHT

Copyright Massachusetts Institute of Technology.

B.5.4 getmatrix

NAME

getmatrix

SYNOPSIS

getmatrix [configuration_file]

DESCRIPTION

getmatrix combines the results of a number of Poisson-Boltzmann calculations into a set of matrices for use in electrostatic optimization. Each set of calculations must have previously been processed into a single binary file, using the **binpot** program. A configuration file as described below is required.

For most purposes, the configuration file will be generated, and **getmatrix** will be run, by the **matrix_elements.prl** script.

OPTIONS

configuration_file

Name of configuration file to use. Defaults to "getmatrix.cfg".

CONFIGURATION FILE

The syntax of the configuration file must be exact, or errors may result. These errors may be detected, but also may not be. Therefore, it is strongly suggested that the configuration file is not editted by hand unless the user is experienced and confident of the syntax. It is much better to allow the configuration file to be generated by the **matrix_elements.prl** script.

The configuration file must conform exactly to the following format, with each entry falling on it's own line. There are several things to be aware of: (1) Some entries are only read if certain options on previous lines are specified; including these lines without the appopriate options will cause errors. (2) Some sets of entries are repeated over a integer specified on a previous line; these groups are designated below, and the entire set should be entered for one state before entering the set for the next.

CHARMM CRD file (string)

Name of CRD file describing the molecule of interest.

Data Directory (string)

Name of directory containing data from PB calculations.

Component Type (string)

Molecule on which components were calculated. May be none, ligand, or receptor.

Component Directory (string)
Directory containing data for component calculations. This only matters if component type is set to ligand or receptor, but the entry must always be present.

Matrix File (string)
Name of the binary file that will be output.

Verbose Flag (integer)
Set the verbosity level. Current possibilities are 0, 1 or 2.

Receptor Complete Flag (string)
Flag describing whether the receptor calculations are complete. Options are true, partial or false. True uses the receptor calculations for the interaction vector as well as the receptor desolvation energies. Partial uses the receptor calculations for the receptor desolvation energies, but uses the ligand atom calculations for the interaction vector. False ignores receptor calculations completely, setting the receptor desolvation energy to 0, and uses the ligand atom calculations for the interaction vector.

Inner Dielectric (float)
Value of the internal dielectric constant that the PB calculations were done at. This is required for the calculation of Coulombic potentials to be consistent with the solvation potentials.

Multi-conformation Flag (integer)
Currently must be set to 0 (off).

Number of Complex Conformations (integer)
ONLY READ IF MULTI-CONFORMATON FLAG SET

Number of Receptors (integer)

—REPEAT FOR EACH RECEPTOR—

Number of Conformations for Receptor (integer)
ONLY READ IF MULTI-CONFORMATON FLAG SET

Chains defining Receptor (space delimited list of SEGIDs)

———END REPEAT———

Number of Ligands (integer)

—REPEAT FOR EACH LIGAND—

Number of Conformations for Ligand (integer)
ONLY READ IF MULTI-CONFORMATON FLAG SET

———END REPEAT———

—REPEAT FOR EACH LIGAND—

Residual Flag (integer)
Flag specifying whether a residual component to the ligand exists. May be 0 or 1.

Chains defining Ligand (space delimited list of SEGIDs)

Chains defining Ligand Surface (space delimited list of SEGIDs)

Elements of ligand calculated (space delimited list of ATOMNOs)

———END REPEAT———

—REPEAT FOR EACH RECEPTOR—

Elements of receptor calculated (space delimited list of ATOMNOs)
ONLY READ IF COMPONENT TYPE SET TO RECEPTOR

———END REPEAT———

AUTHOR

David F. Green <dfgreen@lms.mit.edu> and Erik Kangas.

BUGS

Please report bugs to the author at <dfgreen@lms.mit.edu>.

COPYRIGHT

Copyright Massachusetts Institute of Technology.

B.5.5 ligand_scan.prl

NAME

ligand_scan.prl – PERL script to perform ligand scanning procedure

SYNOPSIS

ligand_scan.prl [options]

DESCRIPTION

ligand_scan.prl uses the matrices generated from **matrix_elements.prl** to generate and analyzed a database of ligand derivatives to an arbitrary level of detail. The standard levels of operation are: (1) estimated charges on an approximate shape; (2) exact charges on an approximate shape; (3) exact charges and shape. Currently only electrostatic components of the binding free energy are considered. The script will set up and run all necessary computations for every step.

OPTIONS

The mode in which **ligand_scan.prl** must be specified on the command line. Several additional options can also be specified.

-s

Forces computation of single mutations only, overriding the setting in the configuration file.

-l0,l1,-l2,-l3

Perform scanning at level 0 (setup), 1 (estimated charges), 2 (exact chag, or 3. (0) Do setup only. (1) Generate database of derivatives with estimated charges, and rank. (2) Set up and submit QM calculations for top ligands from stage 1. Fit charges to ESP from these calculations and rerank database. (3) Set up and submit binding free energy calculations using the QM shape and charges, then rerank database.

-m

Forces computation of multiple mutations, overriding the setting in the configuration file.

-db db_dir [default scan_data]
Use db_dir to store the database.

-n n_ligands [default 10]
Set the number of ligands to submit to the next level of computation.

CONFIGURATION FILE

All configuration is done through a single configuration file, ligandscan.cfg.

AUTHOR

David F. Green <dfgreen@lms.mit.edu>.

BUGS

Please report bugs to the author at <dfgreen@lms.mit.edu>.

COPYRIGHT

Copyright Massachusetts Institute of Technology.

B.5.6 matrix_elements.prl

NAME

matrix_elements.prl – PERL script to set up and run calculations for electrostatic optimization.

SYNOPSIS

matrix_elements.prl config_file [options]

DESCRIPTION

matrix_elements.prl sets up, executes, and processes the output of all the continuum electrostatic calculations required for electrostatic optimization. One calculation is done for every ligand basis point, as well as a single calculation on all ligand atoms not included as basis points for optimization. In addition, a calculation is done for the bound and unbound states of the receptor. Some simple error checking is done, and jobs will be resubmitted for any calculation determined not to have finished. No checking is done for currently running jobs, so be careful about this. Options in the configuration file can be set to compute receptor elements for use in Type-II optimization. Bound state components can also be computed.

The matrix element calculations each have a single atom charged to +1e, and the difference in bound and unbound state potentials are computed. The bound shape consists of ligand, receptor, and surface segments, and the unbound shape consists of ligand and surface segments. The residual calculation has all non-selected ligand atoms charged to wild-type values, again with the difference in bound and unbound state potentials computed. The bound and unbound receptor calculations have all receptor atoms charged to wild-type values, with the shape either the complex (ligand, receptor, and surface) or the receptor alone. The solvation potentials are computed for each state. The bound state component calculations each have a single atom charged to +1, and the solvation potentials in the bound state are computed.

OPTIONS

By default **matrix_elements.prl** runs in job submission mode. The configuration file is read in, parsed, and the appropriate computations are configuread and submitted. A large number of jobs can be generated, so it is preferrable to have a batch queuing system in place to handle the multiple jobs. **matrix_elements.prl** currently supports both NQS and PBS as the queuing protocol. The operation mode can be changed by the specification of one of the following flags.

-compile

Run script in matrix compilation mode, skipping the check for completion of calculations. This will speed up execution when jobs are known to be complete, but will NOT submit any incomplete jobs, and will bomb with the first incomplete job encountered.

cpu1 cpu2 ...

This is not a flag, but rather a list of computers to which the jobs should be submitted. Jobs are submitted to all machines in the list in a cyclic manner. The default action is to use the local machine.

-help

Print out a brief help message and exit. This option does not require specification of a configuration file.

-print

Print out current configuration as read in from configuration file then exit.

-test

Run script in test mode, doing everything except running the DePhi calculations.

CONFIGURATION FILE

The definition of receptor and ligand, the selection of atoms for which to calculate desolvation elements, as well as all other configuration, including parameters for the continuum calculations, are done through a single configuration file.

AUTHOR

David F. Green <dfgreen@lms.mit.edu> and Erik Kangas.

BUGS

Please report bugs to the author at <dfgreen@lms.mit.edu>.

COPYRIGHT

Copyright Massachusetts Institute of Technology.

B.5.7 optimize

NAME

optimize – optimizes electrostatic charge distributions for ligand-receptor binding

SYNOPSIS

optimize -h
optimize config-file [options]

DESCRIPTION

optimize calculates optimal electrostatic charge distributions for ligand - receptor binding. Affinity and specificity optimization subject to constraints and restraints. Data analysis features.

General sqpecificity and single-decoy specificity optimizations of ligand-charge distribution. It allows the application of many types of constraints and restraints, as well as the use of any number of receptors. It has many built-in data analysis and components analysis features.

The command-line options are parsed and executed from left to right. These commands can be combined in any order and repeatedly to perform complicated functions.

OPTIONS

-a12

Perform Type I and Type II potential analysis on the current charge distribution. Requires -readR file for compute typeII

-BH file

Compute the "best hapten": Produces a ligand charge distribution. If the Receptor were TypeI optimized for this charge distribution, it would be TypeII optimized for the current wild-type charge distribution. The best hapten charges are stored in Qcur. Requires -readR file for compute typeII

-BindBin base

This writes out 4 binary files as would be produced by the 'binpot' program, containing the potentials at the relavent points at the complex atom centers. The 'current' ligand charge distribution is used. 'base-ri' contains the interaction potential of the ligand atoms at the receptor and surface sites; 'base-rd' contains the receptor desolvation potential at the receptor and surface sites; 'base-li' contains the interaction potential at the ligand and surface points; 'base-ld' contains the ligand desolvation potential at the ligand and surface points. In order to produce these files, all components must have been computed. Also, in order to compute the ligand files, all ligand atom centers must be variable (i.e. no residual). This is useful to obtain the potentials from a binding calculation for any set of ligand charges, without performing new binding calculations. These potentials may be input into analysis porgrams to compare specificity and similarity.

-comp?? [n]

Performs a component analysis on the ligand–target interaction energy. Uses the 'current' set of ligand charges for computing the energies, so you may alter the 'current' set to fine-tune the component analysis. The two '?' can be any one of 'A', 'S' or 'B', standing for 'All', 'Sidechain', or 'Backbone'. This specified which atoms on each ligand and receptor, respectively, residue to include in the component analysis. The program then takes all pairs of resides and computes the interaction energy of the designated parts. I.e. -compAS computes the interactions of each receptor sidechain with all atoms from each ligand residue. The [n] option causes the output to be sorted by interaction energy and only the top 'n' interacting pairs are printed (if n $<$ 0 all pairs are sorted and printed). Use of '-comp' by itself is the same as '-compAA'. Component analysis divides the ligand itself into 'groups' instead of just residues. If components are not computed, this function computes the interactions of the ligand groups with the entire receptor.

-compl

The same as '-comp', except that the information for every interaction pair is printed on a separate line. This is particularly useful for exporting data to tables. The 'l' stands for 'list'.

-comp-des

Print components of the desolvation penalty for the current charge distribution in a matrix format, where the elements are the desolvation elements for the individual groups of ligand charges.

-Cur=Ref

Copies the Reference charge set into the Current charge set.

-Cur=Wt

Copies the wild-type charge set into the Current charge set.

-decoy ix

Sets the decoy receptor to receptor 'ix', counting from '1'.

-eval

Display all eigenvalues for a Affinity-optimization (Type 1) for the specified target receptor. Also works with a 1-decoy specificity optimization for the specified target, decoy, secondary target and lambda value. For each eigenvalue, displays the SVD ratio with the largest eigenvalue (for use with the SVD cutoff), the fractional error value (for use with the error cutoff), the maximum contribution to the free energy and the charge coefficient at optimum.

-eval2

Display all eigenvalues for a General Specificity-optimization (Type 2) for the specified target receptor. For each eigenvalue, displays the SVD ratio with the largest eigenvalue (for use with the SVD cutoff), and the fractional error value (for use with the error cutoff). Requires -readR file for compute typeII

-evalB

Save as '-eval', but displays the projection of the interaction potential instear of the charge coefficients.

-evec

Display all eigenvalues and corresponding eigenvectors for a Affinity-optimization (Type 1) for the specified target receptor. Also works with a 1-decoy specificity optimization for the specified target, decoy, secondary target and lambda value.

-evec2

Display all eigenvalues and corresponding eigenvectors for a General Specificity-optimization (Type 2) for the specified target receptor. Requires -readR file for compute typeII

-gui

Enable the graphical user interface. Only '-verbose' and '-help' work in conjunction with this command.

-h, -help

Display this help information.

-info

Display charge, binding and specificity information for each of the reference, wild-type and current charge distributions with respect to each receptor, with the target emphasized.

-o1

Perform a type-1 optimization. This corresponds to a target affinity optimization (if lambda=0) or a single decoy specificity optimization with optional secondary target receptor id lambda != 0. Calls '-info' when complete and stores the optimized charges in the 'current' charge list.

-o1d

Perform N type-1 optimizations (ignoring lambda), one for each basis point, holding all others at wild-type. For each optimization, print out the diagonal desolvation matrix element, the interaction element, the optimized charge, the absolute deviation from the wild-type charge and the gain in affinity for this single point mutation to optimal.

-o1table

Single decoy specificity optimization ramping lambda between the values specified (in the configuration file) with the specified step size. The results are displayed in a table of data at each lambda step.

-o2

Performs a type-2 general specificity optimization for the target receptor. You must have components computed for this option to work. Calls '-info' when complete and stores the resulting charges in the 'current' charge set. Requires -readR file for compute typeII

-parms

Displays the current list of parameters.

-readq file [ref|cur|wt]

Reads the variable ligand charges from an external file. Stores the charges in the 'reference' charge list. The file must be a list of charges as real numbers, like the output of a RESP fit. There must be one charge number for each charge selected for optimization. The function looks at the file suffix. If it finds a '.crd', '.CRD', '.pdb', or '.PDB', is will read the file in using the appropriate format and extract the charge information contained within. All other information in these files will be ignored. Specification of 'ref', 'cur' and 'wt' allow you to read the charges into either of the reference, current or wild-type charge distribution lists.

-readqr file

Reads in a vector of charges or potentials into an internal receptor charge list. This list has the same number of elements as the first receptor has total atom centers. This is designed for reading in the data that was exported by '-RefDesP -writeq file ref' for obtaining the desolvation potential of the 'receptor' in preparation for a type-II analysis.

-readR file

Reads the matrix 'R' matrix from a file into a special interior variable for type II analysis.

-Ref=Cur

Sets the 'reference' charge set equal to the 'current' charge set.

-Ref=DesP

Copies the desolvation potential of the current charges into the reference charge list. Qref = 2 * L * Qcur.

-Ref=Wt

Sets the 'reference' charge set equal to the 'wild type' charge set.

-rev

Prints out the revision history of this program.

-rotdip

Display dipoles in rotated coordinates. The coordinates of each group of atoms are centered on the geometric center of the atoms and the axes are aligned with the principle moments of geometric inertia. The x, y and z axes correspond to the largest, middle and smallest moments, respectively. Use of this flag will allow, for example, all residues of the same conformation of have comparable dipole moments, independent of the position and orientation of the residue in the molecule as a while. The dipole momement of the molecule as a whole will also be rotated in a similar manner.

-silent

Revoves all non-essential output, i.e. no 'progress' info will be displayed. (verbose level -1) This parameter overrides any verbose level specified in the configuration file.

-simtype n

Determines the method used to compare the similarity to two electrostatic potentials (A,B), with N basis points, R = A+B (Note that we typically want A+B=0 so we design A = -B)

n=1: Root Square Deviation [0*,infinity]

n=2: Absolute Deviation [0*,infinity]

n=3: Normalized Absolute Deviation [0*,1]

n=4: Cosine [-1*,1]

n=5: Normalized Root Square Deviation [0*,sqrt(2)]

n=6: Normalized Square Deviation [0*,2]

n=7: Relative Magnitude [0,1*]

Option (6) is the default. The (*) indicates the desired result for good similarity. Note that NSD = 1 + MAG * COS

-starget ix

Sets the secondary target index (counting from 1). This is used in 1-decoy optimizations with the objective function

F = dG(target) - lambda [dG(decoy) - dG(starget)]

usually starget = target.

-target ix

Sets the index of the target receptor, counting from 1. If there is only 1 receptor, this must always be 1.

-verbose

Turn on verbose mode (to level 1). This can be set to higher values in the configuration file. This parameter overrides any verbose level specified in the configuration file.

-writeq file [ref|cur|wt]

Writes the variable ligand charges to an external file. Saves the charges in the 'reference' charge list. The file will be a list of charges as real numbers, each on a separate line. There will be one charge number for each charge selected for optimization. The function looks at the file suffix. If it finds a '.crd', '.CRD', '.pdb', or '.PDB', is will write the file in using the appropriate format. Otherwise, it will use the list format described above. Specification of 'ref', 'cur' and 'wt' allow you to save the charges from either of the reference, current or wild-type charge distribution lists.

-writeL file

Writes the binary matrix L to a file for later reading by '-readR'. This matrix is necessary in type II potential analysis.

AUTHOR

David F. Green <dfgreen@lms.mit.edu> and Erik Kangas.

BUGS

Please report bugs to the author at <dfgreen@lms.mit.edu>.

COPYRIGHT

Copyright Massachusetts Institute of Technology.

B.5.8 protein_scan.prl

NAME

protein_scan.prl – PERL script to perform residue-by-residue optimizations on a protein ligand

SYNOPSIS

protein_scan.prl crdfile [options]

DESCRIPTION

protein_scan.prl uses the matrices generated from **matrix_elements.prl** to perform a series of optimizations on a protein ligand. For every residue of the ligand, an optimization is done constraining the charge to -1, 0 and +1 e. The results can be output in a variety of ways.

OPTIONS

By default **rank_comp.prl** will output a summary of the results by segid. All other output options are specified on the command line. Multiple output options can be given, and all will be performed.

-calc,-calculate
Force re-calculation of the optmization results. This will overwrite any previous results.

-output,-nocalc
Do not do optimization calculations, but rather use previous results. Exit if no previous results have been computed.

-text,-notext
Toggle text output mode.

-latex
Output results in LaTeX table format.

-molscript
Generate a MolScript figure with variable sized spheres representing the degree of improvement on optimization.

-o output_root
Set root of output files to output_root.

-s segid
Do calculations on resdiues of chain segid.

-rx,-ry,-rz rotation
Rotate molecule by rotation in x, y or z before outputing MolScript figure. Only has meaning if -molscript flag set.

-lime max_energy
Limit the maximal sphere size to that of max_energy. Only has meaning if -molscript flag set.

-interactive,-lowres,-midres,-highres
Set mode of MolScript generation, to interactive, a low resolution static figure, a mid resolution static figure, or a high resolution static figure.

-wtref,neutref
Set reference state to wild type or a hydrophobic isostere.

AUTHOR

David F. Green <dfgreen@lms.mit.edu>.

BUGS

Please report bugs to the author at <dfgreen@lms.mit.edu>.

COPYRIGHT

Copyright Massachusetts Institute of Technology.

B.5.9 rank_comp.prl

NAME

rank_comp.prl – PERL script to analyze the results of a component analysis

SYNOPSIS

rank_comp.prl [options]

DESCRIPTION

rank_comp.prl reads the output of component analysis computations are done with **comp_anal.prl** and analyzes the results in multiple ways. Multiple options for sorting the results are given, as are means to output the details of individual interactions.

OPTIONS

By default **rank_comp.prl** will output a summary of the results by segid. All other output options are specified on the command line. Multiple output options can be given, and all will be performed.

-x [cutoff] [default 0.5]
Display records with components greater than cutoff.

-d,-c,-m [n_records] [default 10]
Display n_records records sorted by desolvation (-d), contribution (-c) or mutation (-m).

-dd,-cc,-mm [n_records cutoff] [default 10 0.1]
Display details of n_records records sorted by desolvation (-d), contribution (-c) or mutation (-m). Level of detail set by cutoff.

-ee [cutoff] [default 0.5]
Display individual interactions whose value is greater than cutoff.

-s select_segid
Display results only for components belonging to select_segid.

-o outputfile
Output results to outputfile rather than to the standard output.

-h
Display help information.

AUTHOR

David F. Green <dfgreen@lms.mit.edu>.

BUGS

Please report bugs to the author at <dfgreen@lms.mit.edu>.

COPYRIGHT

Copyright Massachusetts Institute of Technology.

B.6 Sample Configuration Files

All configuration files conform to the same basic format. Parameter specifications are all in the format 'keyword = value', and with extraneous white space ignored. All lines beginning with '`#`' are ignored as comments.

B.6.1 comp_anal.prl (component.cfg)

```
##
## Sample Component Analysis Configuration File
##

## Definition of component analysis type
##
## 'type'       [ binding | stability ]
##               Type of component analysis to do - contribution to
##               binding, or contribution to stability
## 'final'      Definition of final states.  Takes a conma separated
##               list of SEGIDs. For stability analysis, multiple
##               final states may be defined, with each selection
##               treated separately.
## 'reference' Definition of reference states. Takes a comma
##               separated list of SEGIDs.  This only has meaning
##               for analysis of binding.  Multiple reference states
##               may be defined, one for each component of the
##               binding reaction.
type      = binding
final     = ["A","B","C","D"]
reference = ["A","B","C"]
reference = ["D"]

## Batch queue submission parameters
##
## 'subdel'     Command for submission of jobs to batch queue
## 'pause'      Number of seconds to pause between submitting
##               each job to the batch queue.  This prevents
##               locking of the queue in some cases
subdel = /programs/common/bin/subdelphi
pause  = 2
```

B.6.2 matrix_elements.prl (matrix.cfg)

```
##
## Sample Matrix Element Configuration File
##
## The commands can appear in any order, except 'select' which must be
##    the last command.
##
```

```
## General Parameters
##
## 'crd_file'    [Required], name/location of the .crd file to use.
## 'delphi'      [default = /programs/common/bin/delphi.prl]
##                location of the 'delphi' script to use.

crd_file   = complex.crd

## Continuum Electrostatic Calculation Setup Parameters
##
## 'radii.siz' [default = /usr/people/dfgreen/param/delphi/radii.siz]
##      location of the radii.siz file to use
## 'grid'
##      grid spacing to use.
## 'focus'
##      list of focus setps to use.
## 'atoms_center'
##      center for focusing.  There is an additional option
##      'atoms_center=charged' will cause the center of the focussing
##      to be the center of all charged atoms in each run. This will
##      cause the desolvation matrix elements also to be centered on
##      each atom center.
## 'focus_split'
##      focus_split value
## 'innerdiel'
##      protein/molecular dielectric constant
## 'outerdiel'
##      solent dielectric constant
## 'salt'
##      salt concentration
## 'delphi_exec'
##      delphi execputable program.
##
## The 'residual' parameters apply to calculation of the residual
##     Ligand desolvation, and the bound and unbound receptor solvation
##     components only.  Since they may involve many points being
##     charged, the computation parameters may be different.
##
## 'residual_focus'
##     focus levels for residual calculations.
##     defaults to the value of 'focus' if unspecified
## 'residual_grid'
##     grid spacing for residual calculations.
##     defaults to the value of 'grid' if unspecified
## 'residual_atoms_center'
##     atom center for residual calculations.
##     defaults to the value of 'atoms_center' if unspecified
## 'residual_focus_split'
##     focus split for residual calculations.
##     defaults to the value of 'focus_split' if unspecified

radii.siz             = radii.siz
grid                  = 129
focus                 = 23. 92. 184.
atoms_center          = charged
inner_diel            = 4
outer_diel            = 80
residual_grid         = 191
residual_focus_split  = 1
residual_atoms_center = all

## Ligand, Receptor and Surface Definitions
##
## Use Perl list-reference notation to list the segment names.
##
## num_lig [default = 1]
##     The number of ligands you wish to optimize
## lig_segs
##     Segments defining each ligand.  Separate multiple ligands
##     with an entry of "::".
## rec_segs
##     Segments defining the receptor.
## lsrf_segs [default = empty ]
##     Segments defining each ligand surface.
##     Separate multiple ligand surfaces with an entry of "::".
## rsrf_segs [default = empty ]
##     Segments defining receptor surface.
## 'srf_segs'
##     is the same as 'lsrf_segs' which stands for the ligand
##     surface segments.  You may also specify receptor surface
##     segments (i.e. for Type II calculations) using 'rsrf_segs'

num_lig   = 1
rec_segs  = ["A","B"]
lig_segs  = ["C"]
lsrf_segs = []
rsrf_segs = []

## Special Flags and Parameters
##
```

```
## delete [no|yes|full] [default = yes]
##     Deletes extra files in the offset directories to
##     conserve disk space.  'no' deletes none.  'yes' deletes
##     'setup_files/', 'partial_inter/', 'ARCDAT', and all the .phi files.
##     'delete=full' option deletes all files except 'difference.table'
##     and 'potential.bin'.  Do not use this unless you are really low on
##     disk space, because the potential.bin file cannot be regenerated.
## queue [default = pipe]
##     If non-zero it submits each separate delphi job to
##     the local queue (there can be hundreds for a large ligand!).  When
##     all jobs finish, the data will be computed.  If queue=no, each
##     job will be run sequentially in the 'foreground'.
##     'pipe' will queue the job, but cause the job to be run locally
##     in the /tmp/... directory of the local machine.  It handles
##     copying the information to and from the /tmp directory and
##     cleaning up after itself.
## verbose [0 | 1] [default = 1]
##     Prints more information.
## directory [default='data'] subdirectory to store all matrix elements
##     and residual matrix computations.
## component_dir [default='data_C']
##     subdirectory to store bound-state
##     solvation components for doing a component analysis or computing
##     type-II ligands.  If you compute these, this directory MUST be
##     different than that of 'directory' because some of the
##     subdirectories may have the same names (the atomno).
## components [none|ligand|receptor] [default=none]
##     calculates the bound-state
##     solvation for each ligand or receptor atom center and puts the
##     results in subdirectories of 'component_dir'.   Since
##     mathematically it doesn't matter which ones you compute (you get
##     the same interaction matrix out), choose whichever of the two
##     has the fewer atom centers.  However, you must compute all
##     selected ligand basis points if you choose "ligand", if you choose
##     receptor, it will be possible to compute a subset of receptor pts.
## verbose [0..n] [default=0]
##     Non-zero values increase the amount of detail output to
##     the screen.
## rec_desolv [required | optional | off ]
##     Sets whether receptor desolvation
##     calculations will be performed.  With optional setting, jobs will
##     be submitted, but optimizations can be done before completion of
##     these jobs.
## matrix_file [default=matrix.bin]
##     Name of the binary results file
```

```
delete        = yes
queue         = pipe
directory     = data
component_dir = data_C
components    = ligand
verbose       = 0
rec_desolv    = required
matrix_file = matrix.bin

## Definition of ligand and receptor atom as basis points.
##
## All selection lines can take
## conditionals which determine the atoms which will be the ligand
## basis points.  Think of the lines 'OR'ed conditionals.  You can use
## the following keywords to define your selection, together with
## standard Perl conditional notation.
##
## atomno    : Atom number 1...n from .crd
## resno     : Residue number 1...m from .crd (first column)
## resid     : Text residue type
## atomtype  : Text Atom type
## xcoor     : X coordinate
## ycoor     : Y coordinate
## zcoor     : Z coordinate
## segid     : Segment ID text
## charge    : Atomic partial charged (from wmain)
## absres    : Residue number (second column), can be a textual field
##             if you have A,B,C... in the residue 'numbers' here.
##
## The 'select_ligand'/'select_receptor' block must end with
##    a keyword 'end' on a line by itself.  For multiple ligands,
##    separate each ligand's selection string with an entry of "::" on
##    an line by itself.
##
##  'select_receptor' is ignored if 'components != receptor')

select_ligand
  segid eq "C"
end

select_receptor
  (segid eq "A" || segid eq "B") && atomtype ne "CA"
end
```

B.6.3 optimize (optimize.cfg)

```
##
## Sample Optimize Configuration File
##
## The commands can appear in any order, except 'receptors' which must
##    be the first command.

## SPECIFICATION OF LIGAND-RECEPTOR PAIRS
##
## The format is keyword 'receptors' on a line by itself followed by
##    a paired list of
##    'filename name [density] [multlig symmetrization]'
##    for each ligand--receptor
##    complex.  The keyword 'end' on a line by itself terminates the list.
##
## The 'filename' parameters refer to binary matrix files created by
##    the 'compile_matrix.prl' script.  This parameter is the location of
##    the appropriate file.
## The 'file' parameter is just an internal name that will be used to
##    refer to this receptor in the output.
## The optional 'density' parameter represents the average number of A^2
##    covered by each surface point.  This is used in surface potential
##    integrations and has a default value of 1.0.
## The optional "multlig symmetrization" parameter specifies how to deal
##    with multiple ligands.
##
## NOTE: It is assumed that the ligand will be the same in every complex.
##    i.e. the same number of atom centers computed with the same
##    atomtypes, resids and segids.  The conformations, however, do not
##    have to be the same.  It is also assumed that the ligand-surface is
##    similarly conserved

receptors
matrix.bin ComplexName 1.000
end

## LAGRANGIAN OPTIMIZATION PARAMETERS
##
## 'target' Target receptor for affinity/specificity optimization.
##      [default=1], Valid numbers 1..Number of receptors
## 'decoy' Decoy receptor for specificity optimization [default=1]
## 'second_target' Target for specificity optimization which can be
##      different than target for affinity optimization [default=1]
## 'lambda' is the 1-decoy specificity ramping parameter [default=0.0]
##      valid values are lambda \in [0,1]
##
## Type I Optimization minimizes the Lagrangian
##
## (1-lambda) * dG(target) - lambda * ( dG(decoy) - dG(second_target) )
##
## where dG(x) is the electrostatic binding free energy to receptor 'x'.
##

target        = 1
decoy         = 1
second_target = 1
lambda        = 0.0

## MATRIX INVERSION PARAMETERS
##
## 'svd_cut' [default=1e-5]
##      Specifies the minimum ratio of eigen_value/max_eigenvalue
##      to include in the calculation.  All eigenvectors with negetive
##      eigenvalues are always excluded.
## 'err_cut' [default=0.25]
##      Specified the relative maximum error in the eigenvalue
##      allowed.  This value is the standard_deviation / value of
##      the eigenvalue.  I.e. if 'err_cut' is '0.5', then if the
##      eigenvalue is uncertain to more than 50% then the respective
##      eigenvector will be excluded.
## 'null_weight' [default=999.0]
##      Allows you to include the eigenvectors that were excluded
##      by 'svd_cut' and 'err_cut' in order to satisfy constraints better.
##      Each is included with a quadratic penalty of 'null_weight' for
##      being charged and no interaction terms (so there is no favorable
##      contribution from them in the actual optimization process. )
##      A value of '0' means that there is no penalty for charging them
##      and the will be used freely.  A large value (>5) puts a steep
##      penalty on using them and they will be used only when absolutely
##      necessary (i.e. when it would otherwise be impossible to satisfy
##      all constraints).  A value of '999.0' means DO NOT include these
##      vectors for any reason.

svd_cut       = 1e-5
err_cut       = 0.25
null_weight   = 999.0

## OPTIMIZATION CONSTRAINTS
##
```

```
## 'QrConstrain' [yes|no] [default=yes]
##      If yes, then constraints will be applied to residues.
##      If no, IntConstrain and Max|Qres| will be ignored.
## 'IntConstrain' [yes|no] [default=yes] If yes, then the total charge on
##      each residue will be constrained to be an integer.  This will be
##      done exactly or approximately depending on the value of
##      'MaxResIntOpt' and the number of residues being optimized.
## 'MaxResIntOpt' [default=3]
##      This gives the maximum number of residues for which to try
##      ** ALL ** combinations of integer charge constraints.
##      I.e. if 'MaxResIntOpt=3' and residues are constrained to be
##      integers between -1,0,1, then there would be 9 separate
##      optimizations to see which is best.  If the number of residues
##      optimized is  more than 'MaxResIntOpt' but 'IntConstrain' is on,
##      then an approximate optimization will be performed.  In the
##      first pass, there will be no integer constraints, only bounds on
##      the total residue charge of 'Max|Qres|'.  On the second pass,
##      each residue will be forced to the integer charge nearest to the
##      free optimum charge.  This might not be the global optimum charge
##      distribution, however.
## 'Max|Qres|' [default = 1.0]
##      Determines the maximum magnitude of charge that any residue
##      can have (integer optimized or not) for protein residues.
##      ** NOTE ** that if this value is LARGE and IntConstrain=yes, then
##      it may take a very long time for optimization because all residue
##      charges between [-Max|Qres|,Max|Qres|] will be tried.  This
##      parameter is a bound.  There is no way of turning it
##      'off', you can just adjust it to be as large as you wish.
## 'Max|Qi|' [default=0.85]
##      Determines the maximum partial atomic charge on ant atom center
##      for protein residues.  This parameter is a bound.  There is no
##      way of turning it 'off', you can just adjust it to be as
##      large as you wish.
## 'Set|Qtot|' [default=999.0]
##      Constrains the total optimized charge to some value.
##      When 'Set|Qtot|=999.0', the constraint is turned OFF.
## 'DipConstrain' [no|yes] [default=no]
##      If yes, constraints will be applied on the dipole moments of each
##      residue.  If no, Max|Dres|, Max|Dres|o, and Max|Dres|c will be
##      ignored.
## 'Max|Dres|' [default=1.5]
##      Determines the maximum dipole (in each direction) that a residue
##      may have when integer charge constraints are not applied (this also
##      applies to the first pass in the approximate optimization) for
##      protein residues.  Note that residue dipole constraints are applied
##      to the dipole magnitude in each of 5 directions:  x, y, z and also
##      along two of the long diagonals to get an approximate 'spherical'
##      constraint on the total dipole.  This constraint is a bound, as
##      such there is no way to turn it off,  though you can adjust it to
##      be as large as you wish.
## 'Max|Dres|o' [default=1.0]
##      Determines the maximum dipole (in each direction) that a residue
##      may have when it is forced to have zero charge for protein residues
##      This constraint is a bound, as such there is no way to turn it off,
##      though you can adjust it to be as large as you wish.
## 'Max|Dres|c' [default=1.5]
##      Determines the maximum dipole (in each direction) that a residue
##      may have when it is forced to have a nonzero integer charge for
##      protein residues.  This constraint is a bound, as such there is no
##      way to turn it off, though you can adjust it to be as large as you
##      wish.

QrConstrain   = yes
IntConstrain  = yes
MaxResIntOpt  = 3
Set|Qtot|     = 999.0
Max|Qres|     = 1.00
Max|Qi|       = 0.85
DipConstrain  = no
Max|Dres|     = 1.50
Max|Dres|o    = 1.50
Max|Dres|c    = 1.50

## ANALYSIS PARAMETERS
##
## 'sim_menu' Chooses the 'menu option' for determining what functions is
##      used to compart the similarity of potentials on the molecular
##      surface.  The possible options are... [default=1]
##        1.  RSD = \sqrt{ \sum_i (p1_i-p2_i)^2 }
##        2.  ABSD = \sum_i | p1_i - p2_i |
##        3.  NABSD = ABSD / \sum_i ( |p1_i| + |p2_i| )
##        4.  COS = p1 . p2 / (|p1|x|p2|)
##        5.  NRSD = RDS / \sqrt{ \sum_i (p1_i^2 + p2_i^2) }
##
## 'verbose' [0,1] [default=0] Affects the verbosity of the output.

sim_menu   = 1
verbose    = 0

## OPTIMIZE A SUBSET OF LIGAND CHARGES
```

```
##
## Select a subset of all computed ligand residues to optimize,
## leaving all the rest at their wild-type values.  If this section is
## omitted, that is the same as optimizing on ALL computed basis points.
##
## keyword 'select_residues' followed by a list of selections, terminated
##   by the keyword 'end' on a line by itself.
##
## selection lines have the form 'segid resid [resname]'
##   where 'segid' is the segment if of the residue
##   'resid' is the second-column residue id.  This should be the same
##   in all complexes.  'resname' is an optimal text label for comment
##   purposes.
##
##

select_residues
  H 1 CHO
end

## CONSTRAIN CERTAIN BASIS POINTS TO SPECIFIED CHARGES
##
## If you wish, you may constrain certain atom centers to have specified
##    fixed charge values.  this is done in the 'fix_charges' section.
##    This section lists all atoms which shall have fixed charges in the
##    format 'segid resid atomtype [charge]' where 'resid' is the second
##    column residue id and 'atomtype' is the atom type string.
##    '[charge]' is the charge to fix the atom to.  If this is omitted,
##    the atom will be fixed to the respective wild-type charge.  All
##    charges fixed in this way will be reflected in the 'reference'
##    charge distribution.
##

fix_charges
  H 1 C1 0.75
  H 1 C2
end

## APPLY PROPORTIONALITY CONSTRAINTS
##
## You man apply constraints to make pairs of charges
##    proportional to each other.  I.e. force them to be equal
##    or opposite in value.  Use the 'proportionality' section.
## The constraint lines have the form
##    'seg1 resid1 type1 seg2 resid2 type2 [const]'
## The two basis points are specified by their segment ID,
##   atomtype and second column residue id.  If Q1 and Q2
##   denote the charges of these two basis points, respectively,
##   then the constraint takes the form 'Q1 = [const] Q2'
##   If [const] is not specified, then it defaults to '1.0'
##   which corresponds to forcing the charges to be equal.
## You can have any number of these constraints.  You can force
##   more then two atoms to have the same charge by using the
##   transitive property.
##

proportionality
  H 1 C3  H 1 C4 -1.0
  H 1 C5  H 1 C6
end
```

B.6.4 ligand_scan.prl (ligandscan.cfg)

```
##
## Sample Ligand Scanning Configuration File
##

## GENERAL SETUP
##
## 'datafile'
##      File name for output ligand rankings.
## 'multiples' [ 0 | 1 | 2 ]
##      How to treat multiple mutations.
##      0 = single only, 1 = doubles, 2 = all combinations

datafile = ligand_scan.out
multiples = 2

## DEFINE MUTATIONS

## 'select_mutations' (terminated by 'end')
##      Each line contains a definition of a mutation in the format
##      displayed below.  Following the number of atoms involved in
##      the mutation comes a list of all involved atoms, with the
##      mutation ID specified last.
select_mutations
#   SEGID RESID RES     NATOM   ATOM1   ATOM2   MUTATION
#   ----- ----- ---     -----   -----   -----   --------
        D    13 DRP         2     CD1     HD1       H->F
        D    13 DRP         2     NE1     HE1       H->F
        D    13 DRP         2     CZ2     HZ2       H->F
        D    13 DRP         2     CH2     HH2       H->F
        D    13 DRP         2     CE3     HE3       H->F
        D    13 DRP         2     CZ3     HZ3       H->F
end

## 'mutation_rules' (terminated by 'end')
##     Define mutation types.  Each ATOM/RULE pair specifies the
##     identity of the atom before and after the mutation, followed
##     by the operation on the charge of the initial atom to produce
##     the mutant.
mutation_rules
#   MUTATION    NATOM   ATOM1  QRULE1  ATOM2   QRULE2
#   --------    -----   -----  ------  -----   ------
        H->F        2     C,C  +=0.25    H,F   -=0.25
        H->F        2     N,N  +=0.40    H,F   -=0.40
end


## RULES FOR QM CALCULATIONS

## 'qm_define_region' (terminated by 'end')
##      Define region to be considered in the QM calculations.
##      A SEGID/RESID/RES is specified, followed by any atoms to
##      exclude.
qm_define_region
#   SEGID RESID RES     EXCLUDED ATOM LIST
#   ----- ----- ---     ------------------
        D    13 DRP
end

## 'qm_add_atoms' (terminated by 'end')
##      Define any atoms which must be added prior to the QM
##      calculation.  This is generally used to fill the valencies
##      of any aliphatic carbons.  A Z-matrix type specification
##      defines how the atom is initially placed.
qm_add_atoms
#      NEW       BOND   ANGLE     DIHED   MUTATION
#      ---      ------  ------  --------   --------
        H97     CB 1.0  CG 110 CD1  -60        all
        H98     CB 1.0  CG 110 CD1  +60        all
        H99     CB 1.0  CG 110 CD1  180        all
end
```

```
## 'qm_define_aliphatic' (terminated by 'end')
##      Any groups that should be defined as aliphatic for RESP
##      charge fitting are defined here.
qm_define_aliphatic
#   CARBON      Hydrogens
#   ------      ---------
        CB      H97 H98 H99
end

## 'qm_define_non_polar' (terminated by 'end')
##       Any hydrogen atoms that should be treated as non-polar,
##       and thus constrained to 0.0 in the RESP charge fitting,
##       are defined here.
qm_define_non_polar
        H97
        H98
        H99
end
```

Bibliography

[1] J. Antosiewicz, J. A. McCammon, and M. K. Gilson. Prediction of pH-dependent properties of proteins. *J. Mol. Biol.*, **238**:415–436, 1994.

[2] J. Antosiewicz, J. A. McCammon, and M. K. Gilson. The determination of pK(a)s in proteins. *Biochemistry*, **35**:7819–7833, 1996.

[3] G. Archontis, T. Simonson, and M. Karplus. Binding free energies and free energy components from molecular dynamics and Poisson–Boltzmann calculations. Application to amino acid recognition by aspartyl-tRNA synthetase. *J. Mol. Biol.*, **306**:307–327, 2001.

[4] J. G. Arnez and T. A. Steitz. Crystal structure of unmodified tRNA$^{\mathrm{Gln}}$ complexed with glutaminyl-tRNA synthetase and ATP suggests a possible role for pseudo-uridines in stabilization of RNA structure. *Biochemistry*, **33**:7560–7567, 1994.

[5] J. G. Arnez and T. A. Steitz. Crystal structures of three misacylating mutants of *Escherichia coli* glutaminyl-tRNA synthetase complexed with tRNA$^{\mathrm{Gln}}$ and ATP. *Biochemistry*, **35**:14725–14733, 1996.

[6] C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for determining atom-centered charges: The RESP model. *J. Phys. Chem.*, **97**:10269–10280, 1993.

[7] B. H. Besler, K. M. Merz, and P. A. Kollman. Atomic charges derived from semiempirical methods. *J. Comput. Chem.*, **11**:431–439, 1990.

[8] T. Bhattacharyya, A. Bhattacharyya, and S. Roy. A fluorescence spectroscopic study of glutaminyl-tRNA synthetase from *Escherichia coli* and its implications for the enzyme mechanism. *Eur. J. Biochem.*, **200**:739–745, 1991.

[9] J. O'M. Bockris and A. K. N. Reddy. *Modern Electrochemistry*. Plenum, New York, 1973.

[10] C. M. Breneman and K. B. Wiberg. Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformation analysis. *J. Comput. Chem.*, **11**:361–373, 1990.

[11] 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. Comput. Chem.*, **4**:187–217, 1983.

[12] F. K. Brown and P. A. Kollman. Molecular-dynamics simulations of loop closing in the enzyme triose phosphate isomerase. *J. Mol. Biol.*, **198**:533–546, 1987.

[13] A. T. Brunger, C. L. Brooks, and M. Karplus. Active-site dynamics of ribonuclease. *Proc. Natl. Acad. Sci. U.S.A.*, **82**:8458–8462, 1985.

[14] A. T. Brünger and M. Karplus. Polar hydrogen positions in proteins: Empirical energy placement and neutron diffraction comparison. *Proteins*, **4**:148–156, 1988.

[15] S. Cabani, P. Gianni, V. Mollica, and L. Lepori. Group contributions to the thermodynamic properties of non-ionic organic solutes in dilute aqueous solution. *J. Sol. Chem.*, **10**:563–595, 1981.

[16] M. Caffrey, M. Cai, J. Kaufman, S. J. Stahl, P. T. Wingfield, D. G. Covell, A. M. Gronenborn, and G. M. Clore. Three-dimensional solution structure of the 44 kDa ectodomain of SIV gp41. *EMBO J.*, **17**:4572–4584, 1998.

[17] J. A. Caravella. *Electrostatics and Packing in Biomolecules: Accounting for Conformational Change in Protein Folding and Binding.* PhD thesis, Massachusetts Institute of Technology, 2002.

[18] H. A. Carlson, J. M. Briggs, and J. A. McCammon. Calculation of the pK(a) values for the ligands and side chains of *Escherichia coli* D-alanine:D-alanine ligase. *J. Med. Chem.*, **42**:109–117, 1999.

[19] H. A. Carlson, T. B. Nguyen, M. Orozco, and W. L. Jorgensen. Accuracy of free energies of hydration for organic molecules from 6-31G*-derived partial charges. *J. Comput. Chem.*, **14**:1240–1249, 1993.

[20] D. C. Chan, C. T. Chutkowski, and P. S. Kim. Evidence that a prominent cavity in the coiled coil of HIV type 1 gp41 is an attractive drug target. *Proc. Natl. Acad. Sci. U.S.A.*, **95**:15613–15617, 1998.

[21] D. C. Chan, D. Fass, J. M. Berger, and P. S. Kim. Core structure of gp41 from the HIV envelope glycoprotein. *Cell*, **89**:263–273, 1997.

[22] L. E. Chirlain and M. M. Francl. Atomic charges determined from electrostatic potentials: A detailed study. *J. Comput. Chem.*, **8**:894–905, 1987.

[23] 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.

[24] C. Chothia. Hydrophobic bonding and accessible surface area in proteins. *Nature (London)*, **248**:338–339, 1974.

[25] C. Chothia. Nature of accesible and buried surfaces in proteins. *J. Mol. Biol.*, **105**:1–14, 1976.

[26] C. Chothia and J. Janin. Principles of protein–protein recognition. *Nature (London)*, **256**:705–708, 1975.

[27] L. L. Conte, C. Chothia, and J. Janin. The atomic structure of protein–protein recognition sites. *J. Mol. Biol.*, **285**:2177–2198, 1999.

[28] W. D. Cornell, P. Cieplak, C. I. Bayly, and P. A. Kollman. Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *J. Am. Chem. Soc*, **115**:9620, 1993.

[29] W. D. Cornell, P. Cieplek, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A 2nd generation force-field for the simulation of proteins, nucleic acids and organic-molecules. *J. Am. Chem. Soc*, **117**:5179–5197, 1995.

[30] V. Daggett and P. A. Kollman. Molecular-dynamics simulations of active-site mutants of triosephosphate isomerase. *Protein Eng.*, **3**:677–690, 1990.

[31] B. I. Dahiyat and S. L. Mayo. De novo protein design: Fully automated sequence selection. *Science (Washington, D.C.)*, **278**:82–87, 1997.

[32] P. I. W. de Bakker, P. H. Hunenberger, and J. A. McCammon. Molecular dynamics simulations of the hyperthermophilic protein Sac7d from *Sulfolobus acidocaldarius*: Contribution of salt bridges to thermostability. *J. Mol. Biol.*, **285**:1811–1830, 1999.

[33] M. De Maeyer, J. Desmet, and I. Lasters. The dead-end elimination theorem: Mathematical aspects, implementation, optimizations, evaluation, and performance. *Methods Mol. Biol.*, **143**:265–304, 2000.

[34] A. K. Debnath, L. Radigan, and S. Jiang. Structure-based identification of small molecule antiviral compounds targeted to the gp41 core structure of the human immunodeficiency virus type 1. *J. Med. Chem.*, **42**(3203-3209), 1999.

[35] J. R. Desjarlais and T. M. Handel. De novo design of the hydrophobic cores of proteins. *Protein Sci.*, **4**:2006–2018, 1995.

[36] R. L. DesJarlais, R. P. Sheridan, J. S. Dixon, I. D. Kuntz, and R. Venkataraghavan. Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem.*, **29**:2149–2153, 1986.

[37] J. Desmet, M. De Maeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature (London)*, **356**:539–542, 1992.

[38] R. S. DeWitte, A. V. Ishchenko, and E. I. Shaknovich. SMoG: De novo design method based on simple, fast, and accurate free energy estimates. 2. case studies in molecular design. *J. Am. Chem. Soc*, **119**:4608–4617, 1997.

[39] S. B. Dixit, R. Bhasin, E. Rajasekaran, and B. Jayaram. Solvation thermodynamics of amino acids. *J. Chem. Soc., Faraday Trans.*, **93**:1105–1113, 1997.

[40] 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.

[41] D. M. Eckert and P. S. Kim. Design of potent inhibitors of HIV-1 entry from the gp41 N-peptide region. *Proc. Natl. Acad. Sci. U.S.A.*, **98**:11187–11192, 2001.

[42] D. M. Eckert and P. S. Kim. Mechanisms of viral membrane fusion and its inhibition. *Annu. Rev. Biochem.*, **70**:777–810, 2001. Review.

[43] D. M. Eckert, V. N. Malashkevich, L. H. Hong, P. A. Carr, and P. S. Kim. Inhibiting HIV-1 entry: Discovery of D-peptide inhibitors that target the gp41 coiled-coil pocket. *Cell*, **99**:103–115, 1999.

[44] R. Elber and M. Karplus. Multiple conformational states of proteins — a molecular-dynamics analysis of myoglobin. *Science (Washington, D.C.)*, **235**:318–321, 1987.

[45] A. H. Elcock, R. R. Gabdoulline, R. C. Wade, and J. A. McCammon. Computer simulations of protein–protein association kinetics: Acetylcholinesterase-fasciculin. *J. Mol. Biol.*, **291**:149–162, 1999.

[46] M. Ferrer, T. M. Kapoor, T. Strassmaier, W. Weissenhorn, J. J. Skehel, D. Oprian, S. L. Schreiber, D. C. Wiley, and S. C. Harrison. Selection of gp41-mediated HIV-1 cell entry inhibitors from biased combinatorial libraries of non-natural binding elements. *Nat. Struct. Biol.*, **6**:953–959, 1999.

[47] W. Freist. Mechanisms of aminoacyl-tRNA synthetases: A critical consideration of recent results. *Biochemistry*, **28**:6787–6795, 1989. Review.

[48] W. Freist, D. H. Gauss, M. Ibba, and D. Söll. Glutaminyl-tRNA synthetase. *Biol. Chem.*, **378**:1103–1117, 1997. Review.

[49] M. J. Frisch, G. W. Trucks, H. B. Schlegel, P. M. W. Gill, B. G. Johnson, M. A. Robb, J. R. Cheeseman, T. Keith, G. A. Petersson, J. A. Montgomery, K. Raghavachari, M. A. Al-Laham, V. G. Zakrzerwski, J. V. Ortiz, J. B. Foresman, J. Cioslowski, B. B. Stefanov, A. Nanayakkara, M. Challacombe, C. Y. Peng, P. Y. Ayala, W. Chen, M. W. Wong, J. L. Andres, E. S. Replogle, R. Gomperts, R. L. Martin, D. J. Fox, J. S. Binkley, D. J. Defrees, J. Baker, J. P. Stewart, M. Head-Gordon, C. Gonzalez, and J. A. Pople. *Gaussian 94, Revision D.1.* Gaussian, Inc., Pittsburgh, PA, 1995.

[50] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzerwski, J. A. Montgomery, Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, J. 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, 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. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, 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, Revision A.7.* Gaussian, Inc., Pittsburgh, PA, 1995.

[51] N. Froloff, A. Windemuth, and B. Honig. On the calculation of binding free energies using continuum methods: Application to MHC class I protein–peptide interactions. *Protein Sci.*, **6**:1293–1301, 1997.

[52] J. Gao, K. Kuczera, B. Tidor, and M. Karplus. Hidden thermodynamics of mutant proteins — a molecular-dynamics analysis. *Science (Washington, D.C.)*, **244**:1069–1072, 1989.

[53] M. A. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys. J.*, **72**:1047–1069, 1997.

[54] M. K. Gilson and B. Honig. Destabilization of an alpha-helix-bundle protein by helix dipoles. *Proc. Natl. Acad. Sci. U.S.A.*, **86**:1524–1528, 1989.

[55] 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.

[56] M. K. Gilson and B. H. Honig. Calculation of electrostatic potentials in an enzyme active site. *Nature (London)*, **330**:84–86, 1987.

[57] M. K. Gilson, K. A. Sharp, and B. H. Honig. Calculating the electrostatic potential of molecules in solution: Method and error assessment. *J. Comput. Chem.*, **9**:327–335, 1988.

[58] D. B. Gordon and S. L. Mayo. Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J. Comput. Chem.*, **19**:1505–1514, 1998.

[59] P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, and P. S. Kim. High-resolution protein design with backbone freedom. *Science (Washington, D.C.)*, **282**:1462–1467, 1998.

[60] P. B. Harbury, B. Tidor, and P. S. Kim. Repacking protein cores with backbone freedom: Structure prediction for coiled coils. *Proc. Natl. Acad. Sci. U.S.A.*, **92**:8408–8412, 1995.

[61] H. W. Hellinga and F. M. Richards. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **91**:5803–5807, 1994.

[62] R. H. Henchman and J. W. Essex. Free energies of hydration using restratined electrostatic potential derived charges via free energy pertubations and linear response. *J. Comput. Chem.*, **20**:499–510, 1999.

[63] R. H. Henchman and J. W. Essex. Generation of OPLS-like charges from molecular electrostatic potential using restraints. *J. Comput. Chem.*, **20**:483–498, 1999.

[64] Z. S. Hendsch. *Continuum Electrostatic Calculations of Biological Molecules.* PhD thesis, Massachusetts Institute of Technology, 2001.

[65] Z. S. Hendsch, T. Jonsson, R. T. Sauer, and B. Tidor. Protein stabilization by removal of unsatisfied polar groups: Computational approaches and experimental tests. *Biochemistry*, **35**:7621–7625, 1996.

[66] Z. S. Hendsch, M. J. Nohaile, R. T. Sauer, and B. Tidor. Preferential heterodimer formation via undercompensated electrostatic interactions. *J. Am. Chem. Soc*, **123**:1264–1265, 2001.

[67] Z. S. Hendsch and B. Tidor. Electrostatic interactions in the 434 repressor-or1 operator complex. *Manuscript in preparation.*

[68] Z. S. Hendsch and B. Tidor. Do salt bridges stabilize proteins? A continuum electrostatics analysis. *Protein Sci.*, **3**:211–226, 1994.

[69] 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.

[70] B. Honig, K. Sharp, and A.-S. Yang. Macroscopic models of aqueous solutions: Biological and chemical applications. *J. Phys. Chem.*, **97**:1101–1109, 1993. Review.

[71] S. B. Jiang and A. K. Debnath. A salt bridge between an N-terminal coiled coil of gp41 and an antiviral agent targeted to the gp41 cove is important for anti-HIV-1 activity. *Biochem. Biophys. Res. Comm.*, **270**:153–157, 2000.

[72] D. T. Jones. De novo protein design using pairwise potentials and a genetic algorithm. *Protein Sci.*, **3**:567–574, 1994.

[73] W. L. Jorgensen and J. Tirado-Rives. The OPLS potential function for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc*, **110**:1657–1666, 1988.

[74] D. Joseph-McCarthy, J. M. Hogle, and M. Karplus. Use of the multiple copy simulateous search (MCSS) method to design a new class of picornavirus capsid binding drugs. *Proteins*, **29**:32–58, 1997.

[75] J. K. Judice, J. Y. K. Tom, W. Huang, T. Wrin, J. Vennari, C. J. Petropoulos, and R. S. McDowell. Inhibition of HIV type 1 infectivity by constrained α-helical peptides: Implications for the viral fusion mechanism. *Proc. Natl. Acad. Sci. U.S.A.*, **94**:13426–13430, 1997.

[76] E. Kangas. *Optimizing Molecular Electostatic Interactions: Binding Affinity and Specificity.* PhD thesis, Massachusetts Institute of Technology, 2000.

[77] E. Kangas and B. Tidor. Optimizing electrostatic affinity in ligand–receptor binding: Theory, computation, and ligand properties. *J. Chem. Phys.*, **109**:7522–7545, 1998.

[78] E. Kangas and B. Tidor. Charge optimization leads to favorable electrostatic binding free energy. *Phys. Rev. E*, **59**:5958–5961, 1999.

[79] E. Kangas and B. Tidor. Electrostatic specificity in molecular ligand design. *J. Chem. Phys.*, **112**:9120–9131, 2000.

[80] E. Kangas and B. Tidor. Electrostatic complementarity at ligand binding sites: Application to chorismate mutase. *J. Phys. Chem. B*, **105**:880–888, 2001.

[81] M. Karplus and J. A. McCammon. Protein structural fluctuations during a period of 100-ps. *Nature (London)*, **277**:578–578, 1980.

[82] M. Karplus and G. A. Petsko. Molecular-dynamics simulations in biology. *Nature (London)*, **347**:631–639, 1990. Review.

[83] 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: Struct., Funct., Genet.*, **1**:47–59, 1986.

[84] P. Koehl and M. Delarue. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.*, **239**:249–275, 1994.

[85] P. Koehl and M. Levitt. De novo protein design. I. In search of stability and specificity. *J. Mol. Biol.*, **293**:1161–1181, 1999.

[86] P. Koehl and M. Levitt. Structure-based conformational preferences of amino acids. *Proc. Natl. Acad. Sci. U.S.A.*, **96**:12524–12529, 1999.

[87] P. J. Kraulis. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.*, **24**:946–950, 1991.

[88] I. Lasters, M. De Maeyer, and J. Desmet. Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains. *Protein Eng.*, **8**:815–822, 1995.

[89] F. T. K. Lau and M. Karplus. Molecular recognition in proteins — simulation analysis of substrate-binding by a tyrosyl-transfer-RNA synthetase mutant. *J. Mol. Biol.*, **236**:1049–1066, 1994.

[90] A. R. Leach and A. P. Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins: Struct., Funct., Genet.*, **33**:227–239, 1998.

[91] C. Lee and S. Subbiah. Prediction of protein side-chain conformation by packing optimization. *J. Mol. Biol.*, **217**:373–388, 1991.

[92] L.-P. Lee and B. Tidor. Optimization of electrostatic binding free energy. *J. Chem. Phys.*, **106**:8681–8690, 1997.

[93] L.-P. Lee and B. Tidor. Barstar is electrostatically optimized for tight binding to barnase. *Nat. Struct. Biol.*, **8**:73–76, 2001.

[94] L.-P. Lee and B. Tidor. Optimization of binding electrostatics: Charge complementarity in the barnase-barstar protein complex. *Protein Sci.*, **10**:362–377, 2001.

[95] T. S. Lee and P. A. Kollman. Theoretical studies suggest a new antifolate as a more potent inhibitor of thimidylate synthase. *J. Am. Chem. Soc*, **122**:4385–4393, 2000.

[96] L. L. Looger and H. W. Hellinga. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: Implications for protein design and structural genomics. *J. Mol. Biol.*, **307**:429–445, 2001.

[97] J. M. Louis, C. A. Bewley, and G. M. Clore. Design and properties of N_{CCG}-gp41, a chimeric gp41 molecule with nanomolar HIV fusion inhibitory activity. *J. Biol. Chem.*, **276**:29485–29489, 2001.

[98] M. Lu, S. C. Blacklow, and P. S. Kim. A trimeric structural domain of the HIV-1 transmembrane glycoprotein. *Nat. Struct. Biol.*, **2**:1075–1082, 1995.

[99] J. P. Ma, P. B. Sigler, Z. H. Xu, and M. Karplus. A dynamic model for the allosteric mechanism of GroEL. *J. Mol. Biol.*, **302**:303–313, 2000.

[100] 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, 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.

[101] V. N. Malashkevich, D. C. Chan, C. T. Chutkowski, and P. S. Kim. Crystal structure of the simian immunodeficiency virus (SIV) gp41 core: Conserved

helical interations underlie the broad inhibitory activity of gp41 peptides. *Proc. Natl. Acad. Sci. U.S.A.*, **95**:9134–9139, 1998.

[102] A. C. R. Martin. ProFit v.1.8. http://www.bioinf.uk.org/software/profit-/index.html.

[103] J. A. McCammon, B. R. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature (London)*, **267**:585–590, 1977.

[104] M. A. McCarrick and P. A. Kollman. Predicting relative binding affinities of non-peptide HIV protease inhibitors with free energy perturbations calculations. *J. Comput.-Aided Mol. Des.*, **13**:109–112, 1999.

[105] E. A. Merritt and D. J. Bacon. Raster3D: Photorealistic molecular graphics. *Methods Enzymol.*, **277**:505–524, 1997.

[106] V. K. Misra, J. L. Hecht, A.-S. Yang, and B. Honig. Electrostatic contributions to the binding free energy of the λ cI repressor to DNA. *Biophys. J.*, **75**:2262–2273, 1998.

[107] V. K. Misra, K. A. Sharp, R. A. Friedman, and B. Honig. Salt effects on ligand–DNA binding: Minor groove binding antibiotics. *J. Mol. Biol.*, **238**:245–263, 1994.

[108] V. Mohan, M. E. Davis, J. A. McCammon, and B. M. Pettitt. Continuum model calculations of solvation free energies: Accurate evaluation of electrostatic contributions. *J. Phys. Chem.*, **96**:6428–6431, 1992.

[109] F. A. Momany. Determination of partial atomic charges from ab initio molecular electrostatic potentials. application to formamide, methanol, and formic acid. *J. Phys. Chem.*, **83**:592–601, 1978.

[110] R. S. Mulliken. Electronic population analysis on LCAO–MO molecular wave functions I. *J. Chem. Phys.*, **23**:1833, 1955.

[111] A. Nicholls, K. A. Sharp, and B. Honig. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Struct., Funct., Genet.*, **11**:281–296, 1991.

[112] M. Nina, D. Beglov, and B. Roux. Atomic radii for continuum electrostatics calculations based on molecular dynamics free energy simulations. *J. Phys. Chem. B*, **101**:5239–5248, 1997.

[113] M. J. Nohaile, Z. S. Hendsch, B. Tidor, and R. T. Sauer. Altering dimerization specificity by changes in surface electrostatics. *Proc. Natl. Acad. Sci. U.S.A.*, **98**:3109–3114, 2001.

[114] C. O. Pabo. Molecular technology: Designing proteins and peptides. *Nature (London)*, **301**:200, 1983.

[115] E. Paci, A. Caflisch, A. Pluckthun, and M. Karplus. Forces and energetics of hapten-antibody dissociation: A biased molecular dynamics simulation study. *J. Mol. Biol.*, **314**:589–605, 2001.

[116] L. Pauling. *The Nature of the Chemical Bond.* Cornell University Press, New York, Third edition, 1960.

[117] J. J. Perona, M. A. Rould, and T. A. Steitz. Structural basis for transfer RNA aminoacylation by *Escherichia coli* glutaminyl-tRNA synthetase. *Biochemistry*, **32**:8758–8771, 1993.

[118] M. J. Potter, M. K. Gilson, and J. A. McCammon. Molecule pK(a) prediction with continuum electrostatics. *J. Am. Chem. Soc*, **116**:10298–10299, 1994.

[119] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, Cambridge, Second edition, 1992.

[120] M. Prevost, S. J. Wodak, B. Tidor, and M. Karplus. Contribution of the hydrophobic effect to protein stability — analysis based on simulations of the Ile-96-Ala mutation in barnase. *Proc. Natl. Acad. Sci. U.S.A.*, **88**:10880–10884, 1991.

[121] Z. Radic, P. D. Kirchhoff, D. M. Quinn, J. A. McCammon, and P. Taylor. Electrostatic influences on the kinetics of ligand binding to acetylcholinesterase — distinctions between active center ligands and fasciculin. *J. Biol. Chem.*, **273**:23265–23277, 1997.

[122] R. J. Radmer and P. A. Kollman. The application of three approximate free energy calculations methods to structure based ligand design: Trypsin and its complex with inhibitors. *J. Comput.-Aided Mol. Des.*, **12**:215–227, 1998.

[123] G. Rastelli, B. Thomas, P. A. Kollman, and D. V. Santi. Insight into the specificity of thymidylate synthase from molecular-dynamics and free-energy perturbation calculations. *J. Am. Chem. Soc*, **117**:7213–7227, 1995.

[124] V. L. Rath, L. F. Silvian, B. Beijer, B. S. Sproat, and T. A. Steitz. How glutaminyl-tRNA synthetase selects glutamine. *Structure*, **6**:439–449, 1998.

[125] Research Collaboratory for Structural Bioinformatics (RCSB). PROTEIN DATA BANK. http://www.rcsb.org/pdb/.

[126] C. M. Reyes and P. A. Kollman. Structure and thermodynamics of RNA–protein binding: Using molecular dynamics and free energy analyses to calculate the free energies of binding and conformational change. *J. Mol. Biol.*, **297**:1145–1158, 2000.

[127] M. J. Root, M. S. Kay, and P. S. Kim. Protein design of an HIV-1 entry inhibitor. *Science (Washington, D.C.)*, **291**:884–888, 2001. 10.1126/science.1057453.

[128] M. Sanner. msms v.2.5.3. http://www.scripps.edu/pub/olson-web/people-/sanner/html/msms_home.html.

[129] M. Schaefer, , and M. Sommer, M. Karplus. pH-dependence of protein stability: Absolute electrostatic free energy differences between conformations. *J. Phys. Chem. B*, **101**:1663–1683, 1997.

[130] Schrödinger, Inc., Portland, OR. *Jaguar v3.5*, 1998.

[131] G. L. Seibel, U. C. Singh, and P. A. Kollman. A molecular-dynamics simulation of double-helical B-DNA including counterions and water. *Proc. Natl. Acad. Sci. U.S.A.*, **82**:6537–6540, 1985.

[132] T. Selzer, S. Albeck, and G. Schreiber. Rational design of faster associating and tighter binding protein complexes. *Nature Struct. Biol.*, **7**:537–541, 2000.

[133] D. Shanno and R. J. Vanderbei. An interior-point method for nonconvex nonlinear programming. *Computational Optimization and Applications*, **12**, 1999.

[134] K. A. Sharp and B. Honig. Calculating total electrostatic energies with the nonlinear Poisson–Boltzmann equation. *J. Phys. Chem.*, **94**:7684–7692, 1990.

[135] K. A. Sharp, A. Nicholls, R. F. Fine, and B. Honig. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science (Washington, D.C.)*, **252**:106–109, 1991.

[136] K. A. Sharp and B. Honig. Electrostatic interactions in macromolecules: Theory and applications. *Annu. Rev. Biophys. Biophys. Chem.*, **19**:301–332, 1990.

[137] F. B. Sheinerman, R. Norel, and B. Honig. Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol.*, **10**:153–159, 2000.

[138] W. Shu, H. Ji, L. Radigen, S. Jiang, and M. Lu. Helical interactions in the HIV-1 gp41 core reveal structural basis for the inhibitory activity of gp41 peptides. *Biochemistry*, **39**:1634–1642, 2000.

[139] U. C. Singh and P. A. Kollman. An approach to computing electrostatic charges for molecules. *J. Comput. Chem.*, **5**:129–145, 1984.

[140] 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.

[141] 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.

[142] J. J. P. Stewart. MOPAC 7.0 — Public Domain Version.

[143] G. Strang. *Introduction to Applied Mathematics*. Wellesley–Cambridge Press, Wellesley, Massachusetts, 1986.

[144] N. C. J. Strynadka, S. E. Jensen, P. M. Alzari, and M. N. G. James. A potent new mode of inhibition revealed by the 1.7 Å X-ray crystallographic structure of the TEM-1–BLIP complex. *Nature Struct. Biol.*, **3**:290–297, 1996.

[145] Y. C. Sun, D. L. Veenstra, and P. A. Kollman. Free energy calculations of the mutation of Ile96->Ala in barnase: Contributions ot the difference in stability. *Protein Eng.*, **9**:273–281, 1996.

[146] Swiss Institute of Bioinformatics (SIB) and European Bioinformatics Institute (EBI). SWISS-PROT. http://www.ebi.ac.uk/swissprot/.

[147] K. Tan, J.-H. Liu, J.-H. Wang, S. Shen, and M. Lu. Atomic structure of a thermostable subdomain of HIV-1 gp41. *Proc. Natl. Acad. Sci. U.S.A.*, **94**:12303–12308, 1997.

[148] C. Tanford and J. G. Kirkwood. Theory of protein titration curves. I. General equations for impenetrable spheres. *J. Am. Chem. Soc*, **79**:5333–5339, 1957.

[149] J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. The ClustalX windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. **24**:4876–4882, 1997.

[150] B. Tidor and M. Karplus. Simulation analysis of the stability of mutant R96H of T4 lysozyme. *Biochemistry*, **30**:3217–3228, 1991.

[151] B. Tidor and M. Karplus. The contribution of vibrational entropy to molecular association. the dimerization of insulin. *J. Mol. Biol.*, **238**:405–411, 1994.

[152] W. F. van Gunsteren and H. J. C. Berendsen. Computer simulation of molecular dynamics: Methodology, applications and perspectives in chemistry. *Agnew. Chem. Int. Ed. Engl.*, **29**:992–1023, 1990.

[153] H. W. T. van Vlijmen, M. Schaefer, and M. Karplus. Improving the accuracy of protein pK(a) calculations: Conformational averaging versus the average structure. *Proteins*, **33**:145–158, 1998.

[154] R. J. Vanderbei. LOQO: An interior-point code for quadratic programing. *Optimization Methods and Software*, **12**:451–454, 1999.

[155] R. J. Vanderbei. LOQO User's Manual — Version 3.10. *Optimization Methods and Software*, **12**:485–514, 1999.

[156] J. Wang, R. Dixon, and P. A. Kollman. Ranking ligand binding affinities with avidin: A molecular dynamics-based interaction energy study. *Proteins*, **34**:69–81, 1999.

[157] W. Wang, O. Donini, C. M. Reyes, and P. A. Kollman. Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis,

protein–ligand, protein–protein, and protein–nucleic acid noncovalent interactions. *Annu. Rev. Biophys. Biomolec. Struct.*, **30**:211–243, 2001. Review.

[158] J. Warwicker and H. C. Watson. Calculation of the electric potential in the active site cleft due to α-helix dipoles. *J. Mol. Biol.*, **157**:671–679, 1982.

[159] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, Jr., and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc*, **106**:765–784, 1984.

[160] S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.*, **7**:230–252, 1986.

[161] W. Weissenhorn, A. Dessen, S. C. Harrison, J. J. Skehel, and D. C. Wiley. Atomic structure of the ectodomain from HIV-1 gp41. *Nature (London)*, **387**:426–430, 1997.

[162] C. Wild, O. T., M. C., D. Bolognesi, and M. T. A synthetic peptide inhibitor of Human-Immunodeficiency-Virus replication – correlation between solution structure and viral inhibition. *Proc. Natl. Acad. Sci. U.S.A.*, **89**:10537–10541, 1992.

[163] J. Wiorkiewicz and M. Karplus. Personal communication.

[164] L. Xiao and B. Honig. Electrostatic contributions to the stability of hyperthermophilic proteins. *J. Mol. Biol.*, **289**:1435–1444, 1999.

[165] A. S. Yang, M. R. Gunner, R. Sampogna, K. Sharp, and B. Honig. On the calculation of pK(a)s in proteins. *Proteins*, **15**:252–265, 1993.

[166] M. Zacharias, B. A. Luty, M. E. Davis, and J. A. McCammon. Poisson–Boltzmann analysis of the lambda-repressor-operator interaction. *Biophys. J.*, **63**:1280–1285, 1992.

[167] G. Zhou, M. Ferrer, R. Chopra, T. M. Kapoor, T. Strassmaier, W. Weissenhorn, J. J. Skehel, D. Oprian, S. L. Schreiber, S. C. Harrison, and D. C. Wiley. The structure of HIV-1 specific cell entry inhibitor in complex with the HIV-1 gp41 trimeric core. *Bioorgan. Med. Chem.*, **8**:2219–2228, 2000.

Curriculum Vitae

Education

Doctor of Philosophy in Biological Chemistry

Massachusetts Institute of Technology *Sept. 1997 – Sept. 2002*
Cambridge, MA 02139
Thesis Title: Optimization of Electrostatic Binding Free Energy: Applications to the analysis and Design of Ligand Binding in Protein Complexes.

Bachelor's of Science in Chemistry (Honours, Minor in English Literature)

Simon Fraser University *Sept. 1993 – May 1997*
Burnaby, BC, Canada V5A 1S6

Research Positions

Massachusetts Institute of Technology — Cambridge, MA 02139

Laboratory of Prof. Bruce Tidor *Nov. 1997 – present*
Development and application of continuum electrostatic methods for the analysis and design of ligands for protein targets.

Simon Fraser University — Burnaby, BC, Canada

Laboratory of Prof. B. Mario Pinto *May – Dec. 1996*
Computational analysis of the conformational preferences of peptide mimics of carbohydrates.

Laboratory of Prof. Roland K. Pomeroy *May – Aug. 1995*
Synthesis and characterization of Group VIII transitional metal complexes containing dative metal–metal bonds.

Laboratory of Prof. B. Mario Pinto *May – Aug. 1994*
NMR studies of the configurational equilibria of *N*-arylglucopyranosylamines and *N*-arylcyclohexylamines - solvent and substituent effects.

Teaching Positions

Massachusetts Institute of Technology — Cambridge, MA 02139

Teaching Assistant: Organic Chemistry I (5.12) *Feb. – May 1998*
Teaching Assistant: Laboratory Chemistry (5.310) *Sept. – Dec. 1997*

Publications

1. D. F. Green and B. Tidor. “Analysis and optimization of electrostatic contributions to binding.” *Current Protocols in Bioinformatics. Manuscript in preparation.*

2. D. F. Green, Brian A. Joughin and B. Tidor. “Design considerations for “action-at-a-distance” interactions that enhance binding affinity.” *Manuscript in preparation.*

3. B. A. Joughin, D. F. Green and B. Tidor. “Improving electrostatic complementarity through “action-at-a-distance” electrostatic interactions.” *Manuscript in preparation.*

4. D. F. Green and B. Tidor. “Design of improved protein inhibitors of HIV-1 cell entry: Optimization of electrostatic interactions at the binding interface.” *Manuscript in preparation.*

5. D. F. Green and B. Tidor. “Optimizing electrostatic and steric interactions at a binding interface: Method development and application to a D-peptide inhibitor of HIV-1 cell entry.” *Manuscript in preparation.*

6. D. F. Green and B. Tidor. “A comprehensive evaluation of *ab initio* charge determination methods for use in continuum electrostatic calculations.” *Manuscript in preparation.*

7. D. F. Green and B. Tidor. “Electrostatic optimization in an enzyme active site: A study of glutaminyl-tRNA synthetase.” *Manuscript in preparation.*

8. C. Mattos, J. D. Cohen, D. F. Green, B. Tidor and M. Karplus. “X-ray structural and simulation analysis of protein mutants: R96H in T4 lysozyme.” *Submitted.*

9. F. Jiang, H. A. Jenkins, D. F. Green, G. P. A. Yap and R. K. Pomeroy. “A novel metal-chain extension reaction: synthesis of $(X)[Os(CO)_3(CNBut)]_nMn(CO)_5$ (X = Cl, Br, I ; n = 1, 2, 3).” *Can. J. Chem* **80** (3): 281-291 (2002).

10. R. J. Batchelor, D. F. Green, B. D. Johnston, B. O. Patrick and B. M. Pinto. “Conformational preferences in glycosylamines. Implications for the exo-anomeric effect.” *Carbohyd. Res.* **330** (3): 421-426 (2001).

11. K. D. Randell, B. D. Johnston, D. F. Green and B. M. Pinto. “Is there a generalized reverse anomeric effect? Substituent and solvent effects on the configurational equilibria of neutral and protonated *N*-arylglucopyranosylamines and *N*-aryl-5-thioglucopyranosylamines.” *J. Org. Chem.* **65** (1): 220-226 (2000).