Protein Design with Hierarchical Treatment of Solvation and Electrostatics

by

Karl J. M. Hanf

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

Doctor of Philosophy in Biological Physics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2002

Author . . . .
Department of Physics
July 25, 2002

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

Certified by . . . .
Alexander van Oudenaarden
Assistant Professor of Physics
Thesis Co-supervisor

Accepted by . . . .
Thomas J. Greytak
Professor of Physics
Associate Department Head for Education

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

Bruce Tidor . . . .
Thesis Committee Chair

Alexander Van Oudenaarden. . . . .
Physics Department Co-Supervisor

Mehran Kardar . . . .
Reader

Xiao-Gang Wen . . . .
Reader

Protein Design with Hierarchical Treatment of Solvation and Electrostatics

by

Karl J. M. Hanf

Submitted to the Department of Physics
on July 25, 2002, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Biological Physics

Abstract

A detailed treatment of the electrostatic energy of biomolecules in solution is used for two applications that require consideration of large numbers of states: multiple-site titration and protein design. The continuum electrostatic model is combined with covalent, van der Waals, and non-polar energy terms, and the statistical mechanical basis for this model is reviewed. Multiple-site titration is modeled with four titratable residues of the protein barstar. A full enumeration of the titration states is used to predict pH-dependent properties of the system, and the effects of several simplifying assumptions are evaluated. The analytical continuum electrostatics (ACE) method, a computationally inexpensive approximation of the electrostatic free energy, is evaluated in the context of predicting group terms of the binding free energy. A primary source of error in the ACE prediction of atomic solvation energies is identified and ameliorated. A procedure is developed which optimizes the parameters of the ACE method in order to minimize its errors as compared to finite-difference solution of the linearized Poisson–Boltzmann equation. Parameter sets optimized on a "testing" biomolecular binding system yield reduced average errors for related biomolecular systems. Finally, a protein design method is developed which uses the dead-end elimination and A* discrete search algorithms to systematically search large numbers (10^{24}) of structures, varying the protein sequence and the side chain conformation at all selected residues. The method is novel in its co-optimization of binding and folding free energies, its use of three levels of increasingly detailed discrete search (sequence, fleximers, and rotamers), and its use of three hierarchical energy functions to successively screen candidate structures identified by the discrete search. Redesigning sets of three and seven residues of the protein barstar, the wild-type sequence, which is experimentally known to bind very tightly to barnase, is ranked very highly by this method (#5 out of 8000, or #89 out of 1.3×10^{9}), unlike that of previous protein design studies. The present method chooses a structure for the wild-type sequence that is very similar to the crystal structure. Several novel sequences predicted to bind more tightly than wild-type barstar are promising candidates for synthesis.

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

Thesis Co-supervisor: Alexander van Oudenaarden
Title: Assistant Professor of Physics

Acknowledgments

I would like to thank my advisor, Bruce Tidor, for always keeping the big picture in mind.

I thank Zachary Hendsch, Justin Caravella, David Green, and Michael Altman (a.k.a. Maltman) for sharing their knowledge; Brian Joughin, Shaun Lippow, and Dan Kamei for their beta testing; and Peter Lee for his inspiring gastronomic prowess.

I would like to thank all the little people on whom I had to step to get where I am today. I would also like to acknowledge the contributions of rich, creamy nougat to this thesis.

Finally, I thank my parents, my wife Diane, and my son Noah for helping me meet the time demands of my research, and for their love.

Contents

Chapter 1

General Introduction

Two crucial functions of proteins are folding into their native conformations and binding to other molecules. These processes are governed by the value of the free energy for each state; for example, binding affinity is determined by the free energy difference of the bound and unbound states of the binding partners. The presence of water has a complex effect on the free energy of molecules in solution. The polar water molecules can generally make better interactions with each other than with any solute molecule; this is the basis of the hydrophobic effect. The polarization of the surrounding water in response to the electric field produced by a solute molecule acts to screen the solute's internal interactions.

We account for the polarization of the water around a solute molecule with a continuum electrostatic model: the water is modeled as a continuum with a high dielectric constant, and a single solute molecule or molecular complex is modeled as a low dielectric region of fixed shape, embedded in the high dielectric region, containing point charges at the centers of its atoms. The electrostatic energy of the system, including the effects of mobile ions in the solution, can then be obtained by finite-difference solution of the linearized Poisson–Boltzmann equation [1, 2, 3, 4, 5].

The continuum electrostatic model, along with covalent, van der Waals, and hydrophobic energy terms, can be used to predict free energy differences between states of a solute molecule or molecular complex. In this thesis, we use such free energy differences to model the processes of binding, folding, titration, and conformational change. The thesis begins and ends with applications that require the evaluation of many competing states: multiple-site titration and protein design. In the middle, we develop approximate methods of evaluating the electrostatic energy that are computationally faster than finite-differences solution of the Poisson–Boltzmann equation, and therefore can be used to evaluate the large numbers of states required for analysis of titration, minimization, or protein design.

In Chapter 2, we review the statistical mechanical basis for our model, beginning with the full quantum mechanical partition function for a solution consisting of explicit solvent, ion, and solute molecules. We show that the states of the system are populated according to their Gibbs free energies. We show the assumptions and reasoning which allow us to consider a single solute molecule or molecular complex, to separate out the electrostatic energy term, and to treat the solute as a dielectric continuum. We derive the Poisson–Boltzmann equation from the Poisson equation. Finally, we describe how the linearized Poisson–Boltzmann equation can be solved by a finite-difference method, and how this method can be used to calculate free energy differences for solvation, titration, binding, folding, or conformational change of a biomolecule.

In Chapter 3, we apply the continuum electrostatic method to the problem of predicting the pH-dependent properties of a protein with multiple titratable side chains. Titration events at each protonation site — the release of a hydrogen ion from an acidic side chain, or the reverse — are dependent on the titration state of the other sites. For four titratable residues of the protein barnase, we calculate the full partition function including every combination of the four residues' titration states. The predictions of this model are compared to experiment, and to the "null model" of completely independent titrating groups. Finally, we evaluate several simplifying approximations that have been used to make the evaluation of exponentially larger numbers of titratable residues tractable.

In Chapter 4, we introduce the analytical continuum electrostatics (ACE) method [6, 7], which is a computationally inexpensive approximation to the more costly finite-difference solution of the Poisson–Boltzmann equation. Such an approximation has value for applications that require large numbers of electrostatic free energy calculations, such as multiple site titration, minimization, and protein design. We identify and mitigate a primary source of error for the ACE method, and incorporate a treatment of non-zero ionic strength.

In Chapter 5, we develop and test a procedure which optimizes the parameters of the ACE method in order to minimize its errors as compared to finite-difference Poisson–Boltzmann results. Optimized parameter sets give significantly lower errors in components of the electrostatic binding free energy. Only in some cases does the transfer of an optimized parameter set from a "training" to a "testing" protein binding pair result in reduced errors, but training and testing within a family of eight variant Zif268 protein/DNA binding pairs achieves an average reduction of -25% for our error function.

In Chapter 6, we develop a protein design method which uses discrete search algorithms to systematically search extremely large numbers (10^{24}) of structures, varying the amino acid type and the conformation at each selected protein residue. Our design method has these novel features: (1) it can optimize the binding free energy while maintaining a stable folding free energy, (2) it uses three stages of increasingly detailed discrete search in order to evaluate a diversity of sequences as well as multiple structures for each sequence, (3) it includes more accurate solvation and electrostatic free energy terms than methods previously used with discrete searches, and (4) it uses a hierarchy of three energy functions to successively screen candidate structures identified by the discrete search. The three hierarchical energy functions apply Coulombic, ACE, and finite-difference Poisson–Boltzmann electrostatics, respectively. We apply our design method to three residues of the protein barstar in order to enhance its binding to its partner barnase; then to three residues of gp41; and finally to seven residues of barstar. Our method ranks the wild-type barstar sequence very highly, which validates the method because barstar is experimentally known to bind very tightly to barnase. Several novel sequences predicted to bind more tightly than wild-type barstar are promising candidates for synthesis.

Chapter 2

Theory

2.1 Statistical Mechanics of Solvent, Solutes, and Mobile Ions

In this section, we will develop the theory necessary to calculate free energy differences for molecular processes taking place in solution, such as the formation of a binding complex by solute molecules. Other processes of interest are protein folding and conformational changes of a single solute molecule. These and all other processes of interest are described by changes of the free energy, rather than absolute free energies. The physical system of interest is a solution consisting of N_v solvent molecules, N_u molecules of each solute species in the list $\{u\}$, and N_m molecules of each mobile ion species in the list $\{m\}$. Each solute species u in the list $\{u\}$ can be a single molecule or a bound complex of molecules.

The $(\{N_i\}, P, T)$ ensemble, in which the number of each species of molecule, the pressure, and the temperature are held constant, corresponds to the usual situation in biological systems. However, at constant pressure and sufficiently large volume, the $(\{N_i\}, V, T)$ ensemble gives a statistical mechanical description of the system identical to that of the $(\{N_i\}, P, T)$ ensemble, because the average fluctuation $\overline{\Delta V}$ of the volume about its equilibrium value V_0 is negligible [8]. This will allow us to compute Helmholtz free energies

$$F(\{N_i\}, V, T) = -k_{\mathrm{B}}T \ln \mathcal{Q}(\{N_i\}, V, T) \tag{2.1}$$

rather than Gibbs free energies

$$G(\{N_i\}, P, T) = -k_{\mathrm{B}}T \ln \mathcal{Q}(\{N_i\}, P, T) \tag{2.2}$$

because, for a sufficiently large volume and a constant pressure,

$$G = F + PV_0 \tag{2.3}$$

Therefore, we can use the more convenient $(\{N_i\}, V, T)$ ensemble when developing the partition function.

The full quantum mechanical partition function $\mathcal{Q}$ of the solution (containing one solvent species, various mobile ion species, and various solute molecule species) is

$$\mathcal{Q} = \sum_{\{x\}} e^{-\beta \mathcal{H}(\{x\})} \tag{2.4}$$

where $\mathcal{H}$ is the exact quantum mechanical Hamiltonian for the system, and $\{x\}$ are all degrees of freedom of the system.

We assume that there are no interactions between solute molecules except in such binding complexes. That is, each solute molecule is free in solution but is not affected by the other solute molecules. This is true in the limit of low solute concentrations $(\sum_u N_u \ll N_v)$, and is called the dilute solution limit [9]. The dilute solution limit also implies that the solute molecules do not affect the bulk structure of the solvent.

In the dilute solution limit, and assuming that molecules of a given species are indistinguishable, the Hamiltonian is separable into $\mathcal{H}^v$ for the solvent-solvent interactions (including ions), $\mathcal{H}^u$ for the internal states of each solute molecule j of each species u, and $\mathcal{H}^{u,v}$ for the interactions of each solute molecule of each species u

with the region of solvent (and ions) around it [8]:

$$\mathcal{H}(\{x\}) = \mathcal{H}^v(\{x^v\}) + \sum_u \sum_j \mathcal{H}^u(\{x^u_j\}) + \sum_u \sum_j \mathcal{H}^{u,v}(\{x^v, x^u_j\}) \tag{2.5}$$

where $\{x^u_j\}$ are the degrees of freedom of one solute molecule, number j of species u. This allows the separation of the partition function into factors $\mathcal{Q}^v$ for the solvent-solvent interactions (including ions) and $\mathcal{Q}^u_0$ for a single solute molecule of species u together with an arbitrarily large region of solvent (and ions) surrounding it:

$$\mathcal{Q} = \mathcal{Q}^v \prod_u \frac{1}{n_u!} (\mathcal{Q}^u_0)^{n_u} \tag{2.6}$$

where

$$\mathcal{Q}^v = \sum_{\{x^v\}} e^{-\beta \mathcal{H}^v} \tag{2.7}$$

and

$$\begin{aligned} \mathcal{Q}^u_0 &= \sum_{\{x^u\}} \frac{\sum_{\{x^v\}} \left(e^{-\beta \mathcal{H}^v}\right) e^{-\beta \mathcal{H}^{u,v}} e^{-\beta \mathcal{H}^u}}{\sum_{\{x^v\}} \left(e^{-\beta \mathcal{H}^v}\right)} \\ &= \left\langle \sum_{\{x^u\}} e^{-\beta \mathcal{H}^{u,v}} e^{-\beta \mathcal{H}^u} \right\rangle_v \end{aligned} \tag{2.8}$$

where the notation $\langle\, \rangle_v$ is an expectation value over all solvent states, Boltzmann-weighted by the term $e^{-\beta \mathcal{H}^v}$ for the solvent-solvent interactions.

At this point, it is convenient to explicitly allow conformational freedom to the solute molecule, by allowing a solute molecule of species u to have a set of rigid conformations, $\{x^{u,i}\}$ where i is the conformation number, and the allowed values of i are denoted $\{i_u\}$. A rigid conformation is a fixed internal geometry of the molecule's atoms, which can be defined by specifying values for the bond lengths and angles (which are typically fixed), and dihedral angles. Nuclear, electronic, and vibrational degrees of freedom, because they have much higher frequency modes than large changes of dihedral angles, are assumed to be separable in the Hamiltonian. Let us separate those degrees of freedom, as well as the 6 translational and rotational degrees of freedom for the whole molecule, out of the solute intramolecular Hamiltonian:

$$\mathcal{H}^u = \mathcal{H}^{u,\text{nuc+elec+vib}} + \mathcal{H}^{u,\text{rot}} + \mathcal{H}^{u,\text{trans}} + \mathcal{H}^{u,\text{conf}} \tag{2.9}$$

Since we assume these terms are independent of the solvent, they come out of the expectation value in Equation 2.8:

$$\mathcal{Q}_0^u = \mathcal{Z}^{u,\text{trans}} \sum_{\{i_u\}} \mathcal{Z}^{u,i,\text{int}} e^{-\beta \mathcal{H}^{u,i,\text{conf}}} \left\langle e^{-\beta \mathcal{H}^{u,i,v}} \right\rangle_v \tag{2.10}$$

where

$$\mathcal{Z}^{u,\text{trans}} = V \left(\frac{2\pi m_u k_{\text{B}} T}{h^2} \right)^{3/2} \tag{2.11}$$

is the translational partition function of solute species u, which has molecular mass m_u, and h is Planck's constant; and $\mathcal{Z}^{u,i,\text{int}}$ is the internal partition function for conformation i of solute species u, including nuclear, electronic, rotational, and vibrational degrees of freedom.

Let us define the eigenvalue $G_0^{u,i} = \mathcal{H}^{u,i,\text{conf}}$ as the internal conformational energy of a molecule of solute species u in conformation i.

We also identify the eigenvalue of the average solute-solvent interaction energy of a molecule of solute species u in conformation i with the solvation chemical potential:

$$\Delta\mu_{\text{solv}}^{u,i} = -k_{\text{B}} T \ln \left\langle e^{-\beta \mathcal{H}^{u,i,v}} \right\rangle_v \tag{2.12}$$

This chemical potential is the free energy of transfer of one solute molecule of species u in conformation i from vacuum into the solvent. Re-writing the partition function, we have

$$\mathcal{Q}_0^u = \mathcal{Z}^{u,\text{trans}} \sum_{\{i_u\}} \mathcal{Z}^{u,i,\text{int}} e^{-\beta\left(G_0^{u,i} + \Delta\mu_{\text{solv}}^{u,i}\right)} \tag{2.13}$$

Finally, from Equations 2.6, 2.3, and 2.13, and defining $G^v = PV_0 + k_{\text{B}} T \ln \mathcal{Q}^v$, the Gibbs free energy of the entire solution is

$$G = G^v + k_{\mathrm{B}}T \sum_u \left[\ln n_u! - n_u \ln \mathcal{Z}^{u,\mathrm{trans}} - n_u \ln \left(\sum_{\{i_u\}} \mathcal{Z}^{u,i,\mathrm{int}} e^{-\beta\left(G_0^{u,i} + \Delta\mu_{\mathrm{solv}}^{u,i}\right)} \right) \right] \tag{2.14}$$

2.1.1 Covalent, Non-polar, Electrostatic Energy Components

The internal conformational free energy $G_0^{u,i}$ of a molecule of solute species u in conformation i includes the bond, angle, and torsion (dihedral) strain, electrostatic, and van der Waals energies internal to the solute molecule. In practice, $G_0^{u,i}$ can be calculated using quantum mechanics or with an empirical molecular force field such as the PARAM19 [10] parameter set used with the molecular modeling package CHARMM [11]. By "electrostatic", we mean the electrostatic energy of the average charge distribution of the molecule. The van der Waals interaction of induced dipoles, as well as the effect of the Pauli exclusion principle, can be calculated in practice with the Lennard–Jones 6–12 potential. The method of calculating the electrostatic energy will be discussed in Section 2.1.2. The electrostatic component of $G_0^{u,i}$ includes only interactions internal to the solute molecule, not the screening effect of the solvent, so it is equivalent to the energy of the solute molecule in vacuo. Aside from the electrostatic and van der Waals components, empirical parameter sets model the rest of the true quantum mechanical Hamiltonian with force constants for bond, angle, and torsional strain energy between the covalently bonded atoms.

Assume that the Hamiltonian can be separated into electrostatic and non-polar terms:

$$\mathcal{H}^{u,i,v}(\{x^{u,i}\},\{x^v\}) = \mathcal{H}^{u,i,v}_{\mathrm{non-polar}}(\{x^{u,i}\},\{x^v\}) + \int \rho_{u,i}(\vec{x})\Phi^v(\vec{x})d\vec{x} \tag{2.15}$$

where $\rho_{u,i}(\vec{x})$ is the average charge distribution of the solute molecule, and $\Phi^v(\vec{x})$ is the net electrostatic potential of a particular configuration $\{x^v\}$ of the solvent. Now we can also split the solvation chemical potential $\Delta\mu_{\mathrm{solv}}^{u,i}$ into electrostatic and non-polar terms:

$$\Delta\mu_{\mathrm{solv}}^{u,i} = \Delta\mu_{\mathrm{solv,\ non-polar}}^{u,i} + \Delta\mu_{\mathrm{solv,\ ES}}^{u,i} \tag{2.16}$$

where we let the non-polar term be the solvation chemical potential for a hypothetical solute molecule in the same configuration $\{x^{u,i}\}$ but with its charge $\rho_{u,i}(\vec{x})$ set to zero:

$$\Delta\mu^{u,i}_{\mathrm{solv,\ non-polar}} = -k_{\mathrm{B}}T \ln \left\langle e^{-\beta \mathcal{H}^{u,i,v}_{\mathrm{non-polar}}(\{x^{u,i}\},\{x^{v}\})} \right\rangle_{v} \tag{2.17}$$

This can also be thought of as the effect of creating a "cavity" in the solvent with the shape of the solute molecule. It is difficult to compute this term carefully; in fact, how much of it comes from van der Waals, electrostatics, and entropy of the solvent molecules is still not fully understood [12, 13, 14]. In practice it has been found, from experimental free energies of transfer of hydrocarbons from a non-polar hydrocarbon environment into water, that this "cavity" or "hydrophobic" term scales roughly with the solvent-accessible surface area (SAS) of the solute molecule [12, 15, 16, 17]. (The SAS is defined as the locus of points of closest approach of the center of a water-molecule-size sphere (1.4 Å) to the solute, as the sphere is rolled over the union of the van der Waals radius volumes of the solute atoms [18].)

Our definition of $\Delta\mu^{u,i}_{\mathrm{solv,\ non-polar}}$ means that the electrostatic term must be expressed as an average weighted by the non-polar free energy,

$$\Delta\mu^{u,i}_{\mathrm{solv,\ ES}} = -k_{\mathrm{B}}T \ln \left[\frac{\left\langle e^{-\beta(\mathcal{H}^{u,i,v}_{\mathrm{non-polar}}(\{x^{u,i}\},\{x^{v}\}) + \int \rho_{u,i}(\vec{x})\Phi^{v}(\vec{x})d\vec{x})} \right\rangle_{v}}{\left\langle e^{-\beta \mathcal{H}^{u,i,v}_{\mathrm{non-polar}}(\{x^{u,i}\},\{x^{v}\})} \right\rangle_{v}} \right] \tag{2.18}$$

We can simplify the notation by defining the operator $\langle\ \rangle_{v,uv\ \mathrm{non-polar}}$ as the expectation value over the solvent states, Boltzmann-weighted by the solvent-solvent and the non-polar solute-solvent interactions. States in which solvent atoms overlap with solute atoms are suppressed by this weighting. So,

$$\Delta\mu^{u,i}_{\mathrm{solv,\ ES}} = -k_{\mathrm{B}}T \ln \left\langle e^{-\beta \int \rho_{u,i}(\vec{x})\Phi^{v}(\vec{x})d\vec{x}} \right\rangle_{v,uv\ \mathrm{non-polar}} \tag{2.19}$$

Assuming that the charge distribution of the solute molecule does not depend on the solvent configuration,

$$\Delta\mu^{u,i}_{\text{solv, ES}} = -k_{\text{B}}T \ln \left\langle e^{-\beta \int \rho_{u,i}(\vec{x})\Phi^v(\vec{x})d\vec{x}} \right\rangle_{v,uv \text{ non-polar}} \tag{2.20}$$

Consider the process of "turning on" the solute molecule charge distribution $\lambda\, \rho_{u,i}(\vec{x})$ by gradually raising λ from 0 to 1. We will use the functional derivative of $\Delta\mu^{u,i}_{\text{solv, ES}}$ from Equation 2.20 with respect to $\rho_{u,i}(\vec{x})$, which is

$$\begin{aligned} \frac{\partial \Delta\mu^{u,i}_{\text{solv, ES}}}{\partial \rho_{u,i}(\vec{x})} &= \frac{\left\langle \Phi^v(\vec{x}) e^{-\beta \int \rho_{u,i}(\vec{x})\Phi^v(\vec{x})d\vec{x}} \right\rangle_{v,uv \text{ non-polar}}}{\left\langle e^{-\beta \int \rho_{u,i}(\vec{x})\Phi^v(\vec{x})d\vec{x}} \right\rangle_{v,uv \text{ non-polar}}} && (2.21) \\ &= \frac{\sum_{\{x_v\}} \Phi^v(\vec{x}) e^{-\beta(\mathcal{H}^v + \mathcal{H}^{u,v})}}{\sum_{\{x_v\}} e^{-\beta(\mathcal{H}^v + \mathcal{H}^{u,v})}} && (2.22) \end{aligned}$$

We are treating one solute molecule of species u in fixed conformation i, so the only degrees of freedom are for the solvent $\{x_v\}$. We can still multiply on the top and bottom by $e^{-\beta\mathcal{H}^u}$ to make it plain that the sums are based on the full Hamiltonian, and therefore the expression is the expectation value of $\Phi^v(\vec{x})$:

$$\begin{aligned} \frac{\partial \Delta\mu^{u,i}_{\text{solv, ES}}}{\partial \rho_{u,i}(\vec{x})} &= \frac{\sum_{\{x_v\}} \Phi^v(\vec{x}) e^{-\beta\mathcal{H}}}{\sum_{\{x_v\}} e^{-\beta\mathcal{H}}} && (2.23) \\ &= \langle \Phi^v(\vec{x}) \rangle && (2.24) \end{aligned}$$

Let us gradually turn on the solute molecule charge distribution by gradually raising λ from 0 to 1. Let $\langle \Phi^v(\vec{x};\lambda) \rangle$ be the expectation value of the electrostatic potential at $\vec{x}$ due to the solvent when the solute charge distribution is $\rho_{u_i}(\vec{x};\lambda) = \lambda\rho_{u,i}(\vec{x})$. Let the functional variation be simply $\partial\rho_{u_i}(\vec{x};\lambda) = \rho_{u_i}(\vec{x})\partial\lambda$. Now we can integrate along λ from 0 to 1:

$$\Delta\mu^{u,i}_{\text{solv, ES}} = \int \mathrm{d}\vec{x} \rho_{u_i}(\vec{x}) \int_0^1 \mathrm{d}\lambda \langle \Phi^v(\vec{x};\lambda) \rangle \tag{2.25}$$

2.1.2 Continuum Electrostatics

Let us first consider the solvent molecules apart from the ions, both of which contribute to $\langle \Phi^v(\vec{x};\lambda)\rangle$. The sum over solvent degrees of freedom implied in $\langle \Phi^v(\vec{x};\lambda)\rangle$ means that this expectation value of the electrostatic potential due to the solvent will, for a polar solvent such as water, oppose the electrostatic potential due to the solute. This is the behavior of a polarizable medium, i.e. a dielectric. So, to calculate $\langle \Phi^v(\vec{x};\lambda)\rangle$, we will employ a continuum dielectric model. In this model, the solvent is no longer treated as explicit atoms; it is treated as a dielectric medium of high dielectric constant $\epsilon_{\rm s}$. Recall that we are treating one solute molecule infinitely diluted in the solvent, so we can divide all space into two regions, solute and solvent, whose boundary surface (called the molecular surface) is defined as the locus of points of closest approach of the surface of a water-molecule-size sphere (1.4 Å) to the solute as the sphere is rolled over the union of the van der Waals radius spheres of the solute atoms. The solute region is treated as a region of low dielectric constant ϵi containing the charge distribution $\rho_{u,i}(\vec{x})$. We let this charge distribution consist of partial atomic point charges at the centers of the solute atoms, as parametrized in the PARAM19 [10] parameter set of CHARMM [11]. We previously assumed that (1) the electronic and vibrational parts of the Hamiltonian are independent of the electrostatic part, and (2) a discrete conformation i does not deform due to the electric field. A convenient way to relax these assumptions, which we use, is to assign the solute region a dielectric constant $\epsilon_{\rm i} > 1$, to model movement of the solute in response to the electric field.

The total electrostatic potential $\langle \Phi(\vec{x};\lambda)\rangle$ is given by the Poisson equation [19],

$$\nabla \cdot (\epsilon(\vec{x})\nabla \langle \Phi(\vec{x};\lambda)\rangle) = -4\pi\lambda\rho_{u,i}(\vec{x}) \tag{2.26}$$

where

$$\epsilon(\vec{x}) = \begin{cases} \epsilon_{\rm i} \text{ if } \vec{x} \text{ is in the solute region,} \\ \epsilon_{\rm s} \text{ otherwise} \end{cases} \tag{2.27}$$

The total electrostatic potential includes the potential due to the solvent $\langle \Phi^v(\vec{x};\lambda)\rangle$, and the Coulombic potential of the fixed solute point charges,

$$\langle\Phi(\vec{x};\lambda)\rangle = \langle\Phi^v(\vec{x};\lambda)\rangle + \lambda\int \mathrm{d}x' \frac{\rho_{u,i}(\vec{x}')}{\epsilon_\mathrm{i}|\vec{x}-\vec{x}'|} \tag{2.28}$$

for $\vec{x}$ inside the solute volume. Solution of the Poisson equation (Equation 2.26) and use of Equation 2.28 to remove the potential due to the solute charges yields the electrostatic potential $\Phi^v(\vec{x};\lambda)$ due to solvent.

Now we will show how the mobile ions of species $\{m\}$ alter the calculation of $\langle\Phi^v(\vec{x};\lambda)\rangle$. For ion species m, let q_m be the charge of the ion, and let $c_m(\vec{x};\lambda)$ be its equilibrium concentration (i.e., number density). The charge distribution of the ions is

$$\rho_\mathrm{ion}(\vec{x};\lambda) = \sum_m q_m c_m(\vec{x};\lambda) \tag{2.29}$$

Assuming that the ionic concentration is small allows us to neglect the possibility of two ions overlapping, so that we may treat the ions in a mean-field sense. It also allows us to keep the dielectric constant constant at ϵ_s in the solvent, regardless of ionic concentration. With mobile ions, the Poisson equation becomes

$$\nabla\cdot(\epsilon(\vec{x})\nabla\langle\Phi(\vec{x};\lambda)\rangle) = -4\pi\lambda\rho_{u,i}(\vec{x}) - 4\pi\rho_\mathrm{ion}(\vec{x};\lambda) \tag{2.30}$$

We require the concentrations $c_m(\vec{x};\lambda)$ to be zero inside the Stern layer, defined as the locus of points of closest approach of the center of an ion-size sphere (2.0 Å, typically) to the solute, as the sphere is rolled over the union of the van der Waals radius volumes of all solute atoms [20, 21].

To calculate the mobile ion concentration $c_m(\vec{x};\lambda)$, require the chemical potential of mobile ion species m to be constant in $\vec{x}$:

$$\begin{aligned}\mu_m &= \mu_m(\vec{x};\lambda) &(2.31)\\ &= \mu_m^0 + k_\mathrm{B}T\ln a_m(\vec{x};\lambda) + q_m\langle\Phi(\vec{x};\lambda)\rangle &(2.32)\end{aligned}$$

where μ_m^0 is the standard chemical potential of mobile ion species m. We have made the first approximation of Debye-Hückel theory by using $q_m \langle \Phi(\vec{x};\lambda) \rangle$ as the potential of mean force [22]. The activity $a_m(\vec{x};\lambda)$ is assumed equal to the local ionic concentration [23]; if all ion species have low concentrations, this assumption is accurate. Most of the volume of the solvent is not near a solute molecule, so the concentration is the bulk concentration c_m^{b}:

$$\mu_m = \mu_m^0 + k_{\mathrm{B}} T \ln c_m^{\mathrm{b}} \tag{2.33}$$

Combining Equations 2.32 and 2.33, the concentration of mobile ion species m outside of solute molecules' Stern layers is

$$c_m(\vec{x};\lambda) = c_m^{\mathrm{b}} e^{-\beta q_m \langle \Phi(\vec{x};\lambda) \rangle} \tag{2.34}$$

So the net ionic charge density is

$$\rho_{\mathrm{ion}}(\vec{x};\lambda) = \sum_m q_m c_m^{\mathrm{b}} e^{-\beta q_m \langle \Phi(\vec{x};\lambda) \rangle} \tag{2.35}$$

Substituting this into the Poisson equation, Equation 2.30, gives the non-linear Poisson-Boltzmann equation:

$$\nabla \cdot \left(\epsilon(\vec{x}) \nabla \langle \Phi(\vec{x};\lambda) \rangle\right) = -4\pi\lambda\rho_{u,i}(\vec{x}) - 4\pi \sum_m q_m c_m^{\mathrm{b}} e^{-\beta q_m \langle \Phi(\vec{x};\lambda) \rangle} \tag{2.36}$$

If this equation is truly non-linear, there is no way to self-consistently describe free energies of the system's states.

The second Debye–Hückel approximation will allow us to linearize the Poisson-Boltzmann equation, by letting $q_m \langle \Phi(\vec{x};\lambda) \rangle \ll k_{\mathrm{B}} T$. This is true except for ion locations very near the solute molecule. With this approximation, we can expand the exponentials and only keep terms to the first order:

$$\rho_{\mathrm{ion}}(\vec{x};\lambda) = \sum_m q_m c_m^{\mathrm{b}} - \beta \sum_m q_m^2 c_m^{\mathrm{b}} \langle \Phi(\vec{x};\lambda) \rangle \tag{2.37}$$

Since the whole system must be approximately electrically neutral, and the infinite dilution of the solute means that their concentration is much lower than any bulk ion concentration, the first term vanishes. Now we define the modified Debye-Hückel screening parameter $\bar{\kappa}(\vec{x})$, and use its definition to impose the Stern layer:

$$\bar{\kappa}(\vec{x}) = \begin{cases} 0 & \text{inside the Stern layer} \\ \sqrt{\frac{8\pi I}{k_{\mathrm{B}}T}} & \text{elsewhere} \end{cases} \tag{2.38}$$

where $I = \frac{1}{2}\sum_m q_m^2 c_m^{\mathrm{b}}$ is the bulk ionic strength. Putting Equation 2.37 into the Poisson equation (Equation 2.26), we see that the second Debye–Hückel approximation has given us the linearized Poisson–Boltzmann equation:

$$\nabla \cdot (\epsilon(\vec{x})\nabla \langle \Phi(\vec{x};\lambda)\rangle) - \bar{\kappa}^2(\vec{x})\langle\Phi(\vec{x};\lambda)\rangle = -4\pi\lambda\rho_{u,i}(\vec{x}) \tag{2.39}$$

Solution of the linearized Poisson–Boltzmann equation, with the boundary condition $lim_{\vec{x}\to\infty}\langle\Phi(\vec{x};\lambda)\rangle = 0$, yields the electrostatic potential $\langle\Phi(\vec{x};\lambda)\rangle$. The linearity of Equation 2.39 requires that

$$\langle\Phi(\vec{x};\lambda)\rangle = \lambda\langle\Phi(\vec{x};\lambda = 1)\rangle \tag{2.40}$$

And so, since the solute-solute interactions in Equation 2.28 do not depend on λ,

$$\langle\Phi^v(\vec{x};\lambda)\rangle = \lambda\langle\Phi^v(\vec{x};\lambda = 1)\rangle \tag{2.41}$$

Using Equation 2.25, we return at last to the electrostatic component of the solvation chemical potential,

$$\Delta\mu^{u,i}_{\text{solv, ES}} = \frac{1}{2}\int \mathrm{d}\vec{x}\rho_{u_i}(\vec{x})\langle\Phi^v(\vec{x};\lambda = 1)\rangle \tag{2.42}$$

where the potential is obtained from solving the linearized Poisson–Boltzmann equation, Equation 2.39.

For the rest of this work, we will assume the continuum solvent model, so we adopt a simpler notation, $\Phi(\vec{x}) = \langle \Phi^v(\vec{x}; \lambda = 1) \rangle$ if clarity does not require the solute molecule species u and conformation i to be specified. For the electrostatic component of the solvation chemical potential, we will use the notation $\Delta G^{\text{solv}} = \Delta \mu^{u,i}_{\text{solv, ES}}$, where the free energy of solvation ΔG^{solv} of one molecule can be expressed in the units kcal mol^{-1}.

The electrostatic component of the solvation chemical potential ΔG^{solv} is a true free energy, not just an energy. The careful statistical mechanical treatment in this chapter shows why continuum electrostatic calculations result in free energies. Briefly, it is because the dielectric response of the solvent results from the expectation value of the electrostatic field produced by the solvent. This expectation value, in turn, results from the solvent populating its states according to a constant-temperature canonical ensemble. ΔG^{solv} comes from a sum on the solvent states, and so it is entropic as well as enthalpic in origin.

2.2 Finite-Difference Solution of the Poisson-Boltzmann Equation (FDPB)

To solve the linearized Poisson–Boltzmann equation by the finite-difference method (a method abbreviated as FDPB), a region of space is represented as a cubic grid of points [1, 2]. An iterative finite-difference algorithm is then used to solve for the electric potential at every grid point. All FDPB calculations in this work were done with a locally modified version of the program DELPHI [3, 4, 5].

First, boundary conditions must be set: the value of the electric potential must be set at each grid point on the outside boundary of the entire grid. One way to set approximate boundary conditions, which we use, is to superimpose the Debye-Hückel screened Coulombic electic potential (with ϵ_{s}) of every atomic point charge, and evaluate this potential at each boundary grid point.

Next, arrays representing the charge, ionic strength, and dielectric constant must be initialized. The charge distribution, which in our model is a collection of partial atomic point charges, must then be mapped onto the grid points. This is done by distributing the charge of each atom center over the 8 grid points of the cubic grid box surrounding the atom location, with the amount of charge assigned to each of the 8 grid points being determined by a trilinear function of the atom center's position within the grid box [1]. Next, every grid point must be assigned an ionic strength, and every grid line center (the midpoint of every line connecting 2 nearest-neighbor grid points) must be assigned a dielectric constant. As given in Equation 2.38, the ionic strength is set to the desired value of the bulk ionic strength at all grid points outside the Stern layer, and set to zero at all grid points inside the Stern layer. The solute dielectric constant ϵ_i is assigned to every grid line center inside the solvent-accessible surface (SAS) of the solute molecule, and the solvent dielectric constant ϵ_s is assigned to each grid line center outside the SAS. As mentioned in the previous section, the SAS is defined as the locus of points of closest approach of the center of a water-molecule-size sphere (1.4 Å) to the solute, as the sphere is rolled over the union of the van der Waals radius volumes of the solute atoms.

Once the potential, charge, ionic strength, and dielectric constant arrays have been initialized, a finite-difference procedure iteratively relaxes the potential at every grid point until the values converge to those that satisfy the finite-difference representation of the linearized Poisson–Boltzmann equation. To minimize the impact of the approximation used for the boundary conditions, we do several focussing steps, beginning with a grid box in which the solute molecule has a linear dimension only 23% of the edge length of the whole grid. A second focussing step, in which the solute molecule fills 92% of the grid's linear extent, can be the final step, or further focussing steps at greater than 100% fill can be done if the purpose of the calculation is to calculate electrostatic interactions of small groups of atoms rather than the electrostatic free energy of the whole solute molecule.

Finally, the electrostatic free energy is calculated as in Equation 2.42, a sum over all grid points of the charge times the potential.

In order to reduce the error caused by the discretization of the grid, this whole procedure is repeated ten times, with the grid translated by different fractions of the grid size in the three Cartesian directions. The results are then averaged, and their standard error is used to assess the accuracy of the result. A finer grid, with more grid points, will result in a better accuracy.

2.2.1 Calculating Solvation Free Energy with FDPB

To calculate an electrostatic solvation free energy with FDPB, we use the thermodynamic process shown in Figure 2-1. The FDPB procedure is run once on the solvated molecule and again on the molecule in the desolvated state. We define the "desolvated" state to have $\epsilon_{\mathrm{i}} = \epsilon_{\mathrm{s}} = 4$ and zero ionic concentration in all space. The electrostatic solvation free energy is the straightforward free energy difference of the solvated and desolvated states:

$$\Delta G_{\mathrm{ES}}^{\mathrm{solv}} = G_{\mathrm{ES}}^{\mathrm{solvated}} - G_{\mathrm{ES}}^{\mathrm{desolvated}} \tag{2.43}$$

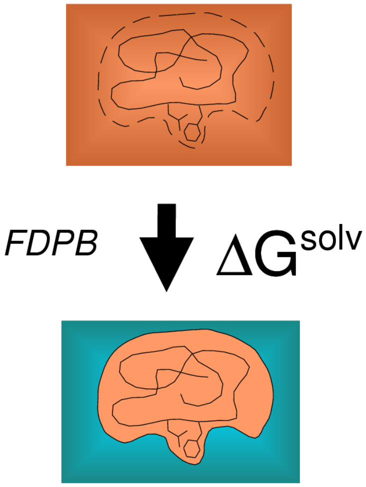

Figure 2-1: Thermodynamic process used to get an electrostatic solvation free energy using FDPB.

2.2.2 Calculating Free Energy Differences with FDPB

To calculate an electrostatic free energy difference of 2 conformations, we use the thermodynamic cycle shown in Figure 2-2. This cycle is necessary because every FDPB free energy result includes a fictitious "grid energy,"

$$G_{\text{ES, FDPB}} = G_{\text{ES}} + G_{\text{grid}} \tag{2.44}$$

which depends on each point charge's magnitude, location relative to the finite-difference

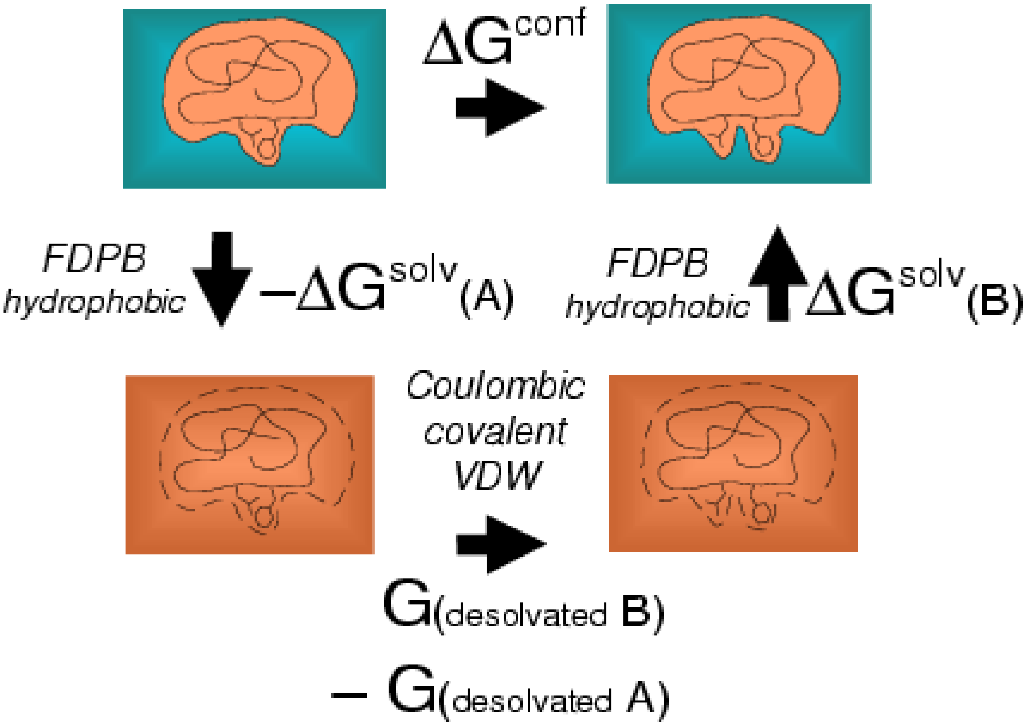

Figure 2-2: Thermodynamic cycle used to get an electrostatic free energy difference for a conformational change using FDPB for the solvation free energies and Coulomb's Law for the energy difference of the conformations when desolvated. To include non-electrostatic energy terms, van der Waals and covalent energy terms would be part of the bottom arrow of the cycle, and a hydrophobic term would be part of each of the solvation free energies. But in practice, the non-electrostatic terms are all just straightforward differences of the 2 conformations.

grid, and local dielectric constant. Recall that the FDPB procedure distributes each partial atomic point charge over the 8 surrounding grid points. The grid energy is partially an artifact of this procedure: it is the interaction of these 8 charges with each other, as well as the self-energy of each point charge. The self-energy of a point charge, which is actually infinite, has a finite value in the FDPB treatment because the solution of the Poisson–Boltzmann equation's finite-difference form has a finite potential at the location of a point charge.

To ensure that all FDPB calculations use the exact same position of the finite-difference grid relative to the solute atoms, "dummy" atoms with zero charge and zero radius are placed at 2 opposite corners of a box large enough to contain all conformations of interest. The electrostatic solvation free energy of one conformation, as given in Equation 2.43, may be calculated straightforwardly using FDPB, because the grid energy is the same in the solvated and desolvated states and therefore cancels out. So the free energy difference of 2 conformations "A" and "B" is calculated using FDPB for the solvation free energies of each conformation, and Coulomb's Law for the energy difference of the conformations when desolvated:

$$\Delta G_{\mathrm{ES}}(B-A) = \left[\Delta G^{\mathrm{solv}}_{\mathrm{ES,\ FDPB}}(B) - \Delta G^{\mathrm{solv}}_{\mathrm{ES,\ FDPB}}(A)\right] + \Delta G^{\mathrm{desolvated}}_{\mathrm{ES,\ Coul}}(B-A) \qquad (2.45)$$

This is the standard way to calculate an electrostatic free energy difference of 2 conformations. Binding, folding, and titration events can be considered as special types of conformational change. Thermodynamic cycles used to calculate electrostatic free energies for these events are described in Section 3.2.6 of Chapter 3 and Section 6.3.5 of Chapter 6.

Chapter 3

Multiple Site Titration: Effects of Approximations

3.1 Introduction

Some protein side chains are titratable; that is, they can bind or release a hydrogen ion. For small acid and base molecules in aqueous solution — including single amino acids such as aspartate, glutamate, and histidine — the probability of having a hydrogen ion bound depends only on the pH of the solution. The pK_a is the pH at which the probability of protonation is 50%; that is, the free energy difference of the protonated and unprotonated states is zero. In a folded protein, the pK_a of each titratable side chain is shifted, because the protein environment of the protonation site is more or less electrostatically favorable for the protonated relative to the unprotonated state.

The exponential dependence of the number of titration states on the number of titrating residues makes it computationally intractable to calculate the continuum electrostatic free energy of each titration state for most proteins. Various simplifying approximations can be employed to make the problem computationally tractable. In this chapter, we evaluate several such simplifications, some of which have been used by previous investigators to calculate effective pK_a values. Taking titration into account when calculating potentially pH-dependent properties, such as folding or binding free energies, presents the same problem. Only a few studies [24, 25] have dealt with this important problem.

We use a system (the binding of barnase to barstar, with the 4 barnase residues Asp54, Glu73, Asp75, His102 allowed to titrate) with few enough ionization states so that enumeration and energy calculation of all titration states is possible, but enough that interactions between the titratable residues can be explored. By calculating the free energy of each titration state we can evaluate the effect of each simplifying approximation. More titratable residues can be treated by methods other than full enumeration of the states: Alexov and Gunner [24] used a titration model similar to ours, but with a Monte Carlo method to avoid full enumeration.

A natural benchmark for evaluating the accuracy of titration models is the "null model" — that is, the assumption that every titratable residue in a protein has the same pK_a as it would free in solution — using an internal protein dielectric constant $\epsilon_{\mathrm{i}} = 4$. Some investigators [26] have roughly matched the accuracy of the null model by using $\epsilon_{\mathrm{i}} = 20$, but we feel it is important to point out that raising ϵ_{i} may simply push inaccurate pK_a shifts closer to zero, rather than calculating the pK_a shifts with a more accurate method. In any attempt to consistently compute values more accurately than the null model, improved modeling of the protonation states' charge distributions and conformations will be more useful than trying to pick the "best" ϵ_{i}. We have chosen to use the usual value $\epsilon_{\mathrm{i}} = 4$ in the present study, so that various methods can be more clearly compared; using a higher ϵ_{i} would tend to wash out the differences between all methods. The rationale is that a low internal dielectric is appropriate if the individual states are being accurately enumerated and quantified; higher values might only be necessary to account approximately for inaccuracies in sampling and energetics.

For binding free energy calculations, the effect of ignoring titration (i.e., assuming the presumed predominant titration state is the only one populated) depends case-by-case on whether any residues near the binding interface have significant pK_a shifts. Some simplifications, such as the use of a single dielectric boundary for all titration states, may also cause systematic errors in binding free energy regardless of the predicted effective pK_a values.

3.2 Theory and Methods

3.2.1 Introduction to the Chemistry of Titration

An arbitrary acid AH, in a dilute aqueous solution, may dissociate into species A^- and H^+ (a hydrogen ion). The equilibrium between the dissociation and the reverse (association) processes can be represented:

$$\mathrm{AH} \rightleftharpoons \mathrm{A}^- + \mathrm{H}^- \tag{3.1}$$

The free energy of the solution can be expressed in terms of the chemical potentials μ_X and number populations n_X of the chemical species X:

$$G = \sum_\mathrm{X} \mu_\mathrm{X} n_\mathrm{X} \tag{3.2}$$

Let the free energy in equilibrium be at a minimum with respect to a small number dξ of the AH molecules dissociating through the chemical reaction in Equation 3.1:

$$\begin{aligned} \mathrm{d}G &= \mu_{\mathrm{A}^-}\mathrm{d}n_{\mathrm{A}^-} + \mu_{\mathrm{H}^+}\mathrm{d}n_{\mathrm{H}^+} + \mu_\mathrm{AH}\mathrm{d}n_\mathrm{AH} \\ &= (\mu_{\mathrm{A}^-} + \mu_{\mathrm{H}^+} - \mu_\mathrm{AH})\mathrm{d}\xi \end{aligned} \tag{3.3}$$

$$\left(\frac{\partial G}{\partial \xi}\right)_{p,T} = \mu_{\mathrm{A}^-} + \mu_{\mathrm{H}^+} - \mu_\mathrm{AH} \tag{3.4}$$

$$0 = \mu_{\mathrm{A}^-} + \mu_{\mathrm{H}^+} - \mu_\mathrm{AH} \tag{3.5}$$

For an ideal dilute solution, the chemical potential of each species X is related to its mole fraction x_X by:

$$\mu_\mathrm{X} = \mu_\mathrm{X}^* + RT \ln x_\mathrm{X} \tag{3.6}$$

For a non-ideal solution, we can define the activity a_{X} to take the place of the mole fraction x_{X} in Equation 3.6:

$$\mu_{\mathrm{X}} = \mu_{\mathrm{X}}^{*} + RT \ln a_{\mathrm{X}} \tag{3.7}$$

For an ideal dilute solution, the activity is simply $a_{\mathrm{X}} = x_{\mathrm{X}}$. We can define an equilibrium constant for the reaction of Equation 3.1 by combining Equations 3.5 and 3.7:

$$K_a'(\mathrm{A}^-) = \frac{a_{\mathrm{A}}^- a_{\mathrm{H}}^+}{a_{\mathrm{AH}}} \tag{3.8}$$

Let [X] be the unitless concentration of chemical species X (with the units (mol L^{-1}) divided out). This concentration [X] is proportional to the mole fraction x_{X}, so for an ideal dilute solution, we define the equilibrium constant as:

$$K_a(\mathrm{A}^-) = \frac{[\mathrm{A}^-][\mathrm{H}^+]}{[\mathrm{AH}]} \tag{3.9}$$

At constant temperature and pressure, the fraction of deprotonated A^- to protonated AH must be determined by the free energy difference of the 2 microstates, which is the free energy of protonation:

$$\Delta G_p(\mathrm{A}^-) \equiv G(\mathrm{AH}) - G(A^- + \mathrm{H}^+) \tag{3.10}$$

$$\frac{[\mathrm{AH}]}{[\mathrm{A}^-]} = \exp(-\beta \Delta G_p(\mathrm{A}^-)) \tag{3.11}$$

where $\beta \equiv 1/k_{\mathrm{B}}T$, k_{B} is Boltzmann's constant, and T is the temperature in Kelvin. The concentration of hydrogen ions in solution $[\mathrm{H}^+]$ (with the units (mol L^{-1}) divided out) defines the pH :

$$\mathrm{pH} \equiv -\log_{10}[\mathrm{H}^+] \tag{3.12}$$

and we similarly define the constant pK_a from the equilibrium constant K_a:

$$\mathrm{p}K_a(\mathrm{A}^-) \equiv -\log_{10} K_a(\mathrm{A}^-) \tag{3.13}$$

Combining Equations 3.9, 3.11, 3.12, and 3.13,

$$\mathrm{p}K_a(\mathrm{A}^-) = +\frac{1}{\ln(10)}\left[-\beta\Delta G_p(\mathrm{A}^-)\right] + \mathrm{pH} \tag{3.14}$$

$$\Delta G_p(\mathrm{A}^-) = +\ln(10)\cdot k_{\mathrm{B}}T\left[\mathrm{pH} - \mathrm{p}K_a(\mathrm{A}^-)\right] \tag{3.15}$$

The pK_a values for fully solvated titratable residues have been experimentally measured: 4.0 for Asp, 4.4 for Glu, 6.3 for His [27]. Approximately the same results are obtained for the pK_a values of blocked single amino acids or short polypeptides such as Ala-"R"-Ala, where "R" is the titratable amino acid type [26, 28, 29, 30]. Residues of denatured proteins in solution are generally found to have the same pK_a values as in short poly- or mono-peptides, although some have been found to differ by about 0.4 pH units [31]. Related but different compounds can have significantly different pK_a values (e.g. 3.75 for formic acid, 4.75 for acetic acid, 4.0 for aspartic acid, and 4.4 for glutamic acid). We take the experimental pK_a value of amino acid type "R" to be the pK_a in solution of a specific model compound, the blocked amino acid N-acetyl-"R" methylamide (CH_3-(CO)-(NH)-(C_α"R")-(CO)-(NH)-CH_3):

$$\mathrm{p}K_a(\text{model Asp}) = 4.0 \tag{3.16}$$

$$\mathrm{p}K_a(\text{model Glu}) = 4.4 \tag{3.17}$$

$$\mathrm{p}K_a(\text{model His}) = 6.3 \tag{3.18}$$

We let the model compound conformation be the same as the titrating protein residue (as in references [27] and [32]). This means that 2 residues of the same amino acid type each have their own model compound conformation.

To calculate pK_a values for a specific protein residue, consider the whole protein to be the same model compound with the rest of the protein attached, and call the protein protonation states PA$^-$ and PAH. Consider the thermodynamic cycle in Figure 3-1: the protonation free energies of the model compound and the residues in the protein (the horizontal arrows) are related by the free energy differences of the vertical arrows, which are, to within a constant that is the same for both sides of the cycle, folding free energies of the protein in the protonated and deprotonated states. Use Equation 3.15 to write the pK_a of the protein residue in terms of the model compound pK_a (A$^-$) :

$$pK_a(PA^-) = pK_a(A^-) - \frac{1}{\ln(10)}\beta\left[\Delta G_p(PA^-) - \Delta G_p(A^-)\right] \tag{3.19}$$

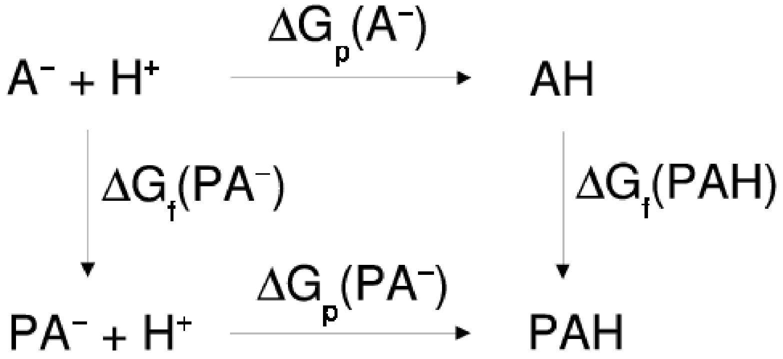

Figure 3-1: Thermodynamic cycle connecting pK_a of a residue in a protein to the experimental pK_a of a model compound (in our case, N-acetyl-“R” methylamide for amino acid type “R”).

3.2.2 Titration Microstates

Carboxylic acids include acetic acid, formic acid, and the protein side chains aspartic acid and glutamic acid. The titratable moiety is the carboxy group (COOH, or COO$^-$ when deprotonated). The deprotonated state, charged -1, has only one microstate. The protonated state, charged 0, has 4 microstates, because the one hydrogen ion can go on either of the 2 oxygen atoms, and there are 2 stable positions the hydrogen ion can occupy relative to its oxygen atom. Quantum mechanical simulations show that the valence electrons of this system are sp^2 hybridized, so the hydrogen atom stays in the plane of the COO atoms, and the O-H bond makes an angle of about 120° from the C-O bond. The preferred position, closer to the other oxygen, is called the "syn" position, and the position further from the other oxygen is called the "anti" position.

Histidine has an imidazole moiety, -NH-CH=N-CH=CH- where the first and last atoms are bonded to form a ring. To form histidine from imidazole, the last hydrogen is removed and the ring is attached, via one CH_2 group, to the protein backbone. Either one of the nitrogens, or both, can be protonated. A hydrogen bonded to either nitrogen has only one stable position. So, histidine has two neutral singly-protonated microstates and one +1 charged, doubly-protonated microstate. Also, because protein crystal structures do not currently differentiate nitrogen atoms from carbon atoms, it is not clear which way histidine rings in protein crystal structures are flipped. Therefore, we double the number of microstates by allowing the ring to be flipped either way, for a total of 4 neutral singly-protonated microstates and two +1 charged, doubly-protonated microstates.

We are using the continuum solvent model with the CHARMM [11] PARAM19 [10] united-atom parameter set. All heavy (non-hydrogen) atoms, and polar hydrogen atoms, of the solute molecule are explicitly represented, with a partial atomic charge and a van der Waals radius. Non-polar hydrogen atoms (for example, all hydrogens on a hydrophobic side chain like valine, $(CH_3)_2$-CH-) are united with the heavy atoms to which they are bonded rather than being represented explicitly. Different radii are used for CH, CH_2, and CH_3 united atoms. Titratable hydrogens are polar and therefore are always represented explicitly. The parameter set specifies partial atomic charges for the deprotonated and protonated states: Aspartate and glutamate (deprotonated aspartic and glutamic acid) end in the carboxyl group -C-COO^- (i.e., -(CH_2)-COO^-); from here on, we will often not refer to the non-explicit hydrogens). The -C-COO^- atoms are given partial atomic charges of -0.16, +0.36, -0.60, and -0.60 respectively in this parameter set. Protonated aspartic and glutamic acids end in -C-COOH, and these atoms are given partial atomic charges of 0.00, +0.70, -0.55, -0.60, and +0.45 respectively.

To explicitly take into account that the protonated and deprotonated states actually each consist of several microstates, we start with Equation 3.11 and populate the protonated and deprotonated microstates x according to the Boltzmann distribution:

$$\exp\left[-\beta \Delta G_p(\mathrm{A}^-)\right] = \frac{[\mathrm{AH}]}{[\mathrm{A}^-]} \tag{3.20}$$

$$\Delta G_p(\mathrm{A}^-) = -k_\mathrm{B} T \log \frac{\sum_{\text{protonated x}} \exp(-\beta \Delta G_\text{micro}(\mathrm{x}))}{\sum_{\text{deprotonated x}} \exp(-\beta \Delta G_\text{micro}(\mathrm{x}))} \tag{3.21}$$

Using Equation 3.15,

$$\ln(10) \cdot \left[\mathrm{pH} - \mathrm{p}K_a(\mathrm{A}^-)\right] = -\log \frac{\sum_{\text{protonated x}} \exp(-\beta G_\text{micro}(\mathrm{x}))}{\sum_{\text{deprotonated x}} \exp(-\beta G_\text{micro}(\mathrm{x}))} \tag{3.22}$$

Since we know $\mathrm{p}K_a(\mathrm{A}^-)$ from experiment, Equation 3.22 is one equation we will need to find all of the relative free energies of the microstates $G_\text{micro}(\mathrm{x})$. The other equations we need will be free energy differences $G_\text{micro}(\mathrm{x}_1) - G_\text{micro}(\mathrm{x}_2)$ among the protonated states, and among the deprotonated states, which can be obtained from certain experiments which we will now discuss. Equation 3.22 serves to connect the $G_\text{micro}(\mathrm{x})$ for protonated states to the $G_\text{micro}(\mathrm{x})$ for deprotonated states.

3.2.3 Barnase and its Four Residues Asp54, Glu73, Asp75, His102

Barnase is an extracellular ribonuclease from *Bacillus amyloliquefaciens*, and barstar is its intracellular inhibitor. Barnase has evolved to be catalytically active (dissociation constant $K_d \approx 10^{-14}$ M), and barnase and barstar together have evolved to bind (as shown in Figure 3-2) tightly and rapidly (association rate constant 3.7×10^8 s^{-1} M^{-1}) because barnase activity inside the cell would be lethal [33, 34].

We will model the titration of 4 barnase residues (Asp54, Glu73, Asp75, His102). We chose these residues because they are near each other and near the binding interface with barstar, and we want to investigate the effects of multiple residues' titration on each other and on binding. The 3 carboxylic acid residues form a negatively charged layer behind the positively charged layer at the binding interface made up of Arg87, Arg83,

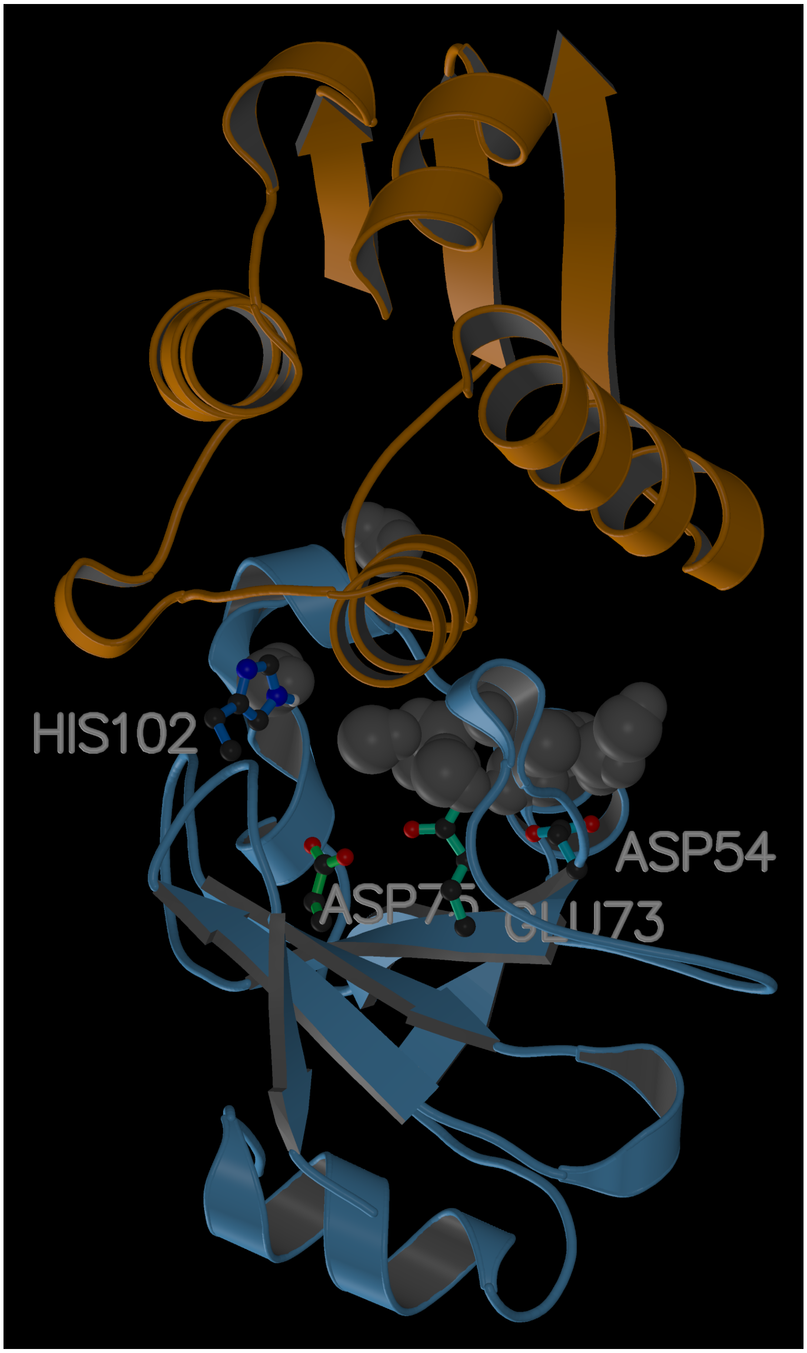

Figure 3-2: Barnase and barstar are shown as blue and orange ribbon diagrams. Explicit interfacial waters are shown as gray van der Waals spheres. The 4 residues of barnase that we allow to titrate are Asp54, Glu73, Asp75, His102. (This figure was made with the molecular graphics program MolScript [35].)

and Lys27. Barnase is a ribonuclease, and this positive charge on its binding region is important for its binding to negatively charged nucleic acids.

3.2.4 Histidine Model Compound Microstate Free Energy Differences

Tanokura [37] used hydrogen NMR to measure the microscopic pK_a values for each of the 2 protonation sites of imidazole and the histidine model compound N-acetylhistidine methylamide. We will denote the microstates of histidine by HisH$_{\delta 1}$, HisH$_{\epsilon 2}$, HisH$_{\delta 1}$H$_{\epsilon 2}$, HisH$_{\delta 1}$flip, HisH$_{\epsilon 2}$flip, and HisH$_{\delta 1}$H$_{\epsilon 2}$flip, where the protonation state is denoted by the PARAM19 atom names for the hydrogens bonded to the N$_{\delta 1}$ and N$_{\epsilon 2}$ nitrogens, and "flip" means that the ring is flipped. Of course, reducing the conformational flexibility of a histidine side chain to the 2 states, flipped and unflipped, is a gross simplification. The idea here is to trust the heavy atom positions from the crystal structure but solve for the uncertainty in titration state and ring orientation. This will only be valid at the pH of the crystal, since heavy atom positions may move as pH changes.

For the histidine model compound, Tanokura found a pK_a of 6.53 for the protonation reaction HisH$_{\epsilon 2}$ + H$^+$ $\rightleftharpoons$ HisH$_{\delta 1}$H$_{\epsilon 2}$, and a pK_a of 6.92 for the protonation reaction HisH$_{\delta 1}$ $\rightleftharpoons$ HisH$_{\delta 1}$H$_{\epsilon 2}$. These values imply a macroscopic pK_a value of 6.38 . We use the standard macroscopic pK_a value of 6.3 as stated earlier, but we will use the difference of Tanokura's microscopic pK_a values to set the free energy difference of HisH$_{\delta 1}$ and HisH$_{\epsilon 2}$ in the histidine model compound:

$$\begin{aligned} G_{\text{micro}}(\text{model HisH}_{\delta 1}) - G_{\text{micro}}(\text{model HisH}_{\epsilon 2}) &= (6.92 - 6.53)\text{pH units} \\ &= 0.53 \text{ kcal mol}^{-1} \qquad (3.23) \end{aligned}$$

$$G_{\text{micro}}(\text{model HisH}_{\delta 1}\text{flip}) - G_{\text{micro}}(\text{model HisH}_{\epsilon 2}\text{flip}) = 0.53 \text{ kcal mol}^{-1} \qquad (3.24)$$

To connect microstates of the histidine model compound with the ring flipped and unflipped, we use only electrostatic free energy, which we calculate as we would for any

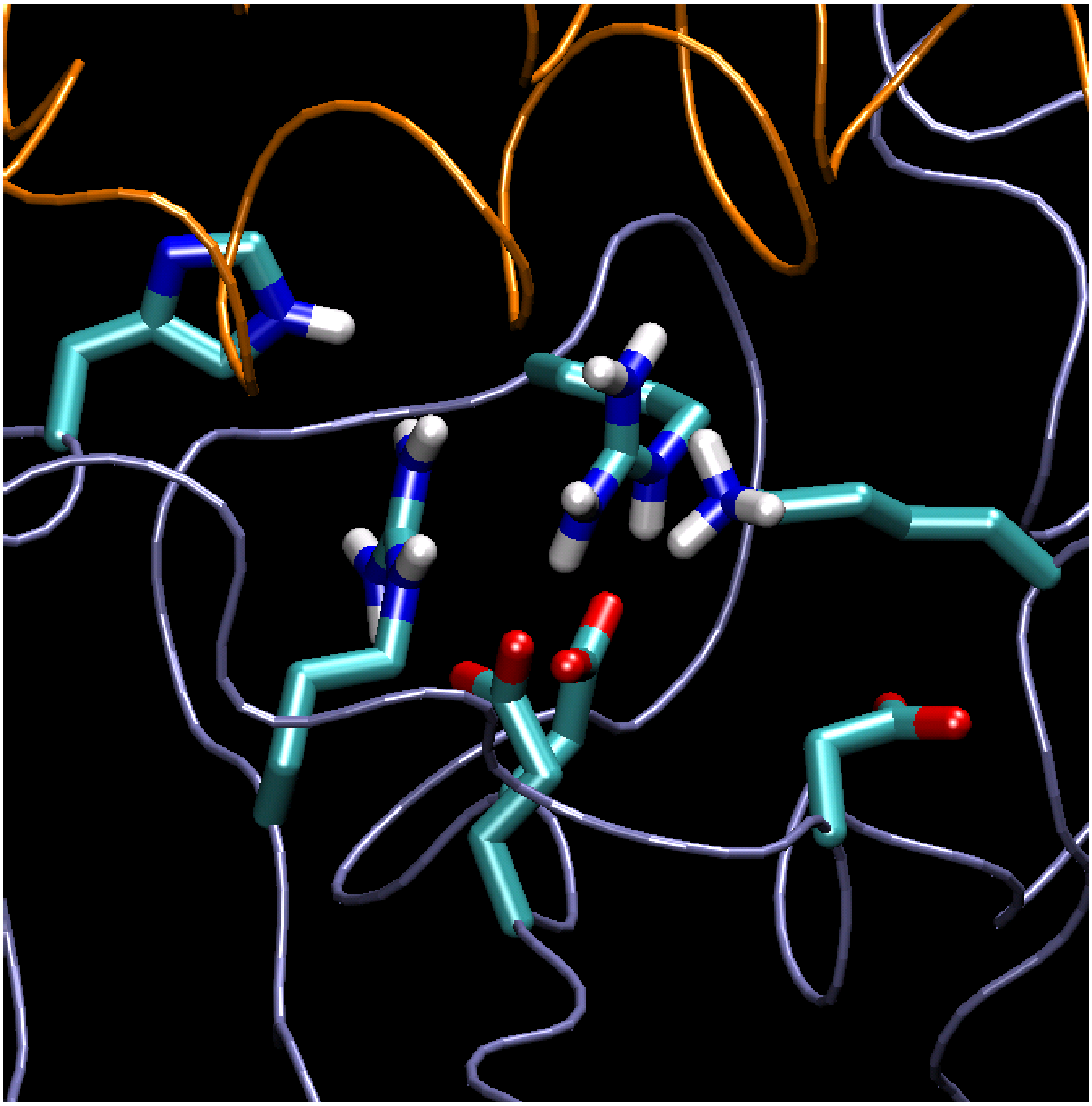

Figure 3-3: The 4 residues we allow to titrate are His102 at the upper left, and Glu73, Asp75, and Asp54 from left to right at the bottom of the picture. These 3 carboxylic acids form a negatively charged layer behind the positively charged layer at the binding interface, made up of Arg87, Arg83, and Lys27 from left to right in the center of the picture. Smooth C$_\alpha$ traces are shown for barnase and barstar in purple and orange. (This figure was made with the molecular graphics program VMD [36].)

other conformational change, as described in Section 2.2.2 of Chapter 2, using FDPB to calculate the solvation free energy of each of the states, and Coulomb's Law to calculate their electrostatic energy difference when desolvated:

$$\begin{aligned}
G_{\mathrm{micro}}&(\text{model His102 H}_{\delta 1}) - G_{\mathrm{micro}}(\text{model His102 H}_{\delta 1}\text{flip}) \\
&= [\Delta G_{\mathrm{solv}}(\text{model His102 H}_{\delta 1}) - \Delta G_{\mathrm{solv}}(\text{model His102 H}_{\delta 1}\text{flip})] \\
&\quad + [G_{\mathrm{desolvated}}(\text{model His102 H}_{\delta 1}) - G_{\mathrm{desolvated}}(\text{model His102 H}_{\delta 1}\text{flip})] \\
&= -\,0.157 \text{ kcal mol}^{-1}
\end{aligned} \tag{3.25}$$

Similarly,

$$G_{\mathrm{micro}}(\text{model His102 H}_{\epsilon 2}) - G_{\mathrm{micro}}(\text{model His102 H}_{\epsilon 2}\text{flip}) = -0.278 \text{ kcal mol}^{-1} \tag{3.26}$$

The 5 Equations 3.22, 3.23, 3.24, 3.25, and 3.26 determine free energies for all 6 microstates of the histidine model compound. Arbitrarily setting the free energy of the $H_{\epsilon 2}$ state to zero will not affect any final results.

$$\begin{aligned}
G_{\mathrm{micro}}(\text{model His102 H}_{\delta 1}) &= +0.53 \text{ kcal mol}^{-1} \\
G_{\mathrm{micro}}(\text{model His102 H}_{\epsilon 2}) &= 0 \text{ kcal mol}^{-1} \\
G_{\mathrm{micro}}(\text{model His102 H}_{\delta 1}\text{H}_{\epsilon 2}) &= +0.70 \text{ kcal mol}^{-1} + \log(10) k_{\mathrm{B}} T (\mathrm{pH} - 7) \\
G_{\mathrm{micro}}(\text{model His102 H}_{\delta 1}\text{flip}) &= +0.37 \text{ kcal mol}^{-1} \\
G_{\mathrm{micro}}(\text{model His102 H}_{\epsilon 2}\text{flip}) &= -0.16 \text{ kcal mol}^{-1} \\
G_{\mathrm{micro}}(\text{model His102 H}_{\delta 1}\text{H}_{\epsilon 2}\text{flip}) &= +0.54 \text{ kcal mol}^{-1} + \log(10) k_{\mathrm{B}} T (\mathrm{pH} - 7)
\end{aligned} \tag{3.27}$$

where, at room temperature,

$$\log(10) k_{\mathrm{B}} T = 1.36 \text{ kcal mol}^{-1} = -1\text{pH unit} \tag{3.28}$$

3.2.5 Aspartic and Glutamic Acid Model Compound Microstate Free Energy Differences

Wiberg and Laidig [38] used ab initio quantum mechanical calculations to obtain a gas phase energy difference of the anti and syn conformations of protonated acetic acid of 5.85 kcal mol^{-1}. We will denote the microstates of aspartic acid by "deprotonated", anti1, syn1, anti2, and syn2, where the number shows whether the hydrogen is bonded to the $O_{\delta 1}$ or $O_{\delta 2}$ atom. Similarly for glutamic acid, except that the hydrogen bonds to the $O_{\epsilon 1}$ or $O_{\epsilon 2}$ atom. We calculated FDPB electrostatic solvation free energies for each protonated microstate of each aspartic or glutamic acid model compound, then combined them to get free energy differences of the model compound microstates in solution:

$$\begin{aligned} G_{\text{micro}}(\text{model anti}) - \mathrm{G}_{\text{micro}}(\text{model syn}) &= [G_{\text{gas}}(\text{model anti}) - \mathrm{G}_{\text{gas}}(\text{model syn})] \\ &+ [\Delta G_{\text{solv}}(\text{model anti}) - \Delta \mathrm{G}_{\text{solv}}(\text{model syn})] \end{aligned} \tag{3.29}$$

$$\begin{aligned} G_{\text{micro}}(\text{model Asp54 anti1}) - G_{\text{micro}}(\text{model Asp54 syn1}) &= 4.31 \text{ kcal mol}^{-1} \\ G_{\text{micro}}(\text{model Asp54 anti2}) - G_{\text{micro}}(\text{model Asp54 syn2}) &= 2.48 \text{ kcal mol}^{-1} \\ G_{\text{micro}}(\text{model Glu73 anti1}) - G_{\text{micro}}(\text{model Glu73 syn1}) &= 3.97 \text{ kcal mol}^{-1} \\ G_{\text{micro}}(\text{model Glu73 anti2}) - G_{\text{micro}}(\text{model Glu73 syn2}) &= 3.38 \text{ kcal mol}^{-1} \\ G_{\text{micro}}(\text{model Asp75 anti1}) - G_{\text{micro}}(\text{model Asp75 syn1}) &= 6.76 \text{ kcal mol}^{-1} \\ G_{\text{micro}}(\text{model Asp75 anti2}) - G_{\text{micro}}(\text{model Asp75 syn2}) &= 3.62 \text{ kcal mol}^{-1} \end{aligned} \tag{3.30}$$

Unlike for histidine, the 2 protonation sites for aspartic and glutamic acids are not chemically distinguishable, so we calculate the free energy difference of the syn1 and syn2 states as we would for any other conformational change, as described in Section 2.2.2 of Chapter 2, using FDPB to calculate the solvation free energy of each of the states, and

Coulomb's Law to calculate their electrostatic energy difference when desolvated:

$$\begin{aligned} & G_{\text{micro}}(\text{model syn1}) - G_{\text{micro}}(\text{model syn2}) \\ & \qquad = [\Delta G_{\text{solv}}(\text{model syn1}) - \Delta G_{\text{solv}}(\text{model syn2})] \\ & \qquad\quad + [G_{\text{desolvated}}(\text{model syn1}) - G_{\text{desolvated}}(\text{model syn2})] \end{aligned} \tag{3.31}$$

$$\begin{aligned} G_{\text{micro}}(\text{model Asp54 syn1}) - G_{\text{micro}}(\text{model Asp54 syn2}) &= -0.321 \text{ kcal mol}^{-1} \\ G_{\text{micro}}(\text{model Glu73 syn1}) - G_{\text{micro}}(\text{model Glu73 syn2}) &= +0.055 \text{ kcal mol}^{-1} \\ G_{\text{micro}}(\text{model Asp75 syn1}) - G_{\text{micro}}(\text{model Asp75 syn2}) &= -0.413 \text{ kcal mol}^{-1} \end{aligned} \tag{3.32}$$

For each aspartic or glutamic acid residue, Equations 3.22, 3.30, and 3.32 give 4 equations which determine free energies for all 5 microstates of each model compound. Arbitrarily setting the free energy of the deprotonated states to zero at pH = 7 will not affect any final results.

$$\begin{aligned} G_{\text{micro}}(\text{model Asp54 deprotonated}) &= 0 - \log(10) k_{\text{B}} T (\text{pH} - 7) \\ G_{\text{micro}}(\text{model Asp54 syn1}) &= +4.36 \text{ kcal mol}^{-1} \\ G_{\text{micro}}(\text{model Asp54 syn2}) &= +4.68 \text{ kcal mol}^{-1} \\ G_{\text{micro}}(\text{model Asp54 anti1}) &= +8.67 \text{ kcal mol}^{-1} \\ G_{\text{micro}}(\text{model Asp54 anti2}) &= +7.15 \text{ kcal mol}^{-1} \end{aligned} \tag{3.33}$$

$$\begin{aligned} G_{\text{micro}}(\text{model Glu73 deprotonated}) &= 0 - \log(10) k_{\text{B}} T (\text{pH} - 7) \\ G_{\text{micro}}(\text{model Glu73 syn1}) &= +3.98 \text{ kcal mol}^{-1} \\ G_{\text{micro}}(\text{model Glu73 syn2}) &= +3.92 \text{ kcal mol}^{-1} \\ G_{\text{micro}}(\text{model Glu73 anti1}) &= +7.95 \text{ kcal mol}^{-1} \\ G_{\text{micro}}(\text{model Glu73 anti2}) &= +7.30 \text{ kcal mol}^{-1} \end{aligned} \tag{3.34}$$

$$G_{\text{micro}}(\text{model Asp75 deprotonated}) = 0 - \log(10) k_{\text{B}} T (\text{pH} - 7)$$

$$\begin{aligned} G_{\text{micro}}(\text{model Asp75 syn1}) &= +4.32 \text{ kcal mol}^{-1} \\ G_{\text{micro}}(\text{model Asp75 syn2}) &= +4.73 \text{ kcal mol}^{-1} \\ G_{\text{micro}}(\text{model Asp75 anti1}) &= +11.08 \text{ kcal mol}^{-1} \\ G_{\text{micro}}(\text{model Asp75 anti2}) &= +8.35 \text{ kcal mol}^{-1} \end{aligned} \tag{3.35}$$

3.2.6 Energy Function

Our energy function includes electrostatic and covalent terms. The electrostatic term is calculated with FDPB. The covalent energy term includes bond, angle, dihedral, and improper dihedral terms according to the PARAM19 parameter set.

van der Waals Term Used as a Screen

All titration microstates were allowed for the model compounds, but for the residues in the protein, some titration microstates were disallowed due to bad van der Waals clashes. For example, the Asp54 syn2 hydrogen position clashes with the hydrogen of the barnase Lys27 backbone amide. The states Asp75 syn1, Asp75 syn2, His102 $H_{\delta 1}$, and His102 $H_{\delta 1}H_{\epsilon 2}$ were also disallowed due to van der Waals clashes. If we had some way to allow conformational flexibility for heavy atoms, rather than just hydrogen atoms, then such states would perhaps be allowed in concert with auxiliary movements. In one study that attempted to do this, Havranek and Harbury [25] combined discrete side chain freedom with multiple protonation states.

We do not include a van der Waals term in our free energy function, however, because we found that van der Waals energy differences between states are not meaningful. If we did include the van der Waals energy, then by consulting the thermodynamic cycle in Figure 3-4, we see that the van der Waals energy difference of interest is

$$[G_{\text{vdW}}(\{i_r\},\ b) - G_{\text{vdW}}(\{i_1\},\ b)] - \sum_i [G_{\text{vdW}}(\text{model } i_r) - G_{\text{vdW}}(\text{model } i_1)] \tag{3.36}$$

Consider the protonation of an Asp or Glu: one hydrogen atom is added, and the

CHARMM atom type changes for the oxygen to which it is bonded. The van der Waals energy difference in Equation 3.36 will contain several contributions which have no physical basis:

1. The hydrogen atom's van der Waals interactions with all other atoms in the protein that are *not* its immediate neighbors are all favorable, and their total magnitude is not small. This unfairly favors protonation, even at a solvent-exposed site where the pK_a should be the same as the model compound.

2. The hydrogen atom will unrealistically favor a position where it can contact more protein atoms for favorable van der Waals interactions, versus a position facing solvent.

3. The change of the oxygen's CHARMM atom type changes its van der Waals radius in the PARAM19 parameter set, which can make the van der Waals energy function penalize it for a van der Waals clash with a neighboring atom even though they were both in the same positions before the addition of the hydrogen. It is not clear that the paramer set's van der Waals radii are realistic; if they are, then these atoms would have to move apart upon protonation to relieve the van der Waals clash.

The underlying reasons that the van der Waals term does not give meaningful energy differences are (1) we do not include the van der Waals interaction of a hydrogen ion free in solution, (2) we do not include a hydrophobic term to account for the van der Waals interactions of explicit protein atoms with implicit water molecules, and (3) we do not allow heavy atoms to move in response to protonation state changes. While the first 2 problems could be addressed, the last is very difficult; neglecting the van der Waals contribution appears to be the preferred option.

Free Energy Differences of Protein Protonation States

The thermodynamic cycle used to calculate free energies for each titration state of the protein is shown in Figure 3-4. A state of the protein is specified by $(\{i_r\},\ b)$, meaning that the microstate i_r is specified for each titrating residue number i, and b is 0 or 1 for the unbound or bound state. The purpose of the cycle is to calculate the free energy of protein state $(\{i_r\},\ b)$ relative to the reference protein state $(\{i_1\},\ b)$, in which every titrating residue is in its microstate #1.

Since we have a model compound for each titrating residue, in the same conformation as the residue, the grid energies cancel out of the terms for the 2 vertical arrows in the top left half of the cycle:

$$\begin{aligned}
&G_{\mathrm{ES}}(\{i_r\},\ b,\ \text{no aux.}) - G_{\mathrm{ES}}(\{i_1\},\ b) - \textstyle\sum_i \left[G_{\mathrm{ES}}(\text{model } i_r) - G_{\mathrm{ES}}(\text{model } i_1)\right] \\
&\quad= G_{\mathrm{ES,\ FDPB}}(\{i_r\},\ b,\ \text{no aux.}) - G_{\mathrm{ES,\ FDPB}}(\{i_1\},\ b) \\
&\qquad - \textstyle\sum_i \left[G_{\mathrm{ES,\ FDPB}}(\text{model } i_r) - G_{\mathrm{ES,\ FDPB}}(\text{model } i_1)\right]
\end{aligned} \tag{3.37}$$

The grid energies also cancel out of each of the two vertical arrows in the bottom right half of the cycle, because they are solvation free energies.

The result of the thermodynamic cycle, with the terms written in the same order as their corresponding arrows in Figure 3-4, is

$$\begin{aligned}
G_{\mathrm{ES}}(\{i_r\},\ b) - G_{\mathrm{ES}}(\{i_1\},\ b) = &-G_{\mathrm{ES,\ FDPB}}(\{i_1\},\ b) \\
&+\textstyle\sum_i \big[\ +G_{\mathrm{ES,\ FDPB}}(\text{model } i_1) \\
&\qquad -G_{\mathrm{micro}}(\text{model } i_1) \\
&\qquad +G_{\mathrm{micro}}(\text{model } i_r) \\
&\qquad -G_{\mathrm{ES,\ FDPB}}(\text{model } i_r)\ \big] \\
&+G^{\mathrm{desolvated}}_{\mathrm{ES,\ FDPB}}(\{i_r\},\ b,\ \text{no aux.}) \\
&-G^{\mathrm{desolvated}}_{\mathrm{ES,\ Coul}}(\{i_r\},\ b,\ \text{no aux.}) \\
&+G^{\mathrm{desolvated}}_{\mathrm{ES,\ Coul}}(\{i_r\},\ b) \\
&+\Delta G^{\mathrm{solv}}_{\mathrm{ES,\ FDPB}}(\{i_r\},\ b)
\end{aligned} \tag{3.38}$$

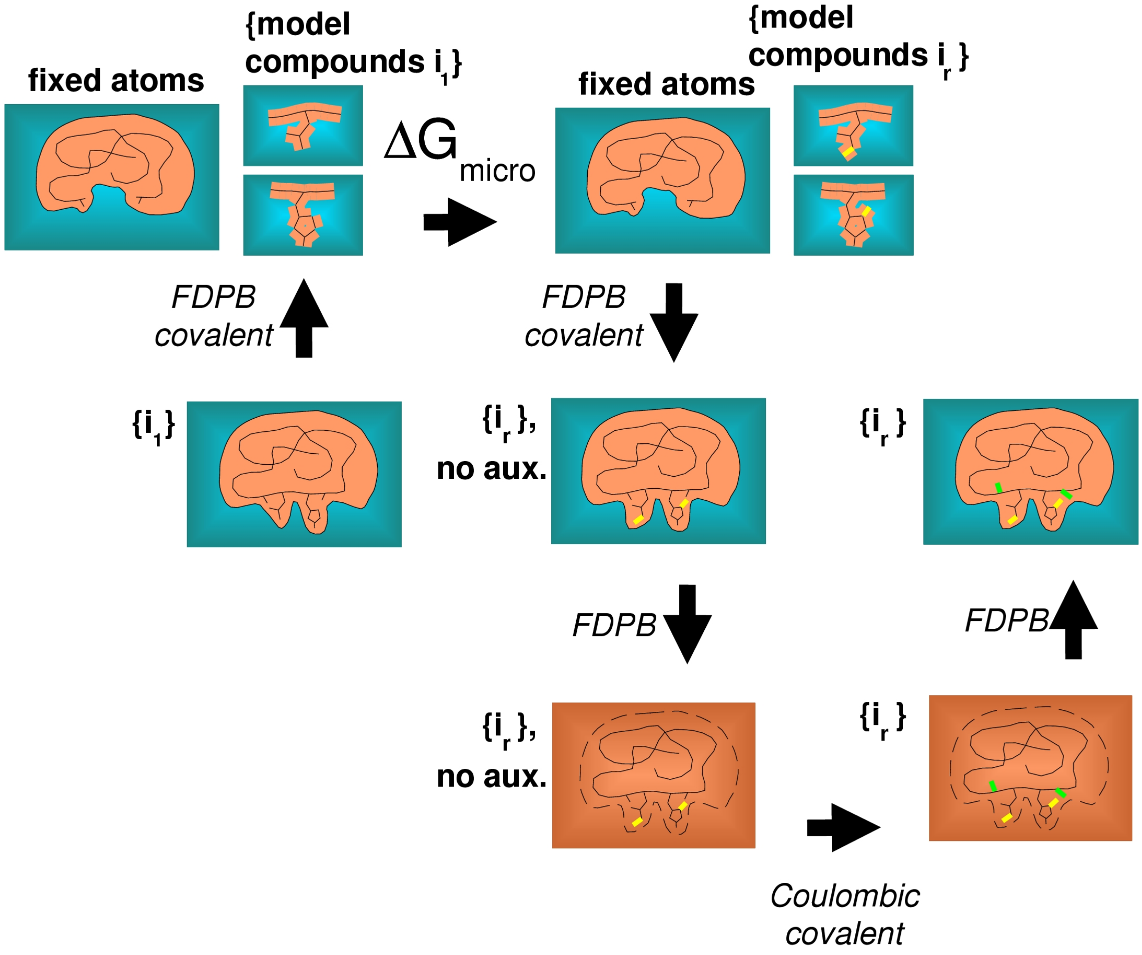

Figure 3-4: Thermodynamic cycle connecting protein state ($\{i_r\}$, b) to reference protein state ($\{i_1\}$, b). The top left half of the cycle changes from protonation state $\{i_1\}$ to $\{i_r\}$ (these changes are shown in yellow); this is calculated using the model compounds, one per titrating residue. The energy of the fixed atoms cancels out. In the middle of the cycle is an intermediate state in which the titrating residues are in their proper microstates ($\{i_r\}$, b), but the other atoms have made no “auxiliary movements” in response to the titration state. The bottom right half of the cycle allows other atoms (hydrogens only, in our model) to make auxiliary movements (these changes are shown in green); this is calculated using desolvated states.

where all terms are determined either by FDPB, Coulomb's Law, or by the values of G_{micro} given in Equations 3.27, 3.33, 3.34, and 3.35. The values of G_{micro} for charged microstates depend linearly on the pH, and this is the source of the pH-dependency of the binding free energy.

The free energies of all bound states are linked to those of all unbound states via the rigid binding free energy of the reference titration state, which can be calculated by FDPB:

$$\begin{aligned}\Delta G_{\mathrm{ES}}^{\mathrm{bind}}(\{i_r\}) &\equiv G_{\mathrm{ES}}(\{i_r\},\ \text{bound}) - G_{\mathrm{ES}}(\{i_r\},\ \text{unbound}) \\ &= G_{\mathrm{ES,\ FDPB}}(\{i_r\},\ \text{bound}) \\ &\quad -G_{\mathrm{ES,\ FDPB}}(\{i_r\},\ \text{unbound barnase}) \\ &\quad -G_{\mathrm{ES,\ FDPB}}(\text{unbound barstar}) \qquad (3.39)\end{aligned}$$

where the grid energy cancels out if, for each of the unbound binding partners, their atom positions and finite-difference grids are made the same as in the bound state by using the "dummy atom" positions determined for the bound state.

3.2.7 Simplifications to Speed pK_a Calculation

We evaluate the effect on pK_a prediction of several simplifying approximations, some of which have been used in previous pK_a prediction studies. First, **simpler charge patterns**: simplifications can be made in the assignment of partial charges for each titration state of each residue type. Methods that allow only two microstates, protonated and deprotonated, lose a good deal of structural accuracy in exchange for reducing the number of states.

Second, there are simplifications that can reduce the number of energy calculations required for N titrating residues with m titration microstates each from m^N to $m \times n$. This is possible if the free energy of any titration state can be decomposed into single-residue and residue-pair terms. In the FDPB method, this can be done by making two assumptions:

1. a **single dielectric boundary** for the protein, which requires either
   (a) fixing all atoms and not adding explicit new hydrogen atoms to the protonated state, or
   (b) using the union of the van der Waals regions of all titration states to define the dielectric boundary for all states.

   The second assumption is

2. no correlation between conformational changes caused by the titration of different residues (i.e., each atom is either fixed, or it moves in response to the titration of only one residue). This can be further simplified by
   (a) allowing only polar hydrogen atoms to move, or even
   (b) allowing no movement at all outside of the titrating residues.

We evaluate the assumption (1.b) in the present work. We do not allow non-hydrogen atoms to move in the present study; so we compare assumption (2.b), **no auxiliary movement**, to our most complete model, in which only hydrogens may move, but they are allowed to relax independently for every titration state of the protein.

On the other hand, assumption (2) can be made less sweeping; for example, the hybrid statistical mechanical/Tanford-Roxby [39] method assumes that residues that are not too near each other interact only in a mean-field sense: the average charge of one depends only on the average charge of the other, thus avoiding consideration of each possible pair of their titration microstates. We do not evaluate this method here; it is less drastic than the form of assumption (2) that we consider, and so it does not save as much computation time; but it has shown promise in other studies [40].

When assumptions (1) and (2) are met, and therefore only $m \times N$ FDPB calculations are required, there are still m^N free energy terms that come out of those FDPB calculations and are needed to do exact calculations. If this calculation is intractably large, very accurate approximations to the full free energy can be calculated using Monte Carlo techniques to partially sample the state space [41, 42].

Simplification: Simpler Charge Patterns

In addition to the set of titration microstates described above (which we will call the **param19 charge pattern**), we also evaluate the effects of using either of two simpler charge patterns, which have only 1 deprotonated and 1 protonated microstate for each titrating residue. These simpler charge patterns, which we will call the **point** and **smeared charge patterns**, have been used in previous pK_a prediction studies. The exact charges used for the 3 methods of charge arrangement are given in Table 3.1.

Table 3.1: Point charges for the 3 charge pattern methods. For the "point" and "smeared" patterns, there are only the 2 microstates given here. For the "PARAM19" pattern, the Asp and Glu protonated charges shown here are for the syn1 or anti1 microstates, the His deprotonated charges shown here are for the $H_{\delta 1}$ microstate, and the His protonated charges shown here are for the $H_{\delta 1}H_{\epsilon 2}$ microstate.

	deprotonated			protonated		
atom	param19	point	smeared	param19	point	smeared
Asp C_β or Glu C_γ	-0.16	-0.16	-0.16	0.00	-0.16	0.00
Asp C_γ or Glu C_δ	0.36	0.36	0.36	0.70	1.36	0.70
Asp $O_{\delta 1}$ or Glu $O_{\epsilon 1}$	-0.60	-0.60	-0.60	-0.60	-0.60	-0.35
Asp $H_{\delta 1}$ or Glu $H_{\epsilon 1}$	N/A	N/A	N/A	0.45	N/A	N/A
Asp $O_{\delta 2}$ or Glu $O_{\epsilon 2}$	-0.60	-0.60	-0.60	-0.55	-0.60	-0.35
His C_β	0.00	0.00	0.00	0.10	0.00	0.10
His C_γ	0.10	0.10	0.10	0.15	0.10	0.15
His $N_{\delta 1}$	-0.40	-0.40	-0.40	-0.30	-0.40	-0.30
His $H_{\delta 1}$	0.30	0.30	0.15	0.35	0.30	0.35
His $C_{\delta 2}$	0.10	0.10	0.10	0.20	0.10	0.20
His $C_{\epsilon 1}$	0.30	0.30	0.30	0.45	0.30	0.45
His $N_{\epsilon 2}$	-0.40	-0.40	-0.40	-0.30	0.60	-0.30
His $H_{\epsilon 2}$	N/A	N/A	0.15	0.35	N/A	0.35

param19 Charge Pattern

We will call our most careful way to arrange the partial atomic charges the **param19 charge pattern**, because we just use the PARAM19 partial atomic charges, including the explicit titrating hydrogen atom. We use all of the titration microstates we described above: 5 for aspartic and glutamic acid, and 6 for histidine. For the 4 barnase residues we are allowing to titrate, some of the microstates are disallowed due to van der Waals clashes. The number of titration states of the whole unbound protein barnase in our model is $4 \times 5 \times 3 \times 4 = 240$. The bound state also has 240 states.

Point Charge Pattern

The **point charge pattern**, employed by Bashford and Karplus [27], uses only 2 titration microstates per titratable residue, protonated and deprotonated. For aspartic acid, the protonated form has a charge of +1 added to the C_γ atom of the carbonyl group. For glutamic acid, the protonated form has a charge of +1 added to the C_δ atom of the carbonyl group. For histidine, the neutral, singly-protonated form has the explicit polar hydrogen $H_{\delta 1}$ bonded to $N_{\delta 1}$ as in the PARAM19 topology; and the positively-charged, doubly-protonated form has a charge of +1 added to the $N_{\epsilon 2}$ atom. For barnase His102, the unflipped deprotonated state is disallowed by a van der Waals clash, so we have the ring flipped for both the deprotonated and protonated microstates. Bashford and Karplus did not address the possibility of histidine rings being flipped the wrong way in the crystal structure.

Smeared Charge Pattern

The **smeared charge pattern**, which we model after that employed by Schaefer, Sommer, and Karplus [26], uses only 2 titration microstates per titratable residue, protonated and deprotonated. For aspartic acid, the protonated form has a total charge of +1 spread over the 4 atoms C_β, C_γ, $O_{\delta 1}$, and $O_{\delta 2}$. Similarly, for glutamic acid, the protonated form has a total charge of +1 spread over the 4 atoms C_γ, C_δ, $O_{\epsilon 1}$, and $O_{\epsilon 2}$. The neutral, singly-protonated form of histidine is represented by one microstate, with some charge at both the $H_{\delta 1}$ and $H_{\epsilon 2}$ positions. The positively-charged, doubly-protonated form of histidine is represented by one microstate with the same partial charges as the $H_{\delta1}H_{\epsilon2}$ microstate used in the PARAM19 charge pattern. As noted above for the **point charge pattern**, the unflipped deprotonated state of barnase His102 is disallowed by a van der Waals clash, so we have the ring flipped for both the deprotonated and protonated microstates. Two differences of our method from that of Schaefer, Sommer, and Karplus are: they did not consider flipping the histidine rings, and they used the PARAM22 all-atom parameter set, so they had partial charges on nonpolar hydrogen atoms as well.

Simplification: Single Dielectric Boundary

The 4 titration microstates of Asp54 all have different dielectric boundaries in both the bound and unbound states, because of the titrating hydrogen atom, and because of the auxiliary movements of Ser28HG and waters #60 and #361 when all hydrogens are allowed to relax (using the CHARMM HBUILD routine [43]) in response to the protonation state of Asp54. The 5 titration microstates of Glu73 all have different dielectric boundaries in the unbound state, because of auxiliary movements. In the bound state, they have only 2 different dielectric boundaries. The 3 titration microstates of Asp75 all share the same dielectric boundary in the bound state, and all share the same dielectric boundary in the unbound state. The 4 titration microstates of His102 all share the same dielectric boundary in the bound state, but have 3 different dielectric boundaries in the unbound state.

We investigated the simplification of using a **single dielectric boundary**. With this simplification, all titration states of the unbound protein use a dielectric boundary enclosing the union of all titration microstates of all titrating residues. One dielectric boundary for all bound states is constructed similarly. Also, when calculating free energies for the state with no auxiliary hydrogen movement (this state is at the middle of the thermodynamic cycle in Figure 3-4), similar constant dielectric boundaries are constructed for the unbound and bound forms of this state.

Simplification: No Auxiliary Movement

In our most complete model, each of the 480 titration states of the whole protein system (240 bound, 240 unbound) is allowed to relax the positions of all of its polar hydrogen atoms, using the HBUILD function of CHARMM.

We also evaluated the simplification of leaving all atoms except for the titrating hydrogens fixed at positions determined by HBUILD on the deprotonated state. This was done by Bashford and Karplus [27]. We call this the **no auxiliary movement** simplification.

3.3 Results and Discussion

3.3.1 Simulated Titration of Barnase Residues

Our full titration model determines free energies for every bound and unbound protein titration microstate. From these, the macroscopic electrostatic binding free energy can be calculated:

$$\Delta G^{\mathrm{bind}} = -k_{\mathrm{B}}T \log \frac{\sum_{\{i_r\}} \exp\left[-\beta G(\{i_r\},\ \mathrm{bound})\right]}{\sum_{\{i_r\}} \exp\left[-\beta G(\{i_r\},\ \mathrm{unbound})\right]} \tag{3.40}$$

Figure 3-5 shows the binding free energy ΔG^{bind} of barnase and barstar vs. pH for the full titration model of the 4 barnase residues (Asp54, Glu73, Asp75, His102), and also with the **single dielectric boundary** and **no auxiliary movement** simplifications. The binding free energy varies significantly with pH, and this conclusion holds even if our model does not accurately predict pK_a shifts. In our titration model, any pH value is possible, and we want to consider a wide enough range of pH values so that all of our residues can protonate, so we report results over the pH range -15 to 15 even though the protein would actually denature as the pH dropped below about 2.

Figure 3-6 shows the populations of the titration microstates of the 4 barnase residues

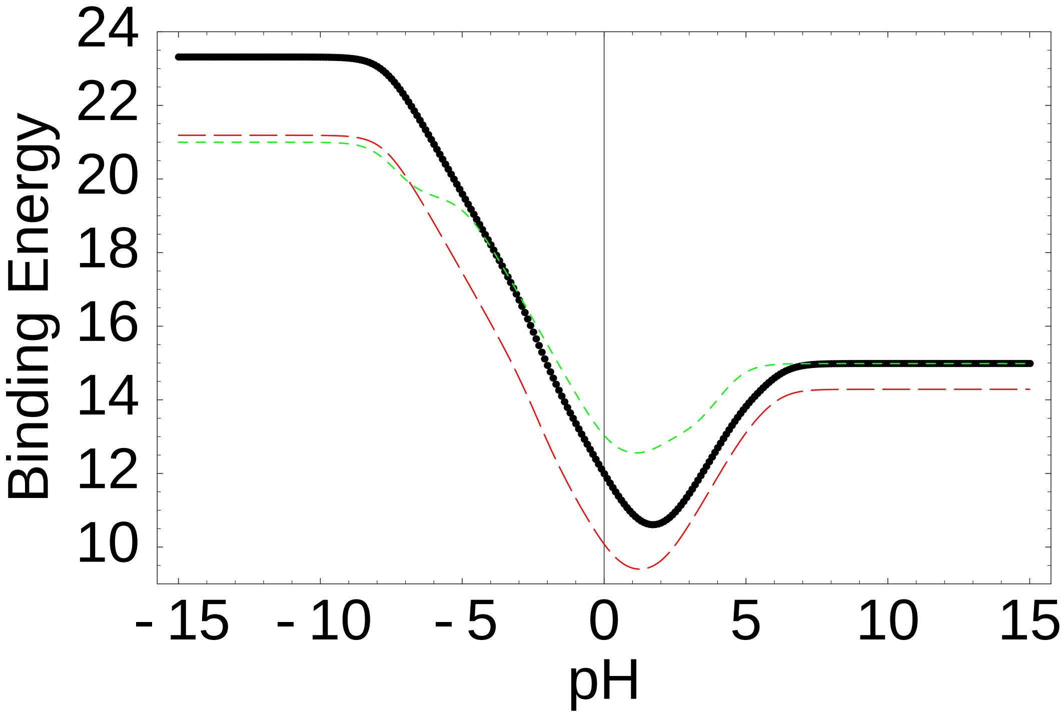

Figure 3-5: Electrostatic binding free energy of barnase and barstar vs. pH when all of our titration states for the 4 barnase residues (Asp54, Glu73, Asp75, His102) are available in both the unbound and bound states. The solid black line is for our most complete model. The long-dashed red line results when FDPB free energies of all states are obtained using a single dielectric boundary. The short-dashed green line results when other hydrogen atoms are kept fixed at their pH= 7 locations, rather than being allowed to relax to different positions for every titration state.

(Asp54, Glu73, Asp75, His102) vs. pH, when unbound and when bound to barstar.

Figure 3-7 shows the total charge of the 4 barnase residues (Asp54, Glu73, Asp75, His102) vs. pH, when unbound and when bound to barstar.

To reduce the populations of all of the protein microstates back to single pK_a values for each residue, we can define the effective pK_a^{eff} of each residue as the pH value at which half of the population has that residue deprotonated, and half has it protonated. Since the titration states of the residues are not independent, a plot of the degree of protonation of one residue vs. pH does not always have the classic sigmoid shape that a single titration site would have, but we are defining the pK_a^{eff} as the pH at which this plot crosses 0.5 . We can also define an "intrinsic" pK_a^{intr} for each residue, as the pK_a that it would have if the other titratable residues in the protein were fixed in their deprotonated reference states. The pK_a^{intr} and pK_a^{eff} values of each of our titrating residues of barnase, in the bound and unbound state, are given in Table 3.2, along with the available experimental values for pK_a^{eff} [31, 44, 33, 45]. The experimental values are not complete, and are contradictory in one case. Only an upper bound can be determined for some pK_a^{eff} values, because proteins denature when the pH drops past about 2. For the barnase residue His102, Buckle *et al.* [33] found a pK_a^{eff} shift upon binding, from 6.3 for unbound barnase to < 5 for barnase in complex with barstar.

The null model is the drastic simplification that all pK_a shifts are zero, so a residue's pK_a^{eff} and pK_a^{intr} are equal to the pK_a of the model compound. This means, of course, that the titration of each residue is not affected by the other residues' titration states, or by whether binding partners are bound to the protein. So ΔG^{bind} does not depend on pH in the null model.

The pK_a shifts

$$\Delta \mathrm{p}K_a \equiv \mathrm{p}K_a^{\text{eff}} - \mathrm{p}K_a(\text{model}) \tag{3.41}$$

depend on the free energies of all microstates, but they correspond approximately (within 0.2 pH units) to the free energy difference between the most populous deprotonated state

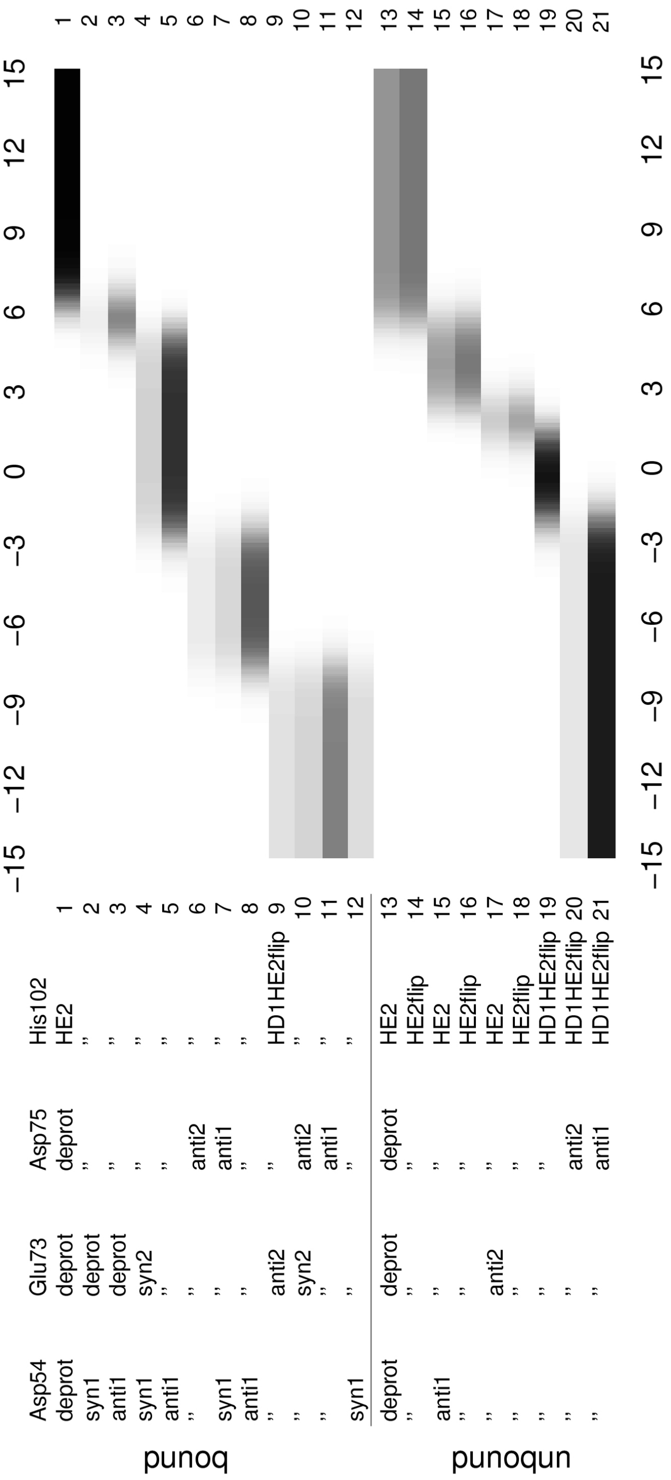

Figure 3-6: Relative populations of the titration microstates of the 4 barnase residues (Asp54, Glu73, Asp75, His102) vs. pH. The grayscale represents 0 (white) to 1 (black). The populations of the bound states at any pH value add up to 1, as do those of the unbound states. States which are not significantly populated at any pH are not shown.

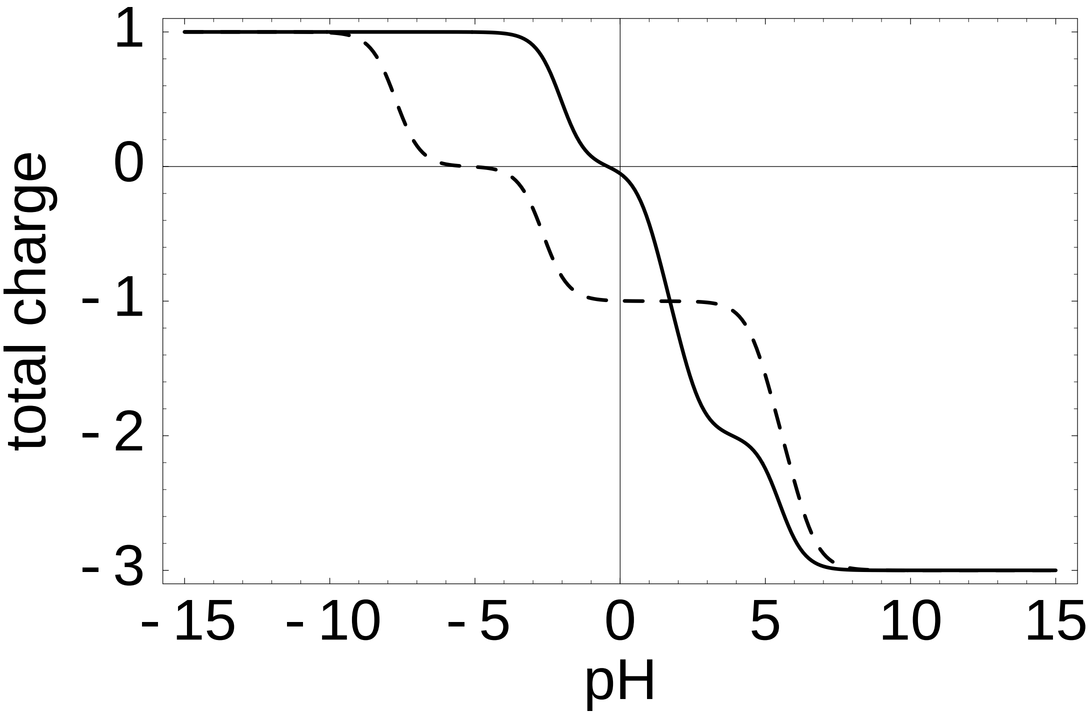

Figure 3-7: Total charge of the 4 barnase residues (Asp54, Glu73, Asp75, His102) vs. pH, when unbound (dotted line) and when bound to barstar (solid line).

Table 3.2: Model compound, intrinsic, effective, and experimental pK_a values for 4 residues of barnase (Asp54, Glu73, Asp75, His102). The pK_a^{intr} and pK_a^{eff} values come from our full titration model. The experimental numbers are values for pK_a^{eff}, so comparing the "predicted" and "experimental" results shows how poorly our full titration model does at predicting pK_a^{eff} values. Note that the null model of titration would give pK_a^{eff} = pK_a(model), so comparing the "model" and "experimental" results shows that the null model does poorly as well.

		model	intrinsic	predicted	experimental
		pK_a(model)	pK_a^{intr}	pK_a^{eff}	pK_a^{eff}
Asp54	unbound	4.0	5.46	5.48	$\leq$ 2.2 [31]
Asp54	bound	4.0	6.08	6.06	
Glu73	unbound	4.4	2.05	2.22	$\leq$ 2.1 [31], 4.8 [44]
Glu73	bound	4.4	5.26	5.09	
Asp75	unbound	4.0	1.81	-2.02	3.1 [31]
Asp75	bound	4.0	3.84	-2.66	
Glu102	unbound	6.3	0.86	1.18	6.3 [33], 6.3 [45], 5.4 [44]
Glu102	bound	6.3	-3.35	-7.74	< 5 [33]

and the most populous protonated state at pH = pK_a^{eff}. These free energy differences have the advantage of being easily broken up into contributions for every interaction, and for the desolvation of the residue due to burial in the protein dielectric cavity. These terms are shown in Table 3.3, with a further break-down of the interactions of the titrating group in Table 3.4.

Table 3.3: Terms of the pK_a shift. In order to break it up into terms, the pK_a shift is approximated by the free energy difference of the most populous deprotonated state $\{i_d\}$ and the most populous protonated state $\{i_p\}$ at pH = pK_a^{eff}, versus the model compounds, for each pK_a^{eff}. That is, these are the terms of

$$[G(\{i_p\}) - G(\{i_d\})] - \sum_i [G(\text{model } \{i_p\}) - G(\text{model } \{i_d\})]$$

Their total, when converted to pH units, equals the "pK_a shift" ΔpK_a to within 0.2 pH units. The "b" column has 0 or 1 for the unbound or bound state. The "ES group inter" column is for the total electrostatic interaction of the titrating residue. The "ES other inter" column is for electrostatic free energy terms not involving the titrating residue; these are from auxiliary movements — hydrogen atoms that have different HBUILD positions in the $\{i_d\}$ and $\{i_p\}$ states. The "ES desolv" column is for the electrostatic free energy of moving the titrating residue from the model compound's dielectric boundary into the protein's much larger dielectric boundary. The "covalent" column is for the covalent free energy term. The "total" column is the total; it corresponds to the "pK_a shift" ΔpK_a to within 0.2 pH units. All terms in this table are in kcal mol^{-1}.

		ES	ES			
		group	other	ES		
	b	inter	inter	desolv	covalent	total
Asp54	0	3.5	-1.7	-3.8	-0.1	-2.1
Asp54	1	1.8	0.0	-4.0	-0.6	-2.8
Glu73	0	13.0	-4.1	-6.0	0.1	3.0
Glu73	1	12.2	-3.7	-9.8	0.5	-0.8
Asp75	0	13.3	2.5	-7.3	-0.3	8.2
Asp75	1	17.5	1.2	-8.1	-1.4	9.2
His102	0	0.9	1.1	4.6	0.0	6.6
His102	1	8.2	-1.5	10.6	1.9	19.2

The electrostatic desolvation term always favors the neutral form of the titrating residue, because the charged form pays a desolvation penalty when buried in the protein.

Table 3.4: Subterms of the pK_a shift that add up to the "ES group inter" term in Table 3.3. In order to break it up into terms, the pK_a shift is approximated by the free energy difference of the most populous deprotonated state $\{i_d\}$ and the most populous protonated state $\{i_p\}$ at pH $=$ pK_a^{eff}, versus the model compounds, for each pK_a^{eff}. That is, these are the terms of

$$[G(\{i_p\}) - G(\{i_d\})] - \sum_i [G(\text{model } \{i_p\}) - G(\text{model } \{i_d\})]$$

The "b" column has 0 or 1 for the unbound or bound state. The "barnase 54,73,75 charge" column has 0 or - for the charge, 0 or -1, of barnase residues Asp54, Glu73, and Asp75. The next 5 columns are for the interactions between the titrating residue and various groups. The "ES group inter" column is for the total electrostatic interaction of the titrating residue. All terms in this table are in kcal mol^{-1}.

		barnase		barnase	barnase			ES
		54,73,75		27,83,87	other	water	barstar	group
	b	charge	inter	inter	inter	inter	inter	inter
Asp54	0	(−−)	-4.0	5.2	2.8	-0.5	.	3.5
Asp54	1	(−−)	-5.3	8.9	2.8	-0.3	-3.6	1.8
Glu73	0	(0 −)	-5.3	9.6	2.0	6.7	.	13.0
Glu73	1	(0 −)	-9.2	23.7	3.7	7.0	-13.0	12.2
Asp75	0	(00)	-0.9	15.5	-2.4	0.0	.	13.3
Asp75	1	(00)	-1.0	28.4	-1.9	0.0	-8.0	17.5
His102	0	(00−)	-1.2	4.2	-2.1	0.0	.	0.9
His102	1	(000)	-0.1	14.2	-1.6	0.1	-4.4	8.2

Three positively-charged barnase residues, Lys27, Arg83, and Arg87 are near the three carboxylic acid residues. Asp54, Glu73, and Asp75 are progressively closer, in that order, to the three positively-charged residues, which naturally favor the negatively-charged, deprotonated forms of the carboxylic acids. Barstar has a net charge of -5, due mainly to 4 carboxylic residues which line the docking region in order to interact with Lys27, Arg83, and Arg87. Due to this net charge, barstar favors the protonated forms of all the titrating residues.

At high pH, where Asp54, Glu73, and Asp75 are all deprotonated and negatively charged, the unfavorable interactions between their negative charges mean that they all favor each other's protonation. Now consider gradually dropping the pH. One of them will have the highest $\mathrm{p}K_a$, and that one will protonate first. Now that it is neutral, it no longer favors the protonation of the other 2 residues so strongly, so their $\mathrm{p}K_a$ values are now lower. Again, one of these 2 will have the higher $\mathrm{p}K_a$, and that one will protonate next. Now that it is neutral, it no longer favors the protonation of the last one of the three, so the third and final $\mathrm{p}K_a$ value drops even lower. This illustrates that $\mathrm{p}K_a^{\mathrm{eff}}$ prediction for multiple interacting titratable groups is very sensitive to the accuracy of the titration states' relative free energies. When the 3 carboxylic acid residues are all deprotonated, their $\mathrm{p}K_a$ values could be quite close to each other, but only the one that happens to be highest will protonate first as the pH is lowered, and this will shift the other residues' $\mathrm{p}K_a^{\mathrm{eff}}$ values down substantially.

In our treatment, Asp54 protonates before the other 2 carboxylic acids in unbound barnase as the pH drops, at $\mathrm{p}K_a^{\mathrm{eff}}(\text{Asp54, unbound}) = 5.48$. Experimentally, Oliveberg *et al.* [31] found that Asp75 protonates first, at $\mathrm{p}K_a^{\mathrm{expt}}(\text{Asp75, unbound}) = 3.1$, and that $\mathrm{p}K_a^{\mathrm{eff}}(\text{Asp54, unbound}) \leq 2.2$. So it would seem that our treatment incorrectly favors the protonated, neutral form of Asp54 over the deprotonated, negatively-charged form. The exact reason is not clear. One might expect that any error is due to Nature having a lower free energy conformation for some state. But, for this to explain our error on $\mathrm{p}K_a^{\mathrm{eff}}(\text{Asp54, unbound})$, Nature would have to have a better conformation for the deprotonated state of Asp54. This does not make sense, because the crystal structure was obtained with all 3 of Asp54, Glu73, and Asp75 deprotonated. The weakest part of our method seems to be that we did not allow heavy atoms to move in response to the protonation state, but the exact causes of our errors are not clear. It should be noted that these simplifications and consequent errors are endemic to nearly all studies of pK_a by calculation.

3.3.2 Simplifications to Speed pK_a Calculation

These simplifications can only be useful if they do not significantly affect predicted pK_a^{eff} values. It does not matter that our predicted pK_a^{eff} do not match experimental values; if a simplification is a valid approximation, it will not significantly affect the relative free energies of any states. So we judge the simplifications by comparing their resulting pK_a^{eff} and ΔG^{bind} predictions to those of our full titration model.

Table 3.5 summarizes the effects of the various simplifications on the predicted pK_a shift, ΔpK_a^{eff}, and binding free energy, ΔG^{bind}. Not all combinations of the simplifications are reported, because there is a logical order to them: assuming **no auxiliary movements** is of no benefit unless one also makes the assumption of a **single dielectric boundary** in order to make the number of required FDPB calculations low enough to be feasible.

Simplification: Simpler Charge Patterns

The **point charge pattern** changes the ΔpK_a values significantly compared to the **param19 charge pattern**, from -6.06 for (Glu73, unbound) to +4.45 for (His102, unbound). The **smeared charge pattern** changes the ΔpK_a values somewhat less, from -2.35 for (Asp75, bound) to +4.56 for (His102, bound). Both methods generally disfavor protonation of the carboxylic acids. An investigation of the reason points to these factors: The **param19 charge pattern** (1) puts +0.45 of the +1 charge out on a new hydrogen; (2) it has 4 choices for where to put the hydrogen; and then (3) the

Table 3.5: Effect of titration model simplifications on predicted pK_a shift, ΔpK_a (in pH units), and binding free energy, ΔG^{bind} (in kcal mol^{-1}).

Method: charge arrangemt.:	expt.	PARAM19			point		smeared	
no aux. mvmnts.: single diel. bndry.:		N N	N Y	Y Y	N N	Y Y†	N N	Y Y†
ΔG^{bind}(pH = 7)		14.93	14.23	14.16	16.13	16.19	21.59	21.70
ΔpK_a of:								
Asp54, unbound	≤-1.8	1.48	1.50	-0.78	-0.92	-2.31	0.17	-1.46
Asp54, bound		2.06	2.01	0.47	1.33	-0.87	1.74	-0.17
Glu73, unbound	≤-1.9	-2.18	-2.44	-11.23	-8.24	-8.89	-3.97	-9.30
Glu73, bound		0.69	0.82	-9.59	-2.11	-11.58	0.87	-11.20
Asp75, unbound	-0.9	-6.02	-6.03	-3.24	-4.35	-4.37	-7.27	-4.30
Asp75, bound		-6.66	-6.62	-1.70	-8.04	-3.35	-9.01	-3.01
His102, unbound	0.0	-5.12	-5.80	-5.63	-0.67	-0.18	-5.50	-4.88
His102, bound	≤-1.3	-14.04	-14.04	-13.85	-13.66	-10.18	-9.52	-7.61

† With the point or smeared charging method, the residue has the same shape when deprotonated or protonated, so if no auxiliary movements are allowed, then all states have the same dielectric boundary.

other hydrogens are all allowed to relax (away from the added hydrogen atom). This procedure often finds a favorable interaction (a hydrogen bond, that is) for the added hydrogen. The **point** or **smeared charge patterns**, on the other hand, (1) put the whole +1 charge on the C atom of the carboxyl group, which is less likely to be able to make favorable interactions with nearby negative or polar groups because (2) it is already bonded to 3 other atoms, and so has less surface facing other groups, (3) negative atoms are not likely to be near this carbon in the crystal structure, and (4) both this carbon and all negatively charged atoms are not hydrogens and are therefore not able to move in our procedure.

The simpler charge patterns also affect the electrostatic binding free energy at pH= 7, because the population of titration microstates changes slightly. The 4 residues are mostly deprotonated at pH= 7, but the **smeared charge pattern** is different from the **param19 charge pattern** for the deprotonated as well as the protonated form of histidine.

Simplification: Single Dielectric Boundary

In the PARAM19 parameter set, the van der Waals surfaces of hydrogen atoms never extend far beyond those of the heavy atoms to which they are bonded. So, even when we allow all hydrogen atoms to make auxiliary movements to stabilize each titration state, the dielectric boundary can not change very much due to the movement of hydrogen atoms. So the carboxylic acid ΔpK_a values change by less than 0.3 pH units when the assumption of a **single dielectric boundary** is made.

The one larger effect that the choice of microstate has on the dielectric boundary is actually from the histidine, because we allow flipped and unflipped versions of the histidine ring, and the ring carbons have a larger van der Waals radius than the nitrogens. This causes the binding free energy to drop by -0.9 kcal mol^{-1} uniformly for all microstates. It also causes the histidine ΔpK_a to drop by -0.68 pH units in the unbound state (using the **param19 charge pattern**). The effect of the **single dielectric boundary** simplification will be small in general, except when heavy atoms on the surface of the protein change position or size in any protonation microstates. Note, however, that with this simplification, a surface histidine could cause an error in the binding free energy even if its pK_a was far from the pH at which the binding free energy is calculated, because all possible titration states affect the dielectric boundary, even states which are never significantly populated.

Simplification: No Auxiliary Movement

The binding free energy at pH= 7 is not affected much by this simplification, mainly because all 4 residues are mostly deprotonated at pH= 7. However, some ΔpK_a values are changed a great deal by this simplification, because it causes a few microstates to be heavily penalized for van der Waals clashes between the titration hydrogen and another hydrogen. (Without this assumption, the other hydrogen would move away to relieve the clash.) The worst example of this is that, for the (Glu73 syn2) microstate, the titration hydrogen is 1.4 Å away from an explicit water molecule's hydrogen atom, earning a 5.85

kcal mol^{-1} energy penalty. Without this assumption, the water hydrogen relaxes to a position 2.1 Å away.

3.4 Conclusion

We model multiple-site titration using a system (4 residues of the protein barnase, unbound and bound to barstar) small enough to allow enumeration of all titration states. To calculate free energy difference between titration microstates, we use experimental values for the pK_a values of Asp, Glu, and His model compounds, the pK_a difference between histidine's two protonation sites; ab initio quantum mechanical gas phase energy difference of the anti and syn conformations of protonated acetic acid; and FDPB electrostatic calculations. We use the van der Waals term to disallow protonation microstates with steric clashes that can not be avoided by relaxation of hydrogen atoms only.

We model the effect of titration on the electrostatic binding free energy of barnase and barstar over a range of pH values. In the bound and unbound states, we define effective pK_a^{eff} values for each of the four residues (Asp54, Glu73, Asp75, His102). Both our full model, and the null model as well, compare poorly to the available experimental pK_a values. Our analysis shows that pK_a predictions can depend delicately on relative energies of titration microstates; as the pH drops, the protonation of one residue forces the pK_a values of neighboring titratable residues further down.

We evaluate several simplifying approximations to speed up the calculation of pK_a values. Most of the simplifications that we tested proved to be inadvisable, because they affect the predicted pK_a values too much. We do not recommend simplifying the charge arrangement any more than the PARAM19 parameter set already has (point charges on all heavy atoms and polar hydrogens). We recommend allowing other hydrogens to move in response to each protonation state. Using a single dielectric boundary proved to be a good approximation. However, when heavy atoms on the protein surface have different positions or sizes for different titration states (as a surface histidine does, because of the way we allow it to flip its ring orientation), the assumption of a single dielectric boundary is somewhat inaccurate: a 0.7 kcal mol^{-1} error in the binding free energy, and a 0.7 pH unit error in one pK_a value. The number of titration states of a protein grows exponentially with the number of titratable residues, so it gets unmanageably large for more than a few titratable residues. The motivation for trying these simplifications was to calculate the free energies of all states using fewer FDPB calculations than one per state.

Movement of totally buried residues has no effect on the dielectric boundary, so assuming a single dielectric boundary is not an approximation in that case. So another possibility is that careful enumeration of the number of unique dielectric boundaries used by all titration states could allow enough savings of calculation time to make the treatment of large numbers of titratable residues feasible. The reason that our most careful titration model does not predict pK_a shifts more accurately than the null model could be that non-hydrogen atoms are not allowed to move in concert with protonation state changes. This important aspect of pH-dependent phenomena should be addressed in future computational titration studies.

Chapter 4

Improvements to the Analytical Continuum Electrostatics Method

4.1 Introduction

Accurate calculation of electrostatic free energies in solvent requires the solution of the Poisson-Boltzmann equation. This can be done by finite-difference or other numerical algorithms, but it is computationally expensive. Computational techniques such as molecular modeling, simulation, and ligand design require thousands of energy calculations. Therefore, such efforts have, until very recently, used very easy-to-compute but inaccurate electrostatic energy functions, usually Coulombic or distance-dependent Coulombic interactions. Using these functions typically neglects desolvation effects altogether.

Several analytical approximations to the solution of the Poisson-Boltzmann equation have been proposed recently, building on the Generalized Born approximation of Still et al. [46]. The Analytical Continuum Electrostatics (ACE) method of Schaefer and Karplus [6, 7] is one such approximation, and it is already implemented in the program CHARMM [11].

In this chapter, we describe improvements we have made to the method. We found that its largest errors were typically caused by the few atoms whose atomic solvation energies are predicted to be incorrectly high. We greatly reduced these errors with an automatic procedure that adaptively determines a maximum atomic solvation energy for each molecular system, based on the distribution of uncorrected atomic solvation energies. We also incorporated the approximate treatment of salt effects proposed by Srinivasan *et al.* [47] into ACE, and found that it predicted the effect of salt on binding free energy terms quite accurately.

4.2 Theory

4.2.1 Generalized Born Methods

The electrostatic free energy G_{ES} of a solute molecule in solution can be divided into the electrostatic free energy of the desolvated state plus a solvation free energy:

$$G_{\mathrm{ES}} = G_{\mathrm{ES}}^{\mathrm{desolvated}} + \Delta G_{\mathrm{ES}}^{\mathrm{solv}} \tag{4.1}$$

The desolvated state is defined to have the solute internal dielectric constant ϵ_{i} over all space, so $G_{\mathrm{ES}}^{\mathrm{desolvated}}$ can be obtained by a simple calculation of Coulombic interactions. For point charges, it also contains infinite self energy terms, so one can instead treat each charge i as uniformly distributed over a sphere of radius R_i, as proposed by Born [48]. For the simple system of a lone charge q uniformly distributed over a sphere of radius R, with uniform dielectric constant ϵ outside the sphere, the solution of the Poisson equation

$$\nabla \cdot (\epsilon(\vec{r})\nabla \cdot \phi(\vec{r}) + 4\pi\rho(\vec{r})) = 0 \tag{4.2}$$

is

$$\phi(r) = \begin{cases} \frac{q}{\epsilon R} & \text{for } r \leq R \\ \frac{q}{\epsilon r} & \text{for } r > R \end{cases} \tag{4.3}$$

$$\begin{aligned}\vec{D}(\vec{r}) &= -\epsilon(\vec{r})\nabla\cdot\phi(\vec{r}) & (4.4)\\ \vec{D}(\vec{r}) &= \begin{cases} 0 & \text{for } r \leq R \\ \frac{q}{\epsilon r^2}\hat{r} & \text{for } r > R \end{cases} & (4.5)\end{aligned}$$

$$\begin{aligned}G &= \frac{1}{8\pi}\int \vec{E}\cdot\vec{D}\mathrm{d}^3x \\ G &= \frac{-q^2}{2\epsilon R} & (4.6)\end{aligned}$$

So, the Born self energy of each atom i when desolvated ($\epsilon(\vec{r}) = \epsilon_{\mathrm{i}}$ everywhere) is

$$G = \frac{-q_i^2}{2\epsilon_{\mathrm{i}}R_i} \tag{4.7}$$

So,

$$G_{\mathrm{ES}}^{\mathrm{desolvated}} = \sum_i \frac{q_i^2}{2\epsilon_{\mathrm{i}}R_i} + \sum_{i<j}\frac{q_iq_j}{\epsilon_{\mathrm{i}}r_{ij}} \tag{4.8}$$

The Born self energies in the first sum are independent of configuration, and so they cancel out of any ΔG terms of interest, such as binding or folding free energies.

Notice that the desolvated free energy in equation 4.8 has "self" terms for each atom, and "pair" terms for each pair of atoms. We now consider the effect of the solvent on the total electrostatic free energy. The electrostatic solvation free energy of the solute molecule $\Delta G_{\mathrm{ES}}^{\mathrm{solv}}$, also called the polarization free energy, can also be constructed as a sum of "self" and "pair" terms,

$$\Delta G_{\mathrm{ES}}^{\mathrm{solv}} = \sum_i \Delta G_i^{\mathrm{self}} + \sum_{i<j}\Delta G_{ij}^{\mathrm{int}} \tag{4.9}$$

where $\Delta G_i^{\mathrm{self}}$ is the favorable interaction between the solvent and atom i, and $\Delta G_{ij}^{\mathrm{int}}$ is the solvent screening of the interaction between atoms i and j.

The Generalized Born (GB) method of Still *et al.* [46] provides an approximation to the atom pair terms $\Delta G_{ij}^{\mathrm{int}}$, given the atomic self-energy terms $\Delta G_i^{\mathrm{self}}$. The atomic self-energy terms, which are the interactions between each charge i and the solvent outside of the solvent-accessible surface (SAS), are equivalent (by the principle of superposition) to the solvation free energy of a hypothetical solute with the same SAS shape but no charges except for charge i. This can be obtained analytically in the case of a spherical solute containing one charge. The Born solvation energy of an atom with charge q uniformly distributed over a sphere of radius R is defined as the free energy difference of moving the charge from the desolvated state (the lone atom immersed in dielectric $\epsilon_{\rm i}$) to the fully solvated state (the lone atom immersed in dielectric $\epsilon_{\rm s}$):

$$\Delta G^{\rm solv} = \frac{-\tau q^2}{2R} \tag{4.10}$$

where

$$\tau \equiv \frac{1}{\epsilon_{\rm i}} - \frac{1}{\epsilon_{\rm s}} \tag{4.11}$$

The same result for $\Delta G^{\rm solv}$ is obtained if the charge q is a true point charge at the center of the sphere R. If the point charge were not at the center, this expression would be the monopole term of a multipole expansion [49, 50]. This motivates the definition of a "solvation radius" b_i defined by

$$\Delta G_i^{\rm self} = -\frac{\tau q_i^2}{2b_i} \tag{4.12}$$

for a charge i at a certain location in a certain SAS shape. Since b_i is the radius of a hypothetical spherical solute, centered on charge i, for which i has the same atomic self-energy as in the actual solute, it offers a convenient measure of the effective degree of burial of charge i in the solute. The definition of the solvation radius is illustrated in Figure 4-1.

Now that we have defined the solvation radii, we will first show how they are used to approximate the interaction terms before describing how they are calculated.

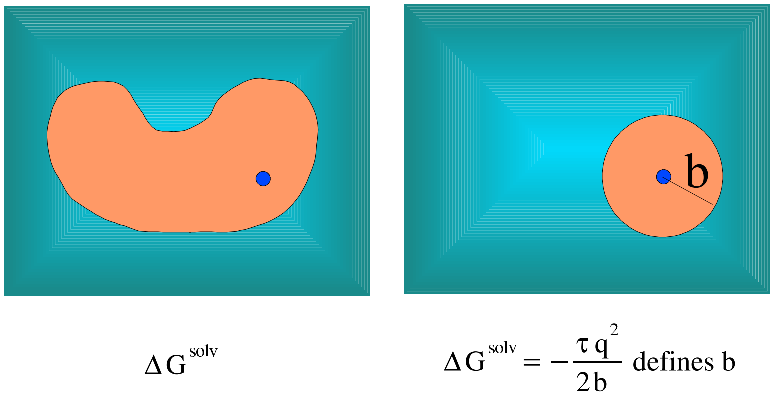
b
ΔG^{solv}
$\Delta G^{\mathrm{solv}} = -\frac{\tau q^2}{2b}$ defines b

Figure 4-1: Definition of Solvation Radius: The solvation radius b of an atomic charge in an arbitrary low-dielectric region is defined as the radius of the spherical region for which the same charge, at the center, would have the same solvation free energy.

4.2.2 Generalized Born Interaction Term

The GB method provides a simple analytical approximation for the atom-pair interaction energies, given the solvation radii of all atoms. The electrostatic solvation free energy $\Delta G_{\mathrm{ES}}^{\mathrm{solv}}$ can be written in a form that combines the atomic self-energies and the atom-pair interaction energies:

$$\Delta G_{\mathrm{ES}}^{\mathrm{solv}} = -\frac{\tau}{2} \sum_i \sum_j \frac{q_i q_j}{f_{ij}^{\mathrm{GB}}} \tag{4.13}$$

(note the full double sum), where the form of f_{ij}^{GB} chosen by Still *et al.* is

$$\begin{aligned} f_{ij}^{\mathrm{GB}} &= (r_{ij}^2 + b_i b_j e^{-D})^{0.5} \\ \text{where } D &= \frac{r_{ij}^2}{4 b_i b_j} \end{aligned} \tag{4.14}$$

The form of f_{ij}^{GB} is not uniquely determined, but this form chosen by Still *et al.* has the correct behavior at the limits of very small or very large separations r_{ij} . At $r_{ij} = 0$, it yields the correct Born energy for a single point charge with magnitude $(q_i + q_j)$. At $r_{ij} \ll b_i = b_j$, it gives the correct Onsager energy for a dipole at the center of a dielectric sphere [49]. At $r_{ij} \gg \max(b_i, b_j)$, it gives the correct Coulombic result. The Generalized

Born equation, Equation 4.13 along with Equation 4.14, approximates the interaction of atoms i and j using only 3 inputs, the solvation radii b_i and b_j, and the separation r_{ij}. This is illustrated in Figure 4-2. Furthermore, the Generalized Born equation has no free parameters, and yet it performs remarkably well. Thus, considerable effort, and many adjustable parameters, have been devoted in recent years to improving the approximation of the atomic self energies rather than tinkering with the Generalized Born equation.

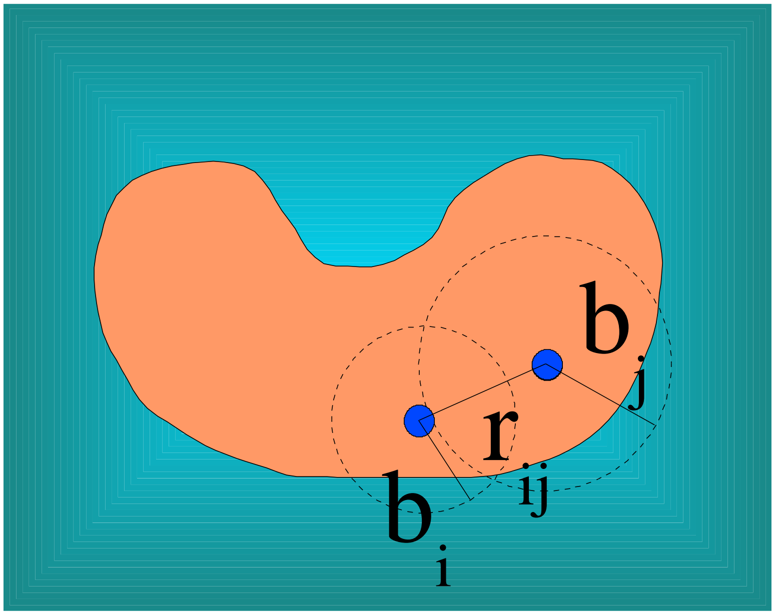

Figure 4-2: Generalized Born Screened Interaction: With no free parameters, the Generalized Born equation approximates the screened interaction between each pair of charges based only on their separation r_{ij} and their solvation radii b_i and b_j.

$$\Delta G_{ij}^{\text{int}} = -\frac{q_i q_j}{\sqrt{r_{ij}^2 + b_i b_j e^{-r_{ij}^2/4b_i b_j}}}$$

4.2.3 ACE Approximation for Atomic Solvation Radii

Some method must be used to supply all atomic solvation radii b_i to be used in the GB equation. To calculate them by finite-difference solution of the Poisson–Boltzmann equation (FDPB) takes as much computer time as calculating all atom-pair interaction terms by FDPB as well. Several analytical methods have been developed to approximate the FDPB result much more quickly: the "pairwise descreening approximation" approach of Hawkins, Cramer, and Truhlar [51, 52], the "ACE" method of Schaefer and Karplus [6, 7], and the "GB/SA Continuum Model" approach of Qiu *et al.* [53]. These methods share two key approximations, the Coulombic field approximation and the pairwise descreening approximation.

Coulombic Field Approximation

The electrostatic energy of the hypothetical state with the whole SAS shape but only atom i charged can be expressed as a volume integral of the electrical energy density over all space, $\int \frac{1}{8\pi}\frac{\vec{D}^2}{\epsilon(\vec{x})}\mathrm{d}^3x$. The Coulombic field approximation lets the dielectric displacement vector $\vec{D}$ due to a charge q at the origin be the spherically symmetric Coulomb field $\vec{D}(\vec{x}) = -\frac{q}{x^2}$ regardless of the solute shape, as shown in Figure 4-3. This is less accurate where the SAS is near the charge or not oriented with its normal pointing toward the charge, because field lines actually bend somewhat from a charge toward the nearest high-dielectric solvent region. Since $\vec{D}$ in the Coulombic field approximation is the same in the solvated and desolvated states, the atomic self-energy simplifies to an integral over all space except the solute region $\mathbf{S}$:

$$\Delta G_i^{\mathrm{self}} = -\frac{\tau}{8\pi}\int_{\not\subset\,\mathbf{S}} \vec{D}^2\mathrm{d}^3x \qquad (4.15)$$

which can be changed into a Born solvation energy term and an integral over the solute region $\mathbf{S}$ except for the van der Waals sphere $\mathbf{V_i}$ of atom i, defined by $|\vec{x}-\vec{x}_i| <= R_i$:

$$\begin{aligned}\Delta G_i^{\mathrm{self}} &= -\frac{\tau}{8\pi}\int_{\not\subset\,\mathbf{V_i}} \vec{D}^2\mathrm{d}^3x + \frac{\tau}{8\pi}\int_{\substack{\subset\,\mathbf{S}\\ \not\subset\,\mathbf{V_i}}} \vec{D}^2\mathrm{d}^3x \qquad (4.16)\\ &= -\frac{\tau q_i^2}{2R_i} + \frac{\tau q_i^2}{8\pi}\int_{\substack{\subset\,\mathbf{S}\\ \not\subset\,\mathbf{V_i}}} |\vec{x}-\vec{x}_i|^{-4}\mathrm{d}^3x\end{aligned}$$

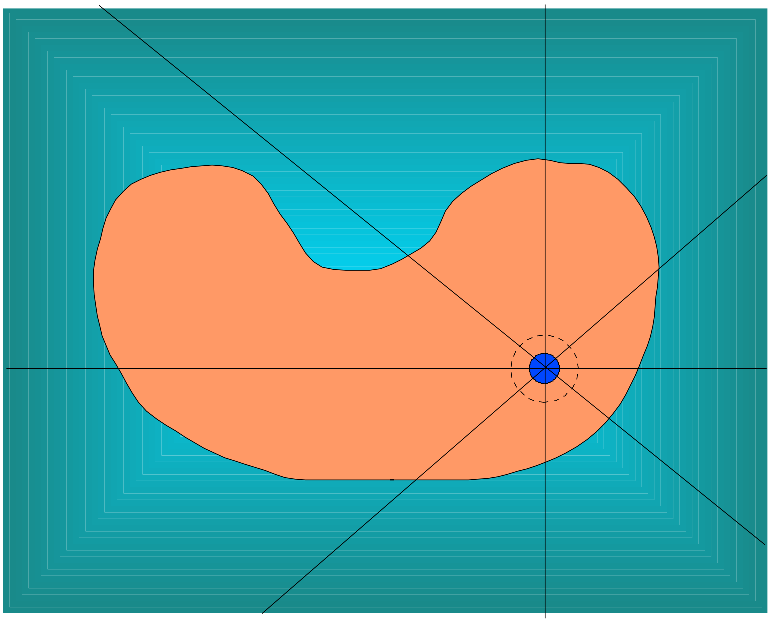

Figure 4-3: The Coulombic Field Approximation of ACE: In order to calculate the atomic solvation energy of an atomic charge, the ACE method assumes that the electric displacement has the Coulombic form

$$\vec{D} \cong \frac{q}{r^2}\hat{r}$$

So the electric field $\vec{E}$ correctly drops in magnitude when crossing into the high dielectric (solvent) region, but the curvature of the field lines (which would depend upon the exact dielectric boundary) is neglected.

In the Coulombic field approximation, a solvation radius depends only on the geometry of the SAS and the location of the atom, and is independent of ϵ_{i} and ϵ_{s}. That is, atomic solvation energies are proportional to τ. We examined the actual dependence of FDPB solvation radii on ϵ_{i} in Figure 4-4. Comparing solvation radii for the two internal dielectric constant values we will use, solvation radii for $\epsilon_{\mathrm{i}} = 4$ average only 3% (with a standard deviation of 1 percentage point) larger than solvation radii for $\epsilon_{\mathrm{i}} = 1$. That standard deviation of 1 percentage point is not noise; a given atom depends more or less strongly on ϵ_{i} across the whole range of ϵ_{i}. And $b(\epsilon_{\mathrm{i}} = 1)$ shows no correlation with $b(\epsilon_{\mathrm{i}} = 4)/b(\epsilon_{\mathrm{i}} = 1)$, so the solvation radii themselves are not correlated with their sensitivity to ϵ_{i}.

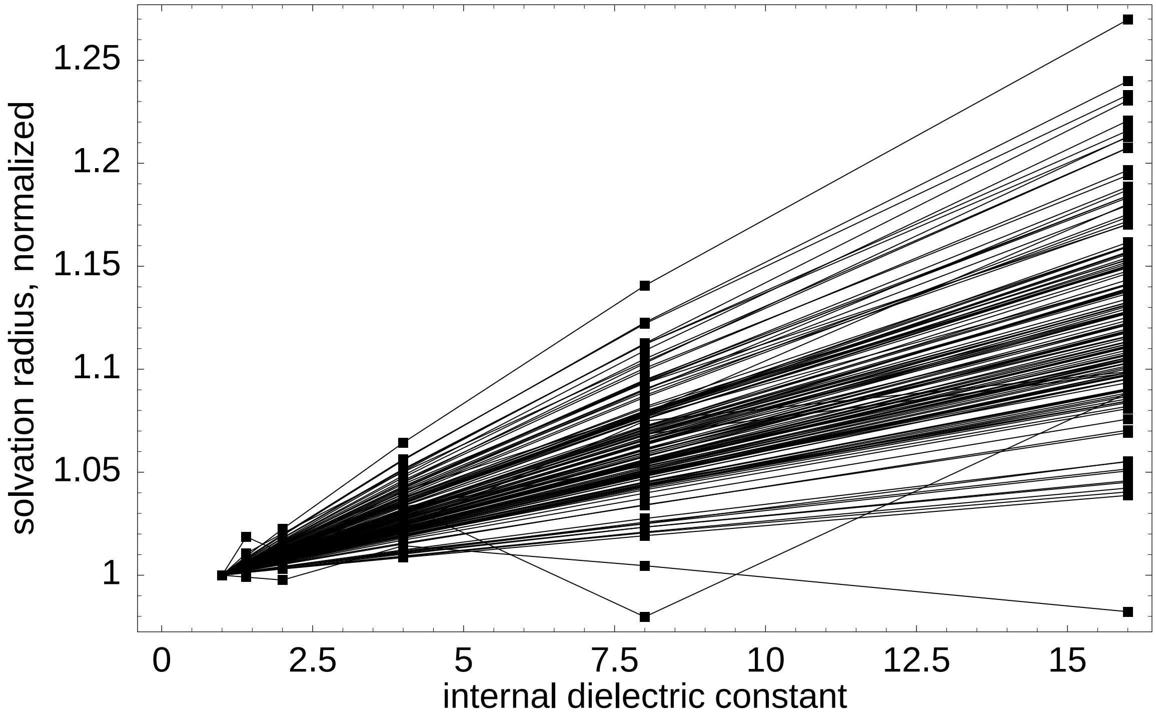

Figure 4-4: Dependence of FDPB solvation radii b_i vs. internal dielectric constant ϵ_{i}. The solvation radii are given as a ratio relative to the value at $\epsilon_{\mathrm{i}} = 1$. In the Coulombic field approximation, solvation radius is independent of ϵ_{i}. Here we see that this is true to within 3% (with a standard deviation of 1 percentage point) between $\epsilon_{\mathrm{i}} = 1$ and $\epsilon_{\mathrm{i}} = 4$. The solvation radi are for a subset of 93 atoms (every tenth atom in the coordinate file) in the A chain of the cyanovirin unit AB, unbound and bound in the domain-swapped homodimer; so there are 186 data points at each ϵ_{i} value.

Pairwise Descreening Approximation

The pairwise descreening approximation replaces the integral over the solute volume $\mathbf{S}$ less the atomic van der Waals volume $\mathbf{V_i}$ with individual terms for each other atom $k \neq i$ in the solute:

$$\Delta G_i^{\mathrm{self}} = -\frac{\tau q_i^2}{2R_i} + \sum_{k \neq i} \Delta G_{ik}^{\mathrm{self}} \tag{4.17}$$

Each $\Delta G_{ik}^{\mathrm{self}}$ is the electrostatic free energy cost of changing the volume of atom k from ϵ_{s} to ϵ_{i} in the presence of charge i. This approximation amounts to ignoring the fact that atoms' van der Waals radii have gaps and overlaps to varying degrees. A cartoon of the pairwise descreening approximation is shown in Figure 4-5.

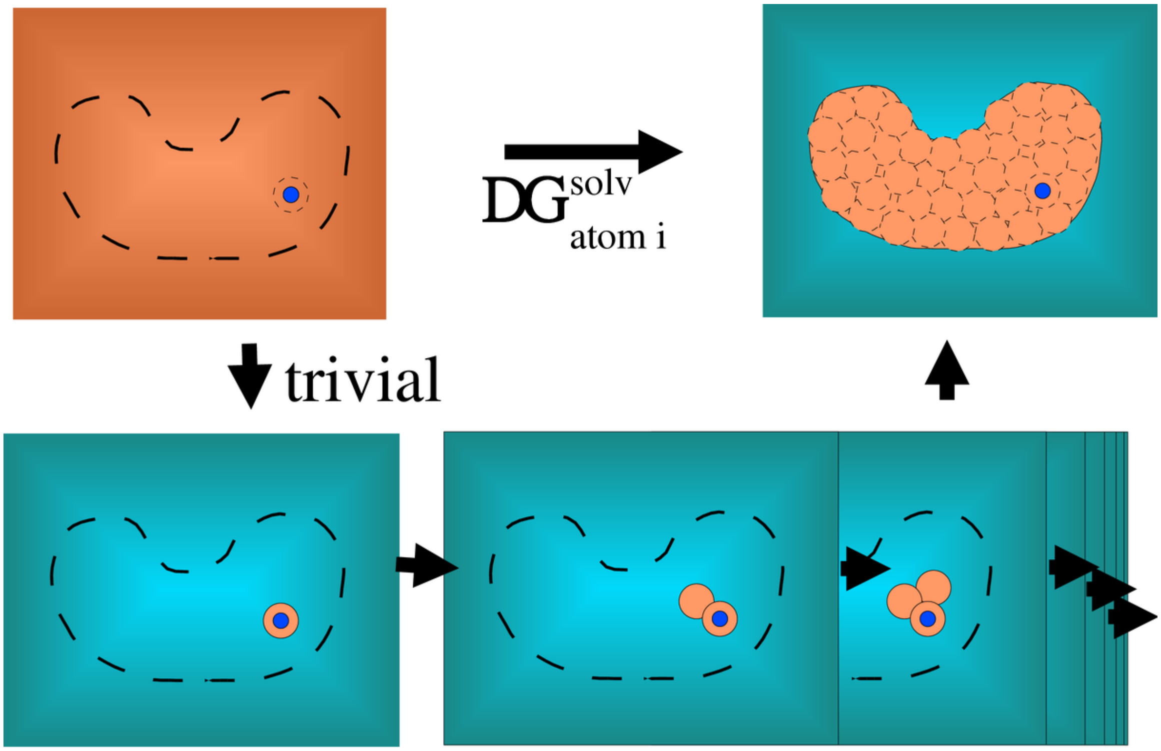

Figure 4-5: The Pairwise Descreening Approximation of ACE: The approximation assumes that every atom has an independent contribution to $\Delta G_i^{\mathrm{self}}$, the solvation free energy of atom i. The contribution of each atom k is the free energy cost of desolvating only the volume of atom k.

Methods other than ACE

As mentioned before, the "pairwise descreening approximation" approach of Hawkins, Cramer, and Truhlar [51, 52], and the "GB/SA Continuum Model" approach of Qiu *et al.* [53] share with the ACE method of Schaefer and Karplus [6, 7] two key approximations, the Coulombic field approximation and the pairwise descreening approximation. The three methods use different approximations for the $\Delta G_{ik}^{\text{self}}$ terms.

The "surface GB model" of Ghosh *et al.* [54] calculates an approximation of a surface integral over the SAS which is equivalent to the above volume integral over the solute volume. The "Modified Tanford-Kirkwood" approach of Havranek and Harbury [25] is an alternative approach to approximating both the atomic solvation energies and the atom-pair interaction energies based on simplifying the actual solute geometry (relative to an atom pair) to a sphere containing the atom pair.

ACE Formulae

In the present work, we chose to use the ACE method of Schaefer and Karplus [6, 7]. In order to calculate atomic solvation energies, ACE treats the charge i as being spread out in a Gaussian pattern rather than a point or spherical-shell charge, and treats the degree of desolvation by atom k as a Gaussian function as well, rather than a spherical step function at its van der Waals radius R_k. The ACE form for the $\Delta G_{ik}^{\text{self}}$ terms is based on the analytical solution of a double integral over these two Gaussian distributions. The first parameter of the ACE model is the density smoothing parameter α, which determines how wide and low these Gaussian distributions are. The remaining parameters, one per atom type, are effective atom volumes V_k which determine the net volume desolvated by each atom k.

We will now summarize the ACE method in Equations 4.18 to 4.31, as given in reference [7] and programmed in CHARMM version 27a2. For a solute with N_{atoms} atoms, the ACE free energy is

$$G_{\mathrm{ES}} = \sum_i G_{i0}^{\mathrm{self}} + \sum_{i \neq k} G_{ik}^{\mathrm{self}} + \sum_{i<j} G_{ij}^{\mathrm{int}} \tag{4.18}$$

$$\begin{aligned} G_{\mathrm{ES}} = & \sum_i q_i^2 \left\{ \frac{1}{2\epsilon_{\mathrm{s}} R_i} - \frac{\tau}{\omega_{ii}^*} \right\} \\ & + \tau \sum_{i \neq k} q_i^2 \left\{ \frac{exp(-r_{ik}^2/\sigma_{ik}^2)}{\omega_{ik}} + \frac{V_k}{8\pi} \left(\frac{r_{ik}^3}{r_{ik}^4 + \mu_{ik}^4} \right)^4 \right\} \\ & + \sum_{i<j} q_i q_j \left\{ \frac{f_{ij}^{\mathrm{nb}}}{\epsilon_{\mathrm{i}} r_{ij}} - \frac{\tau}{(r_{ij}^2 + b_i b_j exp(-r_{ij}^2/4b_i b_j))^{1/2}} \right\} \end{aligned} \tag{4.19}$$

where i, j, and k are atom numbers from 1 to N_{atoms}; where q_i, R_i, and V_i are the partial charge, radius, and volume of atom i; where r_{ij} is the distance between atoms i and j; and where ϵ_{i} and ϵ_{s} are the dielectric constants inside the solute, and outside it in the solvent. We already defined $\tau \equiv (1/\epsilon_{\mathrm{i}}) - (1/\epsilon_{\mathrm{s}})$ in Equation 4.11. The parameters ω_{ii}^*, ω_{ik}, σ_{ik}, and μ_{ik} are defined in Equations 4.23 to 4.31.

The first term of the first sum in Equation 4.19, $q_i^2/(2\epsilon_{\mathrm{s}} R_i)$, is the energy of atom i alone in solvent (solvent dielectric ϵ_{s} filling all space except for the atomic sphere R_i). The next term, $q_i^2 \tau/\omega_{ii}^*$, is the effect of changing the charge distribution from a spherical shell at R_i to a gaussian distribution. The terms of the second sum in Equation 4.19, for $i \neq k$ are the effect of changing the dielectric constant in the volume occupied by atom k from ϵ_{s} to ϵ_{i}. So the sum of all those self terms (the first 2 sums in Equation 4.19) is an approximation of Equation 4.17 for the free energy of atom i in the actual solute shape, with all other charges turned off. The important thing to note is that the second term of the second sum in Equation 4.19, except at small r_{ik}, has the dependence $\sim V_k r_{ik}^{-4}$ because it is an approximation of the integral of $|\vec{x} - \vec{x}_i|^{-4}$ over the solute volume in Equation 4.17.

The first pair term is the Coulombic interaction of atoms i and j in the desolvated state. The second pair term is the generalized Born equation for the screening of the interaction by solvent. Reference [6] lets f_{ij}^{nb} equal zero for bonded and 1-3 atom pairs, 0.4 for 1-4 atom pairs, and 1 otherwise (by setting the `NBXMOD 5 E14FAC 0.4` CHARMM energy parameters). Since we want ACE to approximate the FDPB result, we let $f_{ij}^{\mathrm{nb}} = 1$ for all atom pairs (by setting the `NBXMOD 0 E14FAC 1` CHARMM energy parameters), so that ACE calculates the full interaction between all atom pairs.

The ACE atomic solvation energy is

$$\Delta G_i^{\mathrm{self}} = -\frac{\tau q_i^2}{2R_i} - \frac{\tau q_i^2}{\omega_{ii}^*} + \sum_{k \neq i} G_{ik}^{\mathrm{self}} \tag{4.20}$$

The uncorrected ACE solvation radius of each atom depends on its ACE atomic solvation energy,

$$b_i^{\mathrm{raw}} \equiv \frac{-\tau q^2}{2\Delta G_i^{\mathrm{self}}} \tag{4.21}$$

We will discuss below why and how b_i^{raw} values are corrected to become solvation radii b_i for use in the Generalized Born interactions (the last term of Equation 4.19). Briefly, one can apply a “tangential cutoff at b_0” according to Equation 4.42; or a “flat cutoff at b_0” according to Equation 4.43; or a “flat cutoff at b_a” according to Equation 4.48. We define $b_0 \equiv ((3/4\pi)\sum_i V_i)^{1/3}$, the radius of a sphere with the volume of the whole solute, in Equation 4.41, and our adaptive maximum solvation radius b_a is described in Section 4.3.2. These three types of cutoff are pictured in Figure 4-6.

Very small atomic radii cause large errors, so the atomic radii R_i are defined by:

$$R_i = \max(R_i^{\mathrm{vdW}}, \max_{j \text{ bonded to } i}(R_j^{\mathrm{vdW}} - l_{ij})) \tag{4.22}$$

where l_{ij} is the equilibrium bond length for the atom types of atoms i and j.

The parameters ω_{ii}^*, ω_{ik}, σ_{ik}, and μ_{ik} depend on the atomic radii and volumes, and on the density smoothing parameter α:

$$\frac{1}{\omega_{ik}} = \frac{V_k}{\pi^2(\alpha_{ik}R_k)^4}(Q_{ik} - \arctan Q_{ik}) \tag{4.23}$$

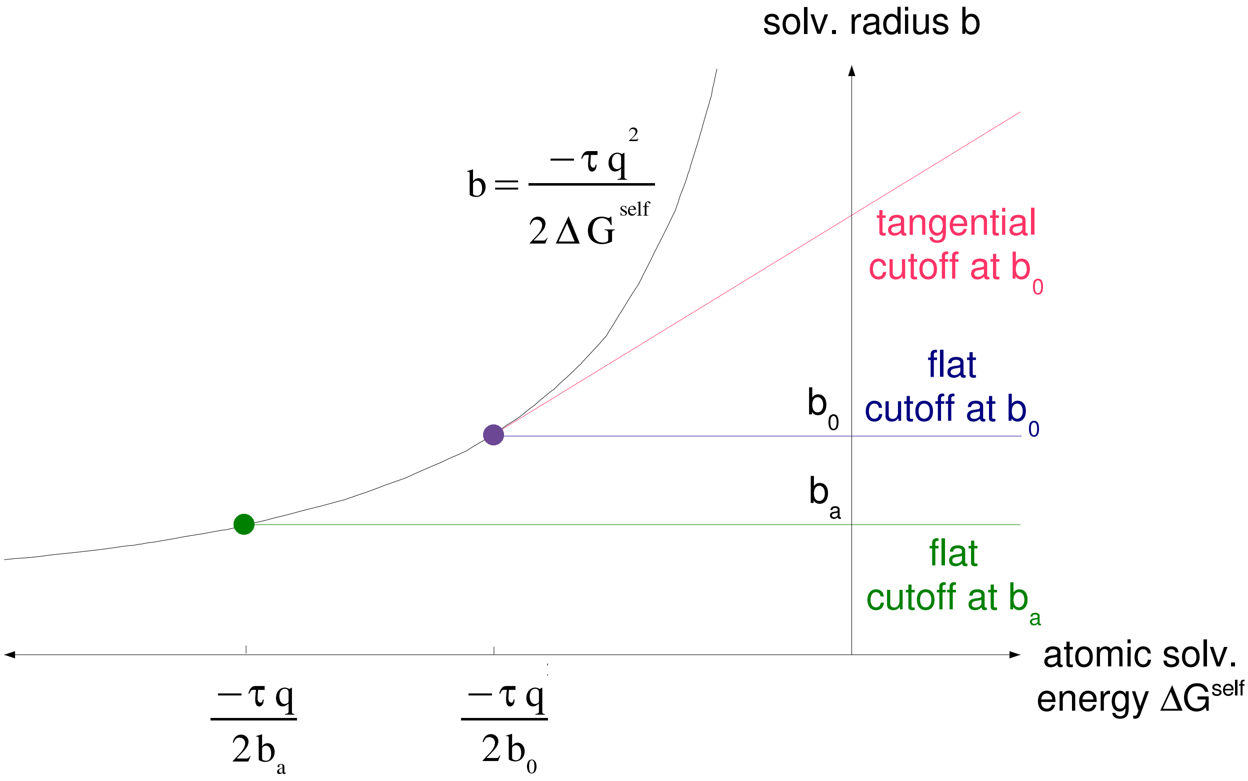

Figure 4-6: Depiction of the three types of solvation radii cutoff used. The curve is the usual definition of the solvation radius, and the three straight lines are three different cutoffs used when determining the solvation radius from the atomic solvation energy. They are defined by Equations 4.42, 4.43, and 4.48. Note that the flat cutoffs are actually connected to the curve by short splines.

$$\frac{1}{\omega_{ii}^*} = \frac{4}{3\pi\alpha_{ii}^4 R_i}(Q_{ii} - \arctan Q_{ii}) \tag{4.24}$$

$$\sigma_{ik}^2 = \frac{3(Q_{ik} - \arctan Q_{ik})}{(3 + f_{ik}^{\mathrm{ACE}})Q_{ik} - 4\ \arctan Q_{ik}}(\alpha_{ik} R_k)^2 \tag{4.25}$$

$$\mu_{ik} = \frac{77\pi\sqrt{2}}{512(1 - 2\pi^{3/2}\sigma_{ik}^3 R_i/(\omega_{ik} V_k))} R_i \tag{4.26}$$

$$Q_{ik} = \frac{q_{ik}^2}{(2q_{ik}^2 + 1)^{1/2}} \tag{4.27}$$

$$q_{ik}^2 = a_i^2(\alpha_{ik} R_k)^2 \tag{4.28}$$

$$f_{ik}^{\mathrm{ACE}} = \frac{2}{q_{ik}^2 + 1} - \frac{1}{2q_{ik}^2 + 1} \tag{4.29}$$

$$a_i = \sqrt{\frac{\pi}{2}}\frac{1}{R_i} \tag{4.30}$$

$$\alpha_{ik} = \max(\alpha, R_i/R_k) \tag{4.31}$$

Using the CHARMM [11] PARAM19 [10] united atom parameter set, both of our protein binding test systems have 17 distinct atom types. So, including α , the ACE model has 18 adjustable parameters. The "default" values we used for these 18 parameters are given in reference [7], distributed with CHARMM version 27a2, and given in Table 4.1 of the present work. They were obtained by minimizing the solute volume fluctuations for a set of 12 protein structures (i.e. making the sum of the Gaussian distributions representing each atom's desolvation as similar as possible to the step function with value 1 inside the SAS and 0 outside) [55]. In Chapter 5, we describe a method to find parameters which minimize errors from FDPB results in a variety of ways.

4.2.4 Generalized Born Salt Treatment

A correction to the Generalized Born equation (Equation 4.13), to account for the presence of salt in the solvent, has been proposed by Srinivasan *et al.* [47]. The factor

Table 4.1: Effective volumes and van der Waals radii for use with ACE and the PARAM19 polar hydrogen parameter set. They are compatible with a density smoothing parameter $\alpha = 1.2$. This table is adapted from reference [7]; these parameters were also distributed with CHARMM version 27a2. They were obtained by Schaefer, Bartels, and Karplus by minimizing the solute volume fluctuations for a set of 12 protein structures [7, 55]. Atom types with only 0 or 1 instances in the test set of reference [7] are omitted.

Atom Type	V ($\AA^3$)	R^{vdW} ($\AA$)	Description
H	0.310	0.80	H which can H-bond to neutral atom
HC	0.529	0.60	H which can H-bond to charged atom
HT[a]	0.0	0.80	TIPS3P water hydrogen
LP[a]	0.0	0.2245	ST2 lone pair
C	12.403	2.10	Carbonyl C
CH1E	12.257	2.365	Extended atom C with one H
CH2E	35.356	2.235	Extended atom C with two H
CH3E	40.947	2.165	Extended atom C with three H
CR1E	18.583	2.10	Extended atom C with one H, in aromatic ring
N	0.0	1.60	Peptide N bound to no H
NR	16.611	1.60	N bound to no H, in aromatic ring
NP	0.0	1.60	Pyrole N
NH1	1.708	1.60	Peptide N bound to one H
NH2	18.677	1.60	Peptide N bound to two H
NH3	15.521	1.60	N bound to three H
NC2	19.336	1.60	Charged guanidinium N bound to two H
O	14.375	1.60	Carbonyl O
OC	16.404	1.60	Carboxy O
OH1	21.427	1.60	Hydroxy O
OH2[b]	29.700	1.7398	ST2 water O
OT[b]	29.700	1.60	TIPS3P water O
S	15.196	1.89	Sulphur bound to no H
FE	0.411	0.65	Iron

[a] Volume set to zero.

[b] Volume derived from water density under standard conditions (1 atm, 298 K).

$\tau = \frac{1}{\epsilon_\mathrm{i}} - \frac{1}{\epsilon_\mathrm{s}}$ is replaced by,

$$\tau_{ij} = \frac{1}{\epsilon_\mathrm{i}} - \frac{e^{-k\bar{\kappa}f_{ij}^\mathrm{GB}}}{\epsilon_\mathrm{s}} \tag{4.32}$$

which can differ for each pair of atoms i, j. Here f_{ij}^GB is the Generalized Born equation term given in Equation 4.14. Srinivasan *et al.* found that a value of 0.73 for the scaling parameter k gave acceptable results for predicting solvation energies. Here

$$\bar{\kappa} = \sqrt{\frac{8\pi \mathrm{e}^2 N_\mathrm{A}(I/(\mathrm{mol\ L}^{-1}))}{1000 k_\mathrm{B} T}} \tag{4.33}$$

is the modified Debye-Hückel screening parameter, e is the fundamental charge, and I is the ionic strength in mol L^{-1}. Note that we use this to correct the atomic self energy ($i = j$) terms as well, where f_{ii}^GB is the atomic solvation radius uncorrected for salt effects, and τ_{ii} substituted into the GB equation yields the atomic solvation energy corrected for salt effects.

4.2.5 Component Analysis

Now that we have reviewed how the electrostatic free energy G_ES is obtained, the electrostatic free energy of binding is simply defined as the difference

$$\Delta G_\mathrm{ES}^\mathrm{bind} \equiv G_\mathrm{ES}^\mathrm{bound} - G_\mathrm{ES}^\mathrm{unbound} \tag{4.34}$$

where the unbound state consists of the two binding partners, infinitely separated by solvent.

We are interested in predicting how each part of a protein-protein binding system contributes to the electrostatic binding free energy. All proteins can be subdivided into these three groups per residue: side chain, amino, and carbonyl. In the polar hydrogen model of PARAM19, an amino group includes C_α, the backbone N, and its H atom; a carbonyl group contains the backbone C and O atoms; and a side chain group contains the atoms C_β and beyond. We choose to include each N- or C-terminal blocking group in its residue's amino group if it contains a nitrogen, or carbonyl group if it contains a carbon.

The electrostatic binding free energy can be constructed out of self (or "solvation") terms for each group m, and interaction terms for each group pair m, n:

$$\Delta G_{\mathrm{ES}}^{\mathrm{bind}} = \sum_{m} \Delta G_{m}^{\mathrm{bind\ solv}} + \sum_{m \neq n} \Delta G_{mn}^{\mathrm{bind\ int}} \tag{4.35}$$

A group solvation binding term $\Delta G_{m}^{\mathrm{bind\ solv}}$ is the cost, with only group m charged, of desolvating the volume of the other binding partner. For groups m and n on opposite binding partners, the group interaction binding term $\Delta G_{mn}^{\mathrm{bind\ int}}$ for this "direct interaction" is equal to their screened interaction in the bound state. For groups m and n on the same binding partner, the group interaction binding term $\Delta G_{mn}^{\mathrm{bind\ int}}$ for this "indirect interaction" is equal to the screening of their interaction by the exclusion of solvent from the volume of the other binding partner.

The group binding terms can, in turn, be constructed out of atomic solvation and interaction terms (i and j are atom numbers):

$$\Delta G_{m}^{\mathrm{bind\ solv}} = \sum_{i \in m} (\Delta G_{i}^{\mathrm{bound}} - \Delta G_{i}^{\mathrm{unbound}}) + \sum_{i \in m} \sum_{j \in m} (\Delta G_{ij}^{\mathrm{bound}} - \Delta G_{ij}^{\mathrm{unbound}}) \tag{4.36}$$

and

$$\Delta G_{mn}^{\mathrm{bind\ int}} = \sum_{i \in m} \sum_{j \in n} (\Delta G_{ij}^{\mathrm{bound}} - \Delta G_{ij}^{\mathrm{unbound}}) \tag{4.37}$$

If groups m and n are not in the same half of the binding complex, the $\Delta G_{ij}^{\mathrm{unbound}}$ terms will be zero.

In our model, the hydrophobic isostere of a group is the same as the group itself except that all of its charges are set to zero. The mutation term of group m, i.e. the change to the binding free energy of mutating group m from its hydrophobic isostere, is:

$$\Delta G_{m}^{\mathrm{bind\ mut}} = \Delta G_{m}^{\mathrm{bind\ solv}} + \sum_{n \neq m} \Delta G_{mn}^{\mathrm{bind\ int}} \tag{4.38}$$

and the two terms on the right are called the group solvation term and the group total-interaction term of the binding free energy. Note that since the mutation terms of both group m and n include their entire interaction, the mutation terms do not sum up to the binding free energy ΔG^{bind}.

The contribution term of group m takes only half of its interactions:

$$\Delta G_m^{\text{bind cont}} = \Delta G_m^{\text{bind solv}} + \frac{1}{2}\sum_n \Delta G_{mn}^{\text{bind int}} \tag{4.39}$$

so that the contribution terms add up to the binding free energy:

$$\Delta G^{\text{bind}} = \sum_m \Delta G_m^{\text{bind cont}} \tag{4.40}$$

4.3 Methods

4.3.1 Structure Preparation

Two protein-protein binding pairs were used in this study: the cyanovirin-N (CV-N) domain-swapped homodimer, and part of the protease-resistant core of the HIV-1 glycoprotein gp41, three inner helices (ABC) binding to one of the outer helices (D).

Several other proteins and protein complexes were used only to test our algorithms for the adaptive maximum solvation radius and for pseudo-symmetry detection. They are the amino-terminal domain of phage 434 repressor (PDB entry 1R69) [56], rabbit uteroglobin (PDB entry 1UTG) [57], arc repressor from bacteriophage P22 (PDB entry 1PAR) [58], human class II MHC protein HLA-DR1 (PDB entry 1DLH) [59], and sperm whale myoglobin (PDB entry 1MBD) [60].

Selected crystallographic subunits from the X-ray crystal structures were used. We use the PARAM19 atomic charge and van der Waals radius parameters in both the FDPB and ACE models. Polar hydrogens were built onto the crystal structure using the HBUILD facility [43] in the CHARMM package [11].

Cyanovirin-N Structure

CV-N, isolated from the cyanobacterium *Nostoc ellipsosporum*, is a potent virucide against human immunodeficiency virus (HIV) in its monomeric form [61, 62]. A domain-swapped homodimeric form is also formed at low pH under certain conditions (reverse-phase HPLC purification, or 26% isopropanol used for crystallization). The monomeric and dimeric forms are both stable at low pH, but only the monomeric form is stable at neutral pH. Our test system is the rigid binding of the two subunits (AB and A'B') making up the domain-swapped homodimer of the 1.5 Å X-ray crystal structure 3EZM [62], pictured in Figure 4-7. It has been shown that the two halves of the homodimer (AB' and A'B) are oriented differently in solution than in the X-ray crystal structure [63], but most of the binding interface is within the halves rather than between them.

We aligned and compared the nuclear magnetic resonance (NMR) structure of the monomer at pH 6.1, 2EZM [61], to one half of the X-ray crystal structure of the homodimer at pH 4.4, 3EZM [62]. There are no major differences: except for the linker between the A and B domains of each unit, no backbone atoms differ by more than about 1 Å. This similarity suggests that the binding of two units to form a homodimer has a mechanism similar to the assumed last stage of monomer folding, the docking of the two domains. So, our rigid binding system can also be considered a simple model for the last step of folding either the dimer or the monomer.

The only contact between the two halves of the homodimer other than the linker region is between the carboxylic acid moieties of Glu41 and Glu41', which directly face each other. Obviously, this is possible at the 4.4 pH of the X-ray crystal structure because these side chains are protonated, and form hydrogen bonds with each other.

Based on an inspection of the 3EZM structure with an eye to satisfying hydrogen-bonding patterns, the protonation states of 4 residues were changed. The His90 and His90' were protonated at $H_{\delta1}$. The carboxylic acid moieties of Glu41 and Glu41', which directly face each other, were both protonated, adding Glu41 $H_{\epsilon1}$ and Glu41' $H_{\epsilon2}$. Glu41 and Glu41' form the only contact between the two halves of the homodimer other than

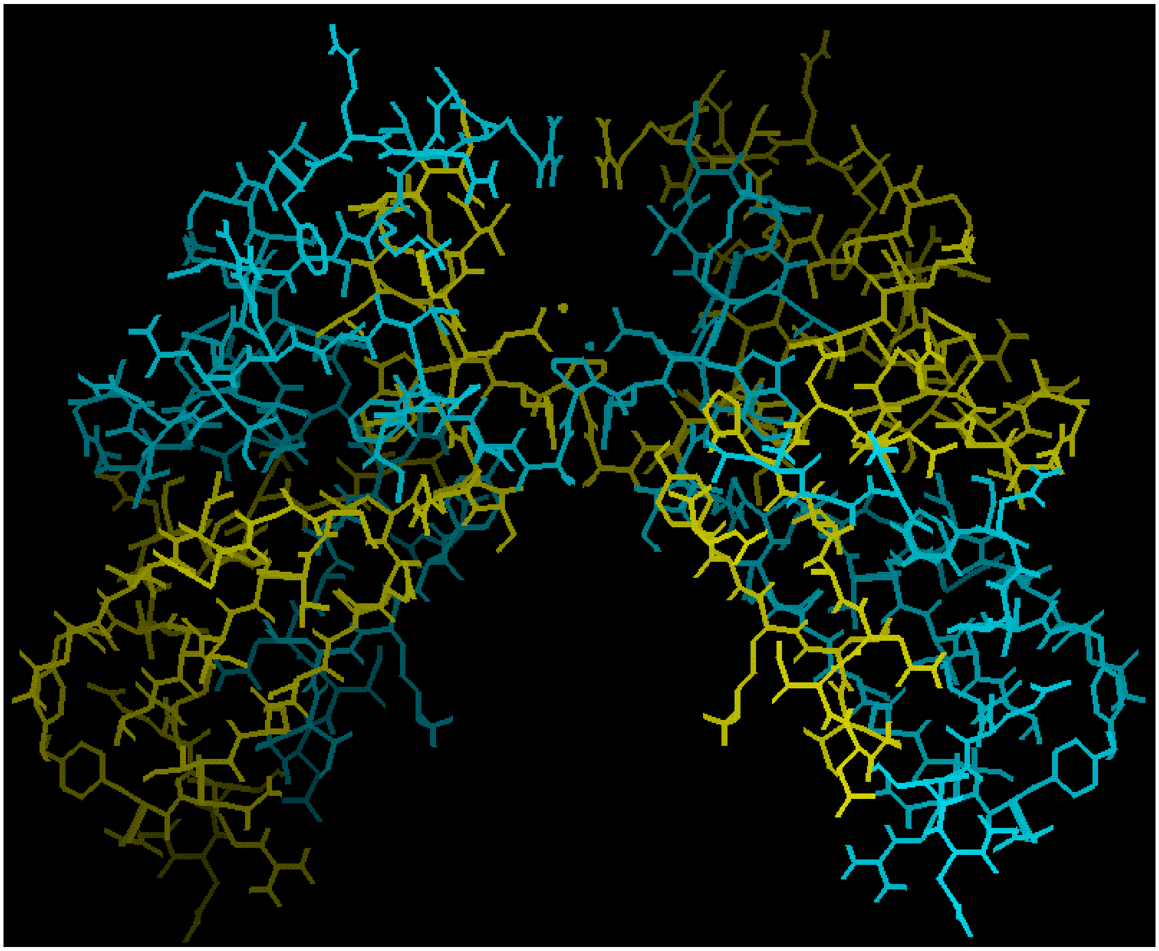

Figure 4-7: Structure of cyanovirin domain-swapped homodimer (from PDB entry 3EZM [62]), color-coded for the rigid binding of the two subunits (AB and A'B'). All bonds are shown except those to the atoms Glu41 $H_{\epsilon 1}$ and Glu41' $H_{\epsilon 2}$ (the only asymmetry in our structure), which are shown as points near the center of the picture. (Figure made with QUANTA from Molecular Simulations, Inc. (San Diego, CA).)

the linker region. to take advantage of hydrogen-bonding opportunities.

The two units are symmetry-related in the 3EZM structure. The only source of assymetry was Glu41 $H_{\epsilon1}$ and Glu41' $H_{\epsilon2}$. The HBUILD function of CHARMM placed all but one of the symmetry-related hydrogen pairs within 0.4 Å of symmetry; the one exception, Thr83' $H_{\gamma1}$, which is far from Glu41 $H_{\epsilon1}$ and Glu41' $H_{\epsilon2}$, was forced into symmetry with Thr83 $H_{\gamma1}$.

gp41 Structure

Our second test system is part of the protease-resistant core of the HIV-1 glycoprotein gp41, the 3 inner helices (ABC) binding to one of the outer helices (D). This rigid binding system, pictured in Figure 4-8, is a simple model of the assumed last stage of folding, in which the outer helices dock onto the inner helices. The coordinates are taken from the 2.0 Å X-ray crystal structure 1AIK [64].

4.3.2 Limiting Large Solvation Radii

Consider the calculation of the ACE atomic solvation energy for a fairly well-buried atom, as illustrated in Figure 4-5: the first term of Equation 4.17 is large and negative for the free energy gain of allowing solvent to fill space right up to the atom's van der Waals radius. The sum in the second term adds small positive terms for the free energy cost of removing solvent from the location of each other atom in the solute. If the atom is fairly well-buried, the sum of these small positive terms must come out to very nearly, but not quite, cancel the large negative first term. So it should not be surprising that in some cases, due to the approximations of the ACE method, the sum in the second term turns out to be larger in magnitude than the first term, in which case the atom would be assigned a positive atomic solvation energy. This is physically impossible. Or, even if the sum in the second term is smaller in magnitude than the first term, but *too* close to it, then the atom would be assigned an atomic solvation energy near zero, and so a very large solvation radius, much bigger than the size of the entire solute molecule. For

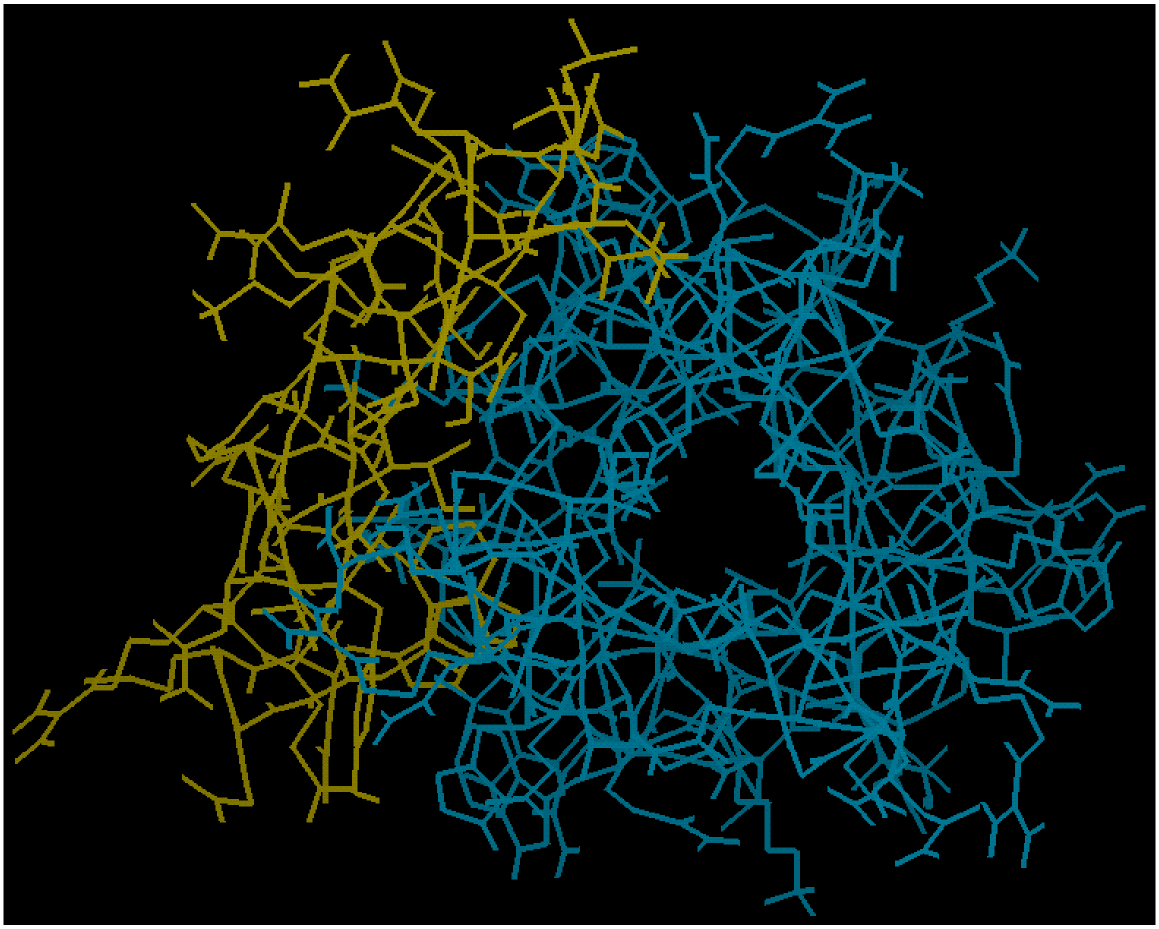

Figure 4-8: Structure of the HIV-1 glycoprotein gp41 (from PDB entry 1AIK [64]), color-coded for the rigid binding of the 3 inner helices (ABC) to one of the outer helices (D). (Figure made with QUANTA from Molecular Simulations, Inc. (San Diego, CA).)

our cyanovirin and gp41 test systems, Figures 4-9 and 4-10 use color-coding to show the solvation radii calculated by ACE as in reference [7] and CHARMM version 27a2, without our improvements. A few atoms, with too-high solvation radii, are a large source of error for the ACE method in this case, because their interactions are essentially unscreened.

Effective System Radius From Sum of Atom Effective Volumes

In the ACE implementation of reference [7] and CHARMM version 27a2, an effective system radius b_0 is used as a maximum reasonable solvation radius. It is based on the sum of all the atom effective volumes V_i, by:

$$b_0 \equiv (\frac{3}{4\pi} \sum_i V_i)^{1/3} \tag{4.41}$$

If atom i has solvation energy greater than $-\tau q_i^2/2b_0$, then the solvation radius is only allowed to increase linearly with solvation energy past that point, rather than inversely. We call this a "tangential cutoff" at b_0 because, as pictured in Figure 4-6, the solvation radius vs. atomic solvation energy curve follows the tangent at b_0 for solvation radii $\geq b_0$:

$$b_i = \begin{cases} -\frac{\tau q_i^2}{2\Delta G_i^{\text{self}}} & \text{if } \Delta G_i^{\text{self}} \leq -\tau q_i^2/2b_0 \\ b_0(2 + \Delta G_i^{\text{self}} \frac{2b_0}{\tau q_i^2}) & \text{otherwise} \end{cases} \tag{4.42}$$

This prevents infinite or negative solvation radii, but we found that it still allows solvation radii that are unreasonably large and therefore greatly degrade accuracy (see Table 4.3). We also tried using a flat cutoff at b_0, as pictured in Figure 4-6, so that the maximum allowed solvation radius is b_0 :

$$b_i = \begin{cases} -\frac{\tau q_i^2}{2\Delta G_i^{\text{self}}} & \text{if } \Delta G_i^{\text{self}} \leq -\tau q_i^2/2b_0 \\ b_0 & \text{otherwise} \end{cases} \tag{4.43}$$

But the flat cutoff at b_0 still allows unreasonably large solvation radii, because the system effective radius is substantially larger than the largest actual solvation radius as calculated

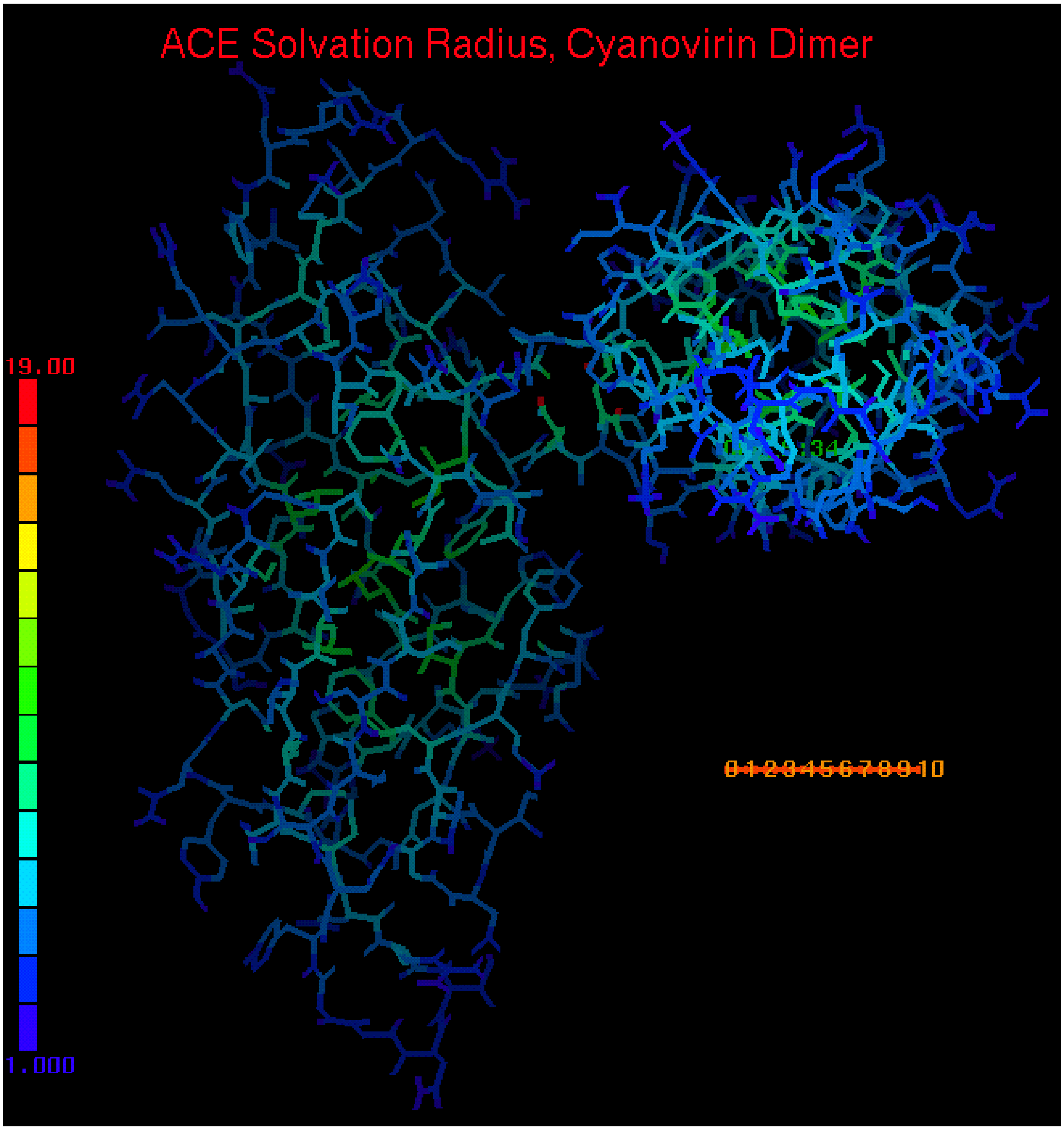

Figure 4-9: Cyanovirin domain-swapped homodimer, color-coded by ACE atomic solvation radius. The color-code key is on the left; a 10-Å long ruler is at the bottom right. These solvation radii are from ACE as in reference [7] and CHARMM version 27a2, without our improvements. The three red atoms' solvation radii are off the scale. (Figure made with QUANTA from Molecular Simulations, Inc. (San Diego, CA).)

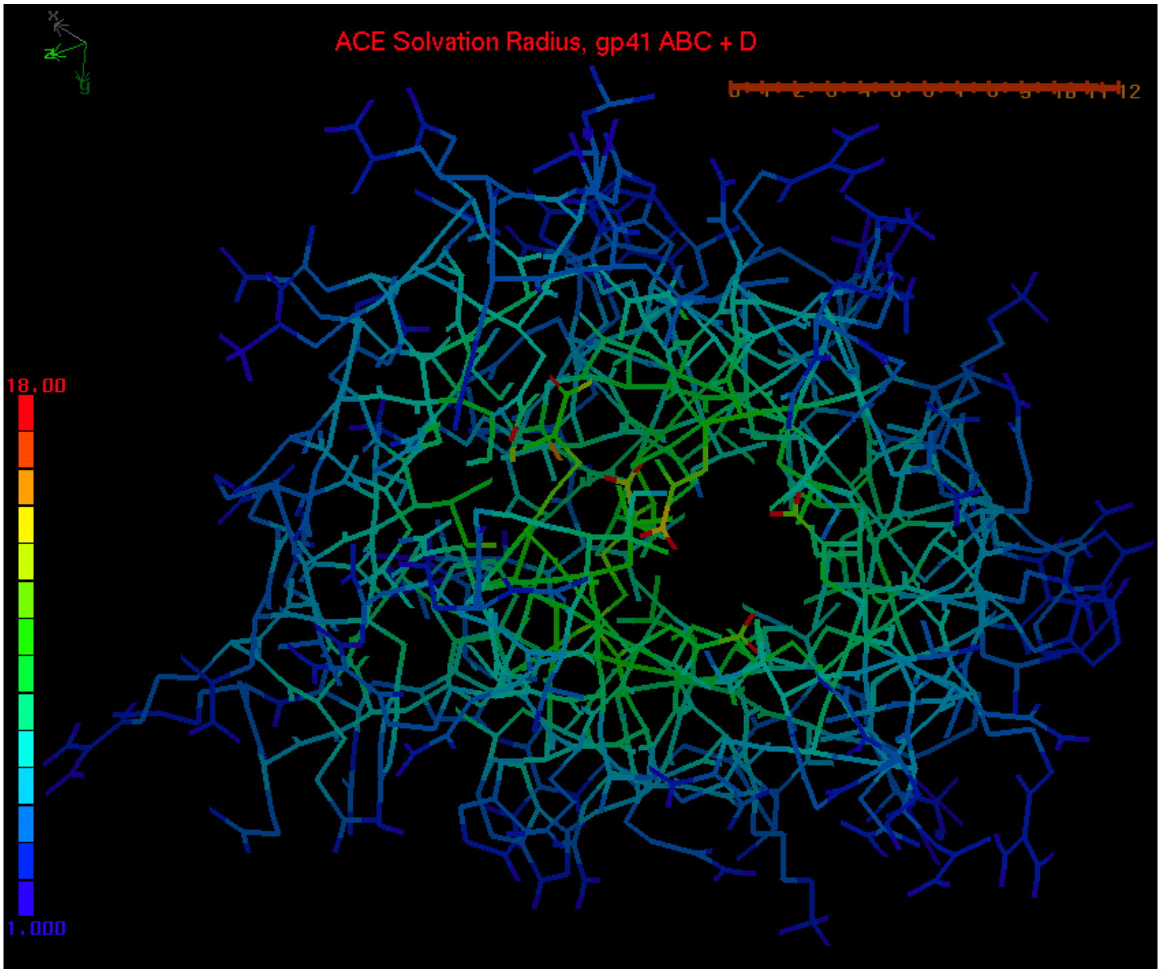

Figure 4-10: Bound complex of ABC:D chains of gp41, color-coded by ACE atomic solvation radius. The color-code key is on the left; a 12-Å long ruler is at the top right. These solvation radii are from ACE as in reference [7] and CHARMM version 27a2, without our improvements. Several red atoms, with too-high solvation radii, are a large source of error for the ACE method. (Figure made with QUANTA from Molecular Simulations, Inc. (San Diego, CA).)

by FDPB. For the bound cyanovirin homodimer, for example, $b_0 = 18.7$ Å is much greater than the largest actual solvation radius of a charged atom, 8.9 Å determined by FDPB. Our method, described in the next section, determines an adaptive maximum solvation radius $b_a = 11.0$ Å.

Adaptive Maximum Solvation Radius

We implemented a procedure, which can run automatically within CHARMM's modified ACE routines, that determines a maximum reasonable solvation radius, b_a, based on the sorted list of uncorrected atomic solvation energies in the particular system under consideration, $\Delta G_i^{\text{self}}, i \in [1, N_{\text{atoms}}]$. The algorithm skips over the negative and infinite uncorrected solvation radii b_i^{raw}, then considers the b_i^{raw} one at a time, starting with the largest one, b_1^{raw}. The adaptive maximum solvation radius b_a is chosen based on the b_i^{raw} for which the scaled rate of decrease

$$B_i \equiv N_{\text{atoms}}^{2/3} \frac{b_i^{\text{raw}} - b_{i+dn}^{\text{raw}}}{dn}, \tag{4.44}$$

where $dn = 4$, drops below a cutoff $B_{\text{cut}} = 20$, and stays below B_{cut} for a number of atoms equal to $1/f_{\text{pass}} = 2\%$ of the total number of atoms, and at least $r_{\text{min}} = 4$ atoms. Using $N_{\text{atoms}}^{2/3}$ like this ensures that the procedure works for arbitrary system size, since the largest true (FDPB) solvation radius scales as $N_{\text{atoms}}^{1/3}$.

Repeating that in formulae, we set

$$b_a = s_2 b_{n_{\text{passed}}}^{\text{raw}}, \tag{4.45}$$

where n_{passed} is the lowest integer i such that

$$B_i < B_{\text{cut}} \text{ for all } i \in [n_{\text{passed}}, n_{\text{passed}} + r], \tag{4.46}$$

where

$$r = \min(r_{\min}, N_{\text{atoms}}/f_{\text{pass}}). \tag{4.47}$$

Having found b_a, our procedure uses a flat cutoff at b_a, as pictured in Figure 4-6, so this is a maximum value for all corrected solvation radii. The parameters $s_1 = 0.99$, $s_2 = 1.01$ leave room for a small spline for uncorrected solvation radii $b_i^{\text{raw}} \in [b_a/s_2, b_a/s_1] = [b^{\text{raw}}_{n_{\text{passed}}}, b^{\text{raw}}_{n_{\text{passed}}} s_2/s_1]$. The spline ensures the continuous derivatives needed for continuous forces in MD applications. The spline affects b_i^{raw} just above $b^{\text{raw}}_{n_{\text{passed}}}$, so the atom that passed (atom number n_{passed}) has the highest solvation radius which is not corrected. We used the values $s_1 = 0.99, s_2 = 1.01, dn = 4, B_{\text{cut}} = 20.0, r_{\text{pass}} = 4, f_{\text{pass}} = 50.0$. The corrected solvation radii b_i are given by:

$$b_i = \begin{cases} -\frac{\tau q^2}{2\Delta G_i^{\text{self}}} & \text{if } \Delta G_i^{\text{self}} \leq -\frac{\tau q^2}{2b_a} s_2 \\ \text{the connecting spline,} & \text{if } -\frac{\tau q^2}{2b_a} s_2 < \Delta G_i^{\text{self}} < -\frac{\tau q^2}{2b_a} s_1 \\ b_a & \text{if } \Delta G_i^{\text{self}} \geq -\frac{\tau q^2}{2b_a} s_1 \end{cases} \tag{4.48}$$

Or, in terms of the uncorrected solvation radius b_i^{raw},

$$b_i = \begin{cases} b_i^{\text{raw}} & \text{if } b_i^{\text{raw}} \leq b_a/s_2 \\ \text{the connecting spline,} & \text{if } b_a/s_2 < b_i^{\text{raw}} < b_a/s_1 \\ b_a & \text{if } b_i^{\text{raw}} \geq b_a/s_1 \end{cases} \tag{4.49}$$

Pseudo-symmetry Detection

The algorithm also detects pseudo-symmetry using the sorted list of uncorrected solvation radii b^{raw}. In a nearly-symmetric homodimer like cyanovirin, for example, almost every atom gets the same solvation radius as its symmetry twin, so the list will consist mostly of pairs of values. If the system has m-fold pseudo-symmetry, the parameter dn above should obey $dn \geq m$ so that B_i vs. i is relatively smooth, rather than dropping to zero periodically because there are sets of m atoms with the same solvation radii.

Here is how our pseudo-symmetry detection algorithm works: If the b_i^{raw} have m

values in a row that are nearly the same ($N_{\text{atoms}}^{2/3}\frac{b_i^{\text{raw}}-b_{i+m}^{\text{raw}}}{m} < B_{\text{symm}} \equiv 0.5$), then m votes are cast for m-fold pseudo-symmetry. The number of atoms that cast votes is $n_{\text{passed}} + r$. If any one m value gets at least 40% of the votes, there may be m-fold pseudo-symmetry, so if $dn < m$, then the determination of b_a is started over again with dn set equal to m.

All of this was done in order to make the procedure fully automatic, so that solvation radii will be limited to reasonable values for any system to which our modified CHARMM ACE code is applied.

4.3.3 Salt Treatment

The approximate salt treatment of Srinivasan *et al.* [47] described in the previous section, was implemented as an option in the CHARMM ACE code. For evaluating this salt treatment, we use $\epsilon_i = 4$, with and without 0.145 M ionic strength, rather than the $\epsilon_i = 1$ used for the rest of our results.

4.3.4 FDPB Procedure

The structure of each bound complex is prepared for FDPB calculation by rotating it so that it fits in the smallest possible cube. To calculate the solvation and interactions of each group or atom, a series of 4 focussing steps were used, in which the entire bound complex's width was 23%, then 92%, then 184%, and finally 368% of the width of the area represented on the FDPB grid. At 23% fill, the boundary conditions on the potential at the edges of the grid are those in solvent fairly far from the protein complex. At 92% fill, the grid has a slightly larger extent than the protein complex. At 184% and 368% fill, only a portion of the protein complex is represented on the grid. For all magnifications except the first, the boundary conditions for the potential at the edges of the grid are taken from the potential map determined by the previous magnification. Interactions between the group at the center of the grid, and any other atom, are determined using the potential, at that atom's position (from the highest-magnification grid on which it fits), created by charging only the group at the center of the grid. These 4 magnification steps for each group were repeated 10 times at a series of small translations, to average out any errors caused by the particular placement of the grid lines with respect to the charges.

The number of grid points to use, 97, was chosen to ensure that the results are converged with respect to the number of grid points used. This was verified by requiring convergence of the group binding contribution terms which converged most slowly (changed the most), when calculated without overfocussing. Using 97 grid points at the final 368% fill yields a final grid spacing of 0.16 Å/grid for the cyanovirin homodimer, and 0.14 Å/grid for the gp41 ABC:D complex.

Convergence of the free energy terms with respect to number of iterations of the FDPB potential relaxation was confirmed by verifying that a group pair interaction got the same value when calculated in either of two ways: (1) the free energy of group 2's charges in the potential created by group 1's charges, or (2) the free energy of group 1's charges in the potential created by group 2's charges.

For each group (side chain, amino, or carbonyl), the FDPB calculations just described (10 translations, 4 magnifications) are done twice: once with that group charged in the dielectric boundary of the bound complex, and once with that group charged in the dielectric boundary of its unbound protein.

All FDPB calculations were repeated for each binding system, using $\epsilon_i = 4$ and $\epsilon_s = 80$, with and without 0.145 M ionic strength, to evaluate the treatment of salt effects.

4.3.5 Timing

For cyanovirin dimer binding, with 262 groups with charged atoms, the 262 × 2 FDPB calculations needed for a binding free energy component analysis by group take about 100 minutes each on a 400 MHz Pentium II processor, for a total of 36.4 processor-days. By comparison, to do a binding free energy component analysis by group or by atom takes ACE less than 1 processor-minute!

The routine which we added to CHARMM to calculate b_a adds an insignificant amount of time to an ACE free energy calculation. We also found it helpful to make unrelated improvements to the CHARMM ACE code to reduce memory usage by 2/3 at no cost in speed. We also added an optional flag which reduces memory usage by an additional 2/3, for a total reduction of 8/9, if forces are not required.

4.4 Results and Discussion

4.4.1 Adaptive Maximum Solvation Radius Results

Figure 4-11 shows ACE vs. FDPB atomic solvation radii for all charged atoms in the cyanovirin homodimer. The ACE solvation radii were calculated with a flat cutoff at $b_a = 11.0$ Å, which affects 3 atoms, Gln41 HE1 on the A subunit (appearing on the $y = b_a$ dotted line, to the left, in Figure 4-11), and Gln50 HE21 on both the A and B subunits (appearing on the $y = b_a$ dotted line, to the right, on top of each other). Table 4.2 shows the FDPB solvation radii for these 3 atoms, and what their ACE solvation radii are limited to using a tangential cutoff at b_0, a flat cutoff at b_0, or a flat cutoff at b_a.

Table 4.2: Atomic Solvation Radii (in Å) for 3 Atoms of Cyanovirin Homodimer. These are the atoms with which ACE has the most trouble. For this system, $b_0 = 18.66585$ and $b_a = 11.01153$. The effect of 3 different cutoffs on atomic solvation radii are shown. These are the only atoms affected by any of the cutoffs, but the choice of cutoff still has a big effect on the overall error of ACE, because atoms with very large solvation radii have incorrectly strong interactions with all other atoms.

atom	FDPB	ACE			
		uncorrected	tangent at b_0	flat at b_0	flat at b_a
Gln50 $H_{\epsilon21}$	5.68868	-27.88325 †	52.27476	18.66585	11.01153
Gln50' $H_{\epsilon21}$	5.68649	-31.73560 †	44.89072	18.66585	11.01153
Glu41 $H_{\epsilon1}$	3.93435	-81.09312 †	43.17975	18.66585	11.01153

† ACE gets nonphysical positive values for the atomic solvation energies, and so a nonphysical negative value for the solvation radii.

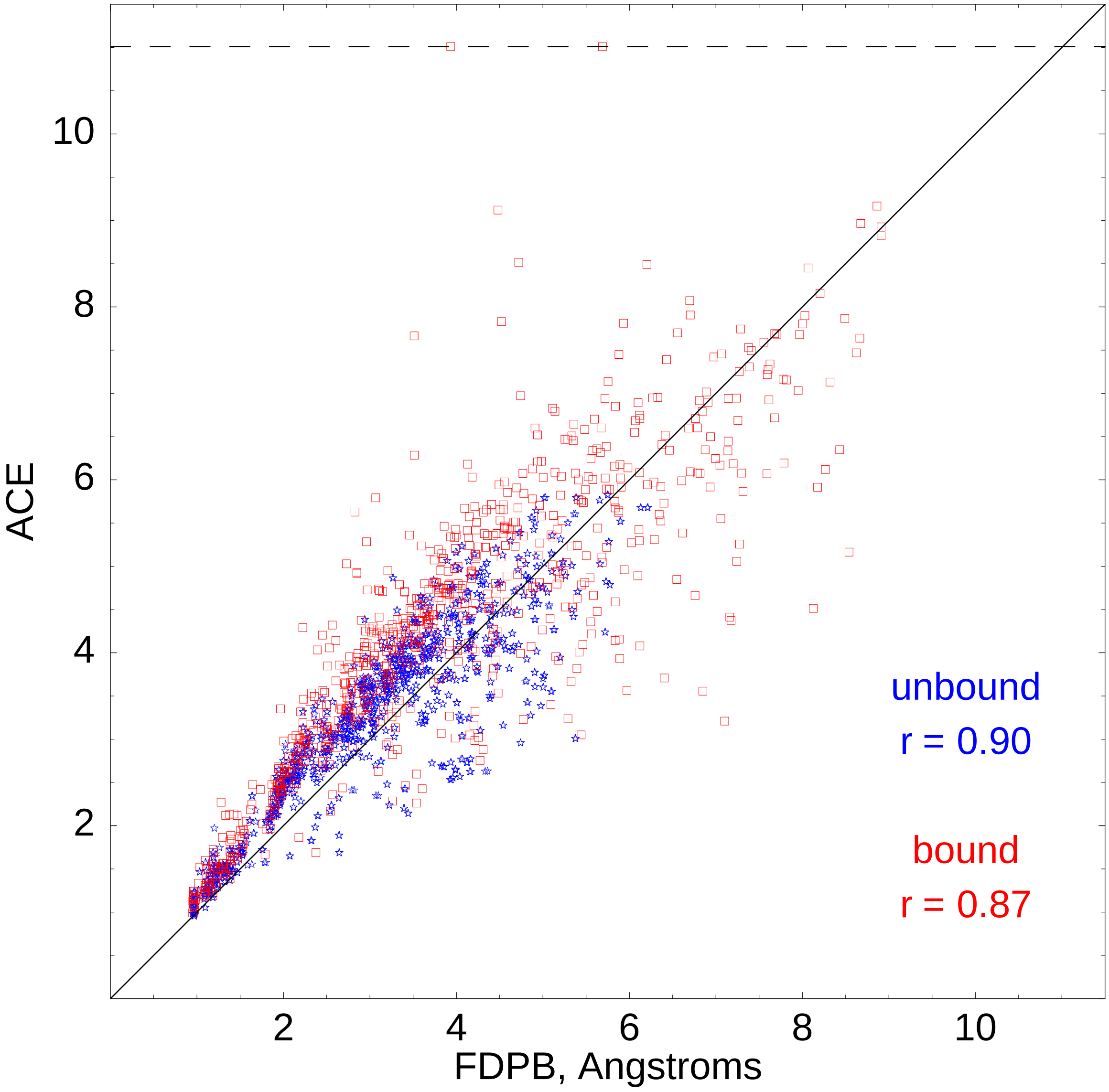

Figure 4-11: Atomic Solvation Radii for all charged atoms of the cyanovirin homodimer (plotted as red squares), and for one rigidly unbound subunit (plotted as blue stars). The ACE radii were subject to a flat cutoff at b_a. This cutoff only affects three atoms, which appear as 2 points on the dotted line at $y = b_a = 11.01153$.

Figures 4-12 and 4-13 illustrate this point with histograms, for 16 protein systems, of all atomic solvation radii b_i calculated with a flat cutoff at b_0. If a tangential cutoff at b_0 were used instead, the atoms in the right-most bin would have even larger solvation radii. The thin vertical line on each plot shows the location of the adaptive maximum solvation radius b_a found by our procedure for each system.

Table 4.3 shows the reduction of errors in terms of the binding free energy by using more stringent cutoffs on atomic solvation radii. Reference [7] and CHARMM version 27a2 use a tangential cutoff at b_0 . Using a flat cutoff (with a small connecting spline) at b_0, so that b_0 is the maximum allowed solvation radius, dramatically reduces the root mean squared deviation (rmsd) errors of the solvation radii and the group interaction binding terms, and reduces the errors somewhat on all terms. Using a flat cutoff at our adaptive maximum solvation radius b_a further reduces the rmsd error of the solvation radii, and slightly reduces the errors of all the energy terms.

4.4.2 Salt Treatment Results

The Srinivasan *et al.* [47] salt treatment does well with their recommended value of the scaling parameter $k = 0.73$. It predicts the salt effect on group binding contribution terms with correlation coefficient $\mathrm{r} = 0.95$ (coefficient of determination $\mathrm{r}^2 = 0.91$), as shown in Figure 4-14. The size of this salt effect is near zero for most groups, but ranges up to about ±0.4 kcal mol^{-1} (with $\epsilon_i = 4$) for some charged groups. The contribution is almost all from "direct" interaction terms (i.e. for group pairs for which the two groups are on opposite binding partners).

4.4.3 Pseudo-symmetry Detection Results

Our algorithm to automatically detect pseudo-symmetry using the ACE atomic solvation energies correctly determines the pseudo-symmetry for all test cases (14 proteins and protein complexes without pseudo-symmetry, 2 homodimers, 1 homotrimer, and 1

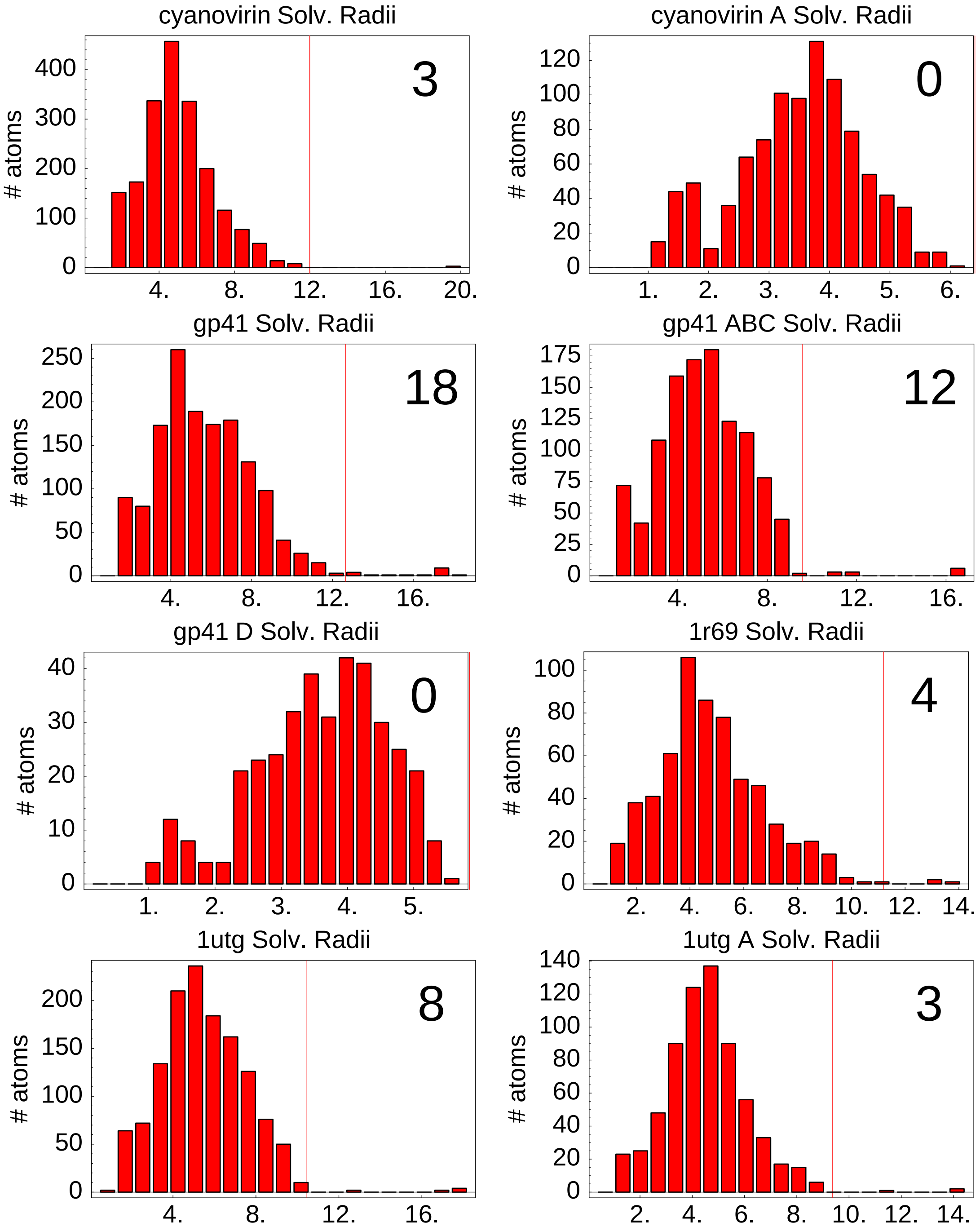

Figure 4-12: Histograms of ACE Atomic Solvation Radii for 8 Test Systems. A vertical line is shown at the adaptive maximum solvation radius b_a determined by our method. The number of atoms affected by a flat cutoff at b_a is shown in large type at the top right of each plot. Both charged and uncharged atoms are included.

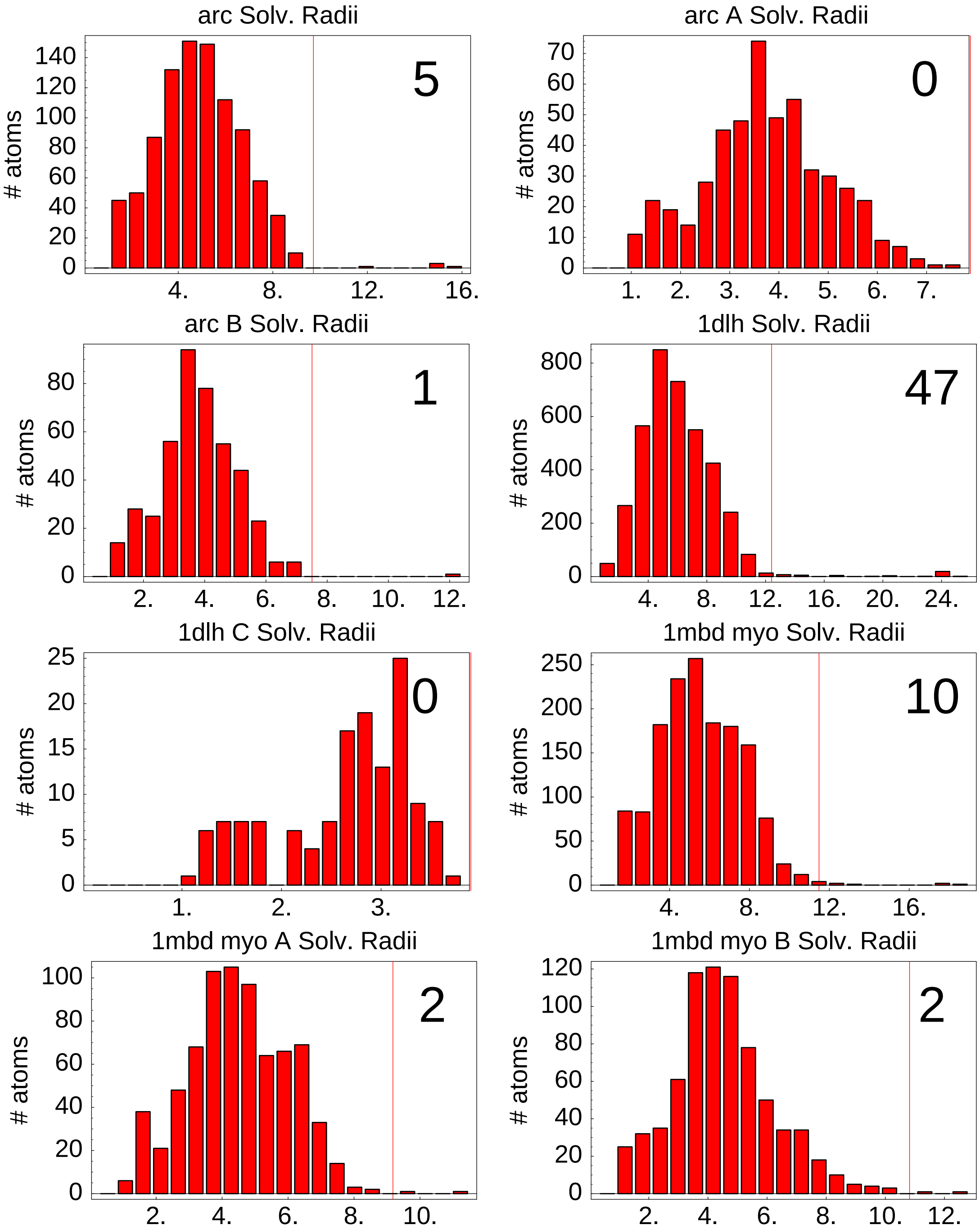

Figure 4-13: Histograms of ACE Atomic Solvation Radii for 8 More Test Systems. A vertical line is shown at the adaptive maximum solvation radius b_a determined by our method. The number of atoms affected by a flat cutoff at b_a is shown in large type at the top right of each plot. Both charged and uncharged atoms are included.

Table 4.3: Errors in binding free energy components for cyanovirin dimer binding. Error terms are for ACE, with the parameters given in reference [7] and with the given cutoff for the atomic solvation radii, vs. FDPB. Both ACE and FDPB used $\epsilon_{\mathrm{i}} = 1$, $\epsilon_{\mathrm{s}} = 80$, no salt.

	rmsd errors of:	
cutoff	solv. radii (Å) b	charged atom pair int. (kcal mol^{-1}) $\Delta G^{\mathrm{bind\ int}}$
tangent at b_0	2.070	0.074
flat at b_0	1.127	0.062
flat at b_a	0.991	0.061

	rmsd errors of:			
	group pair binding terms: (kcal mol^{-1})			
cutoff	$\Delta G^{\mathrm{bind\ int}}$	$\Delta G^{\mathrm{bind\ solv}}$	$\Delta G^{\mathrm{bind\ cont}}$	$\Delta G^{\mathrm{bind\ mut}}$
tangent at b_0	1.70	0.91	1.16	1.18
flat at b_0	0.93	0.89	0.91	1.14
flat at b_a	0.85	0.89	0.92	1.12

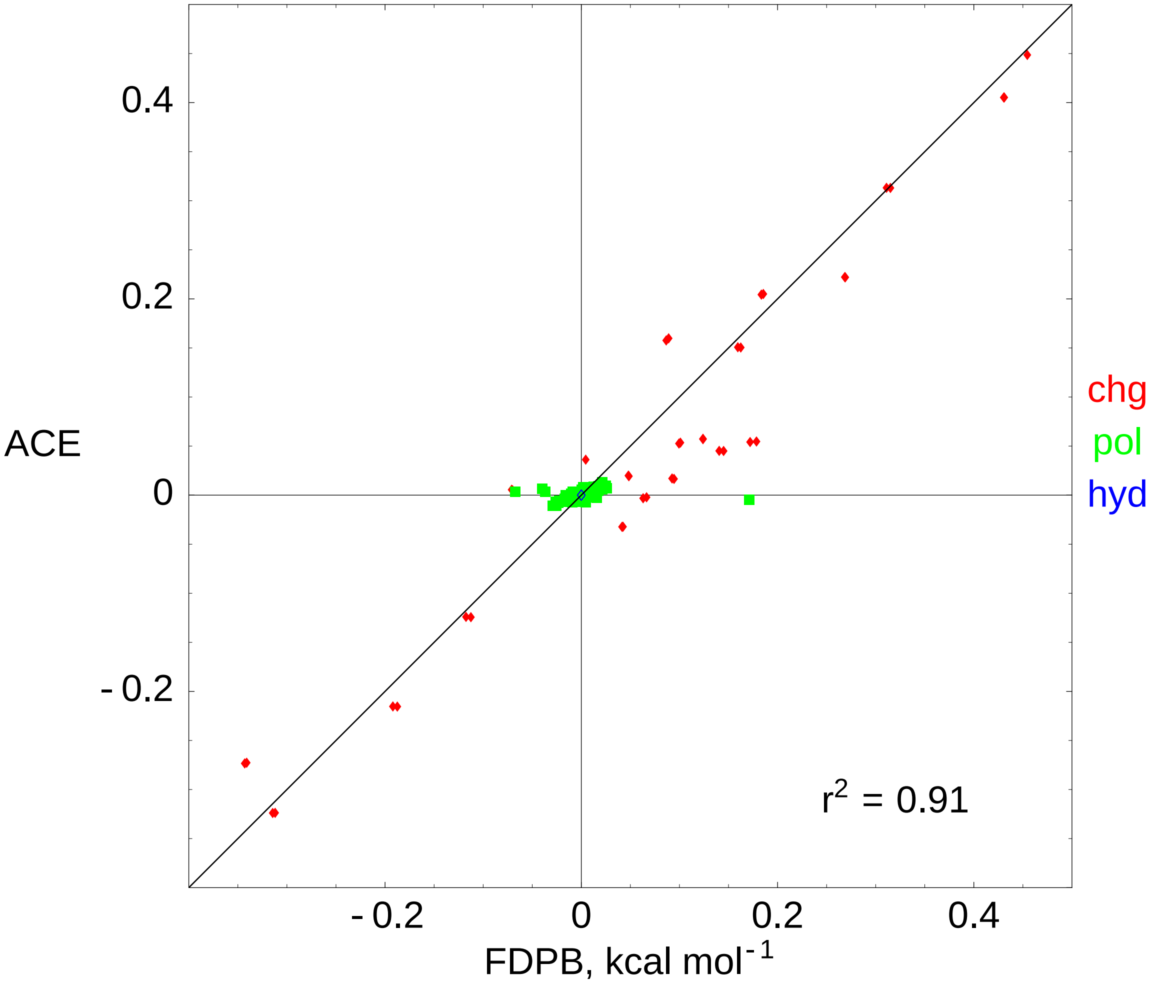

Figure 4-14: Approximate Salt Treatment: The effect of 0.145 M ionic concentration on the binding contribution terms of groups (side chain, amino, carbonyl) of the cyanovirin swapped-domain dimer. The ACE treatment used a flat cutoff at b_a. Both ACE and FDPB used dielectric constants $\epsilon_{\rm i} = 4$, $\epsilon_{\rm s} = 80$. Charged groups are shown as red diamonds; polar groups are shown as green boxes; hydrophobic groups are shown as blue diamonds (and they are all at the origin, of course).

homosexamer), with no fine-tuning of parameters needed. Details are shown in Table 4.4.

Table 4.4: Details of pseudo-symmetry detection algorithm. Atoms at the high end of the sorted list of uncorrected solvation radii vote for m-fold pseudo-symmetry if m solvation radii in a row are similar.

	atoms	votes	votes	votes	votes	cert-	symmetry m	
system, segments	voting	$m = 2$	$m = 3$	$m = 4$	$m = 6$	ainty	predict	actual
1R69	14	2				(0.14)		
1UTG	30	22		4		0.73	2	2[a]
1UTG, A	14	2				(0.14)		
Arc	19	2				(0.11)		
Arc, A	8					(0.00)		
Arc, B	8					(0.00)		
1DLH	104	22				(0.21)		
1DLH, AB	88	10	6			(0.11)		
1DLH, C	3					(0.00)		
1MBD	36	4				(0.11)		
1MBD, A	15	2				(0.13)		
1MBD, B	15					(0.00)		
cyano	37	24	3		6	0.65	2	2[b]
cyano, A	18	4				(0.22)		
gp41	37	4				(0.11)		
gp41, ABC	28	2	15			0.54	3	3[a]
gp41, D	6	2				(0.33)		
gp41, ABCABC	56	12		4	30	0.54	6	6[c]

[a] exact symmetry, from the crystal structure.

[b] nearly exact symmetry, except for 2 hydrogen atoms.

[c] 2 separated copies of the exactly symmetric trimer.

4.5 Conclusion

The ACE analytical approximation of the electrostatic solvation free energy is much faster than finite-difference solution of the Poisson-Boltzmann equation; for example, a component analysis of the binding free energy takes less than one minute of processor time for ACE, versus 5 weeks for converged FDPB results. Atomic solvation radii predicted by ACE correlate very well (r $\geq$ 0.87) with FDPB results. However, we found that the ACE method's largest errors were typically caused by a few atoms whose atomic solvation radii are predicted to be incorrectly high. We found that limiting the solvation radii of these atoms can greatly reduce errors in the group terms of the binding energy. Using a flat cutoff rather than a tangential cutoff on the solvation radii reduces the rmsd error for group interaction terms of the binding free energy from 1.70 to 0.93 kcal mol^{-1} for the cyanovirin homodimer binding system. To further limit the solvation radii, we developed an automatic procedure that adaptively determines a maximum reasonable atomic solvation radius for each molecular system, based on the distribution of uncorrected solvation radii. We implemented the procedure as an option in the ACE routine of the molecular modeling package CHARMM. For the cyanovirin homodimer binding system, this further reduced errors in the solvation radii and the group binding interaction terms. To ensure that our procedure performs properly for symmetric or pseudo-symmetric molecular systems, a fully automated pseudo-symmetry detection algorithm was also developed and implemented. The algorithm correctly detected the presence or absence of pseudo-symmetry for all 18 test cases. We also incorporated the approximate treatment of salt effects proposed by Srinivasan *et al.* into ACE. With no changes or optimization, it predicts the salt effect on group binding contribution terms with correlation r = 0.95. We have made the ACE method more accurate and more flexible. These improvements, combined with the method's speed, are especially valuable when large numbers of electrostatic energy evaluations must be performed, such as for minimization, multiple site titration, or ligand design.

Chapter 5

Optimization of Analytical Continuum Electrostatics Parameters

5.1 Introduction

Development of analytical approximate electrostatic methods such as the Analytical Continuum Electrostatics (ACE) method of Schaefer and Karplus [6] has focused to date on predicting solvation free energies — first for small molecules, then for large molecules. Overall solvation free energy, an aggregate quantity that is important in electrostatics, is the most difficult part of the electrostatic free energy to calculate; however, validation of the methods based only on solvation free energies does not justify their use for prediction of other energy terms [65]. Binding affinity and folding stability, for example, are two free energy terms which should be of paramount interest in ligand design. Furthermore, in ligand design the actual value of the binding free energy is not needed; only the binding free energy differences $\Delta\Delta G^{\text{bind}}$ between possible ligands are of interest. For the folding free energy as well, if an experimental folding free energy is known for one related ligand, then only folding free energy differences $\Delta\Delta G^{\text{fold}}$ are needed. Closely related to the

$\Delta\Delta G^{\text{bind}}$ is the breakdown of ΔG^{bind} into its components, such as the contribution of each group of atoms to the binding.

The ACE method has a number of free parameters; Schaefer *et al.* optimized them to minimize fluctuations in the solute density [55], and validated their use with solvation energies [6]. In this chapter, we develop and apply a novel method to optimize the ACE parameters to minimize any of a wide range of ACE vs. finite difference Poisson-Boltzmann (FDPB) error functions. The terms of interest to us are the effect on ΔG^{bind} of the interaction and solvation terms of all atoms or groups of atoms. By minimizing the error on this large set of small terms, our optimization should have an incentive to get all the details right; whereas optimizing only the total ΔG^{bind}, for example, could allow compensating errors in the components that add up to the correct total.

The parameters are optimized using data from a "training" system. (By "system", we mean a pair of biomolecular binding partners.) Then the parameters can be applied to a "testing" system. In all applications of optimization, there is a danger of "over-training". Over-training means that the performance on the training system is improved, but not in a way that improves the performance on other systems. In our case, for example, much of the error of ACE vs. FDPB comes from a handful of atoms which ACE incorrectly determines to be very desolvated. Suppose that a certain hydrogen atom has this problem, and that there happens to be a nitrogen atom of type N near it. When we optimize the ACE parameters, the algorithm could drop the effective atom volume parameter $V(\text{N})$ nearly to zero only so that this one particular N atom can reduce that one hydrogen atom's desolvation, even though the change to $V(\text{N})$ degrades the accuracy of many other atoms' energy terms by small amounts. What we want instead is for the parameter changes to reduce errors on the training system in ways that would also reduce errors on any other protein.

We found no general method for finding a parameter set that will do well on all protein binding systems. When a parameter set which was optimized to reduce errors on one of our 4 protein-protein binding systems was used on the other systems, the error sometimes improved and sometimes worsened. With a more closely related set of systems

— 8 zinc finger variants — parameters optimized on one system usually performed better on the other systems as well. We conclude that optimization of the ACE parameters can significantly reduce errors when ACE is applied to large families of related systems.

5.2 Methods

5.2.1 Biomolecular Complexes Used for Testing

Cyanovirin-N Domain-Swapped Homodimer

See Section 4.3.1 of Chapter 4.

gp41 Protease-resistant Core

See Section 4.3.1 of Chapter 4.

Arc Repressor Dimer

We modeled the rigid binding of the arc repressor dimer from bacteriophage P22. Coordinates for the A and B chains were takes from the 2.6 Å-resolution crystal structure (PDB entry 1PAR) [58], which includes residues 6 to 53 on the A chain, and 6 to 46 on the B chain. Using the PARAM19 extended-atom parameter set, polar hydrogens were built onto the crystal structure using the HBUILD facility [43] in the CHARMM package [11].

***Bacillus subtilis* Chorismate Mutase Homotrimer**

The isomerase chorismate mutase from the organism *Bacillus subtilis* is a homotrimer with 127 residues per monomer. We modeled the last step of trimer formation by rigidly binding one unit to the other two. The structure is shown in Figure 5-1. Coordinates of the crystallographic subunit consisting of chains A, B, and C were taken from from the 1.9 Å-resolution crystal structure (PDB entry 2CHS) of the trimer [66]. The crystal structure is missing density for the first one and the last 8 to 12 residues of each chain, so the N-termini of the available chains were acetylated, and the C-termini were blocked with N-methyl groups. Using the PARAM19 extended-atom parameter set, the acetyl groups on the N-termini were minimized. Polar hydrogens were built onto the crystal structure using the HBUILD facility [43] in the CHARMM package [11]. Based on visual inspection of all histidine, aspartate, and glutamate residues with an eye to making favorable hydrogen bonding patterns, we chose to doubly protonate all 9 histidines, but leave all aspartates and glutamates deprotonated.

Zinc Finger Protein/DNA Complexes

A family of proteins called zinc fingers bind selectively to DNA sequences. We use 8 such crystal structures, shown in Figure 5-2: variants of the Zif268 protein complexed with DNA. The 8 differ from one another at three DNA base pairs, and at several protein residues near those base pairs. For each structure, the protein is a single chain that consists of three homologous regions ("fingers") [67]. Each of the three finger regions contains a zinc ion coordinated by two histidine and two cysteine residues.

Crystal structures were obtained from PDB entries 1AAY, 1A1F, 1A1G, 1A1H, 1A1I, 1A1J, 1A1K, 1A1L [67, 68]. (Do not confuse the zinc finger structure 1A1K with the gp41 structure 1AIK.) For several atoms which had two possible positions in the crystal structure, we selected the NH2B position in Arg13, OG1B in Thr21, NE2B in Gln34, and CD1B in Leu48. (The selection was arbitrary since neither the X-ray data itself nor a visual inspection show a preference for one position over the other.) Using the PARAM19 extended-atom parameter set, polar hydrogens were built onto the crystal structure using the HBUILD facility [43] in the CHARMM package [11]. Each zinc ion is coordinated by two histidine and two cysteine residues. Based on a visual inspection, both histidines coordinating every zinc ion were chosen to be singly protonated at $N_{\delta 1}$. All of the zinc-coordinating cysteine residues must be deprotonated. A restrained electrostatic potential (RESP) fit to an ab initio quantum mechanical electron distribution was used to

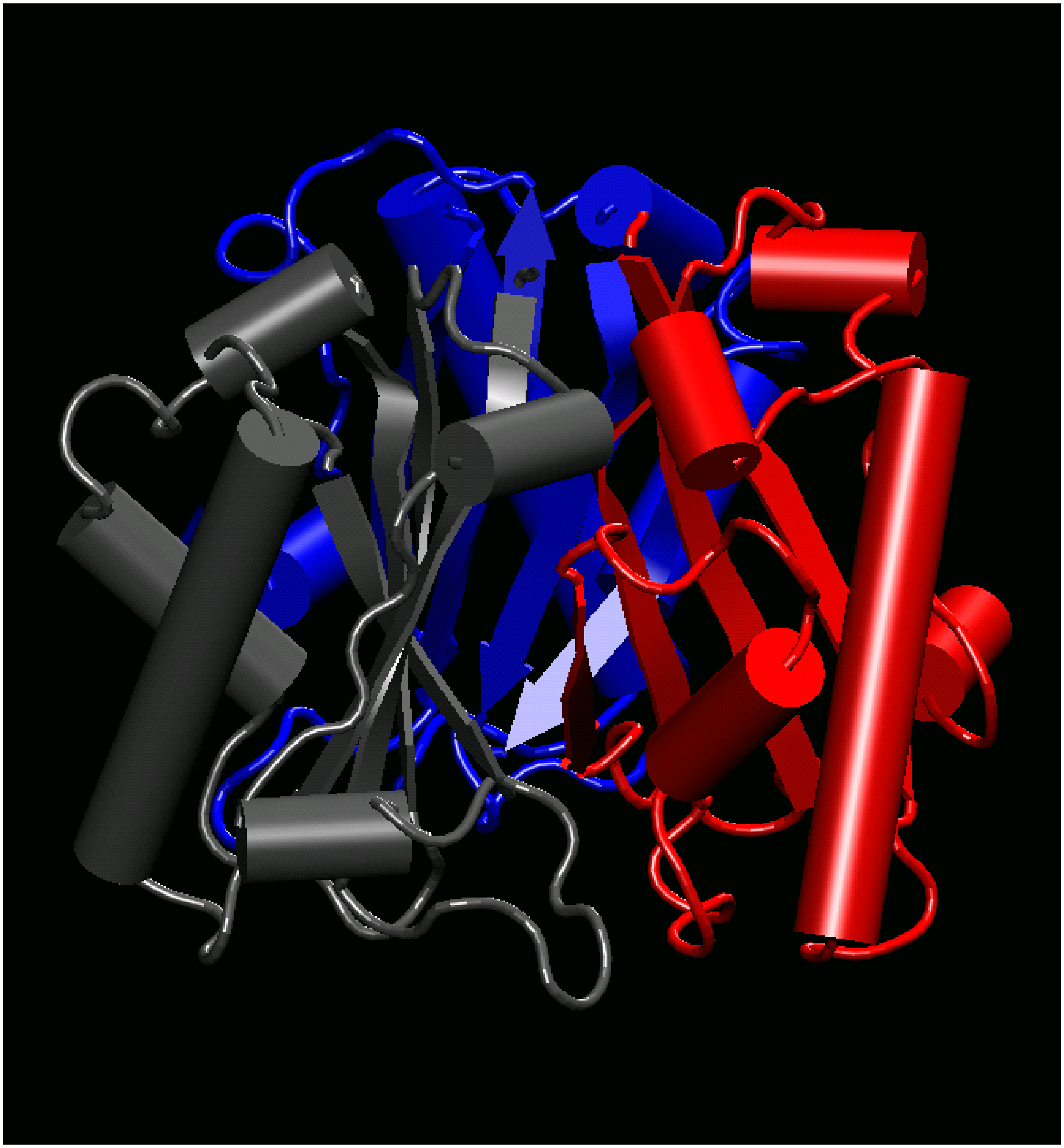

Figure 5-1: Chorismate mutase homotrimer from *Bacillus subtilis*, shown as a secordary structure cartoon, and colored by chain.

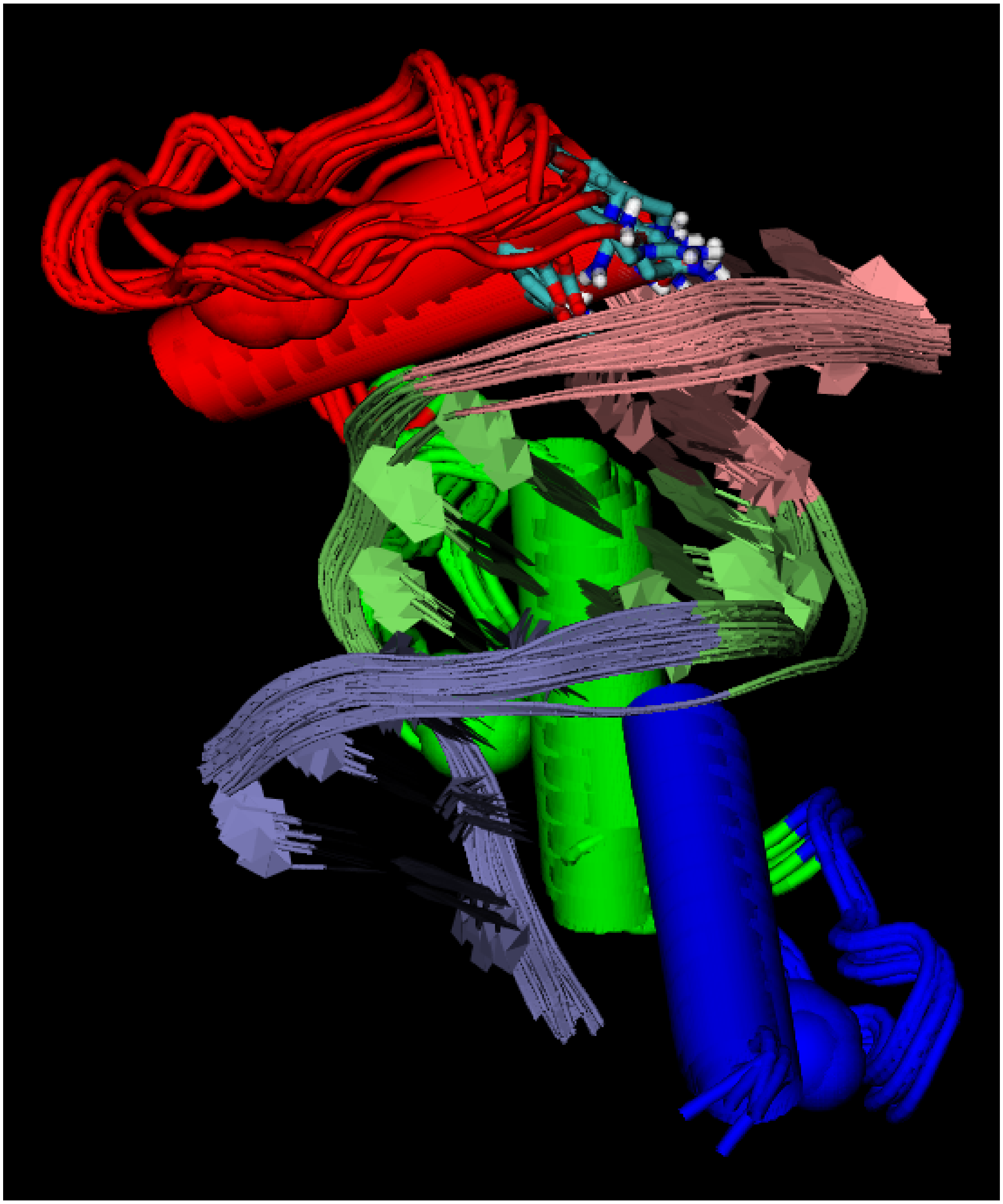

Figure 5-2: Aligned structures of the 8 zinc finger protein/DNA complexes 1AAY, 1A1F, 1A1G, 1A1H, 1A1I, 1A1J, 1A1K, 1A1L, each a variants of Zif268 complexed with DNA. The DNA is shown in cartoon form going from the top right to the bottom left of the picture. The protein is also shown in cartoon form; and the zinc ions are shown as van der Waals spheres. The protein and zinc are colored red, green, and blue for finger regions 1, 2, and 3; the DNA regions are correspondingly colored pink, light green, and steel blue. The three DNA-contacting protein residues which differ among the 8 structures are shown as licorice near the top of the picture.

determine partial atomic charge parameters for a zinc ion coordinated by two histidines and two cysteines [69]. For these atoms, the RESP partial atomic charges are used rather than the PARAM19 partial atomic charges: for the 2 cysteines, +0.154 on C_β and -0.858 on S_γ; for the 2 histidines, +0.144 on C_β, -0.055 on C_γ, -0.213 on $N_{\delta1}$, +0.319 on $H_{\delta1}$, +0.095 on $C_{\delta2}$, -0.415 on $N_{\epsilon2}$, and +0.249 on $C_{\epsilon1}$; and +1.160 on the zinc ion. All aspartates and glutamates in the structures were visually inspected with an eye to satisfying hydrogen bonding patterns, and none of them needed to be protonated. For the only non-zinc-coordinating histidine, His47, we chose the singly protonated state with the hydrogen on the $N_{\epsilon2}$ atom in all 8 structures. In one structure, 1A1G, we chose to flip the His47 ring over. Finally, the structures were all rotated to fit the 1AAY structure in the smallest possible cube, in order to make FDPB calculations more efficient.

In Section 4.2.5 of Chapter 4, we described how we conceptually divide proteins into amino, carbonyl, and side chain groups in order to analyze the components of the binding free energy. DNA segments are similarly divided into base, ribose, and phosphate groups. In order that each of these three groups have a neutral net charge, the C_1' atom is conceptually divided so that +0.2 of its charge is in the ribose group, and +0.06 of its charge is in the base group.

In Table 4.1 of Chapter 4, we gave the ACE effective atom volumes for protein atom types that we use as our "default parameters" and as the starting point of the parameter optimization runs. We extended these volumes to the PARAM19 DNA atom types by choosing the most similar protein atom type where possible: H2, HO, and HZ were set to 0.310 (all volumes are in Å^3) like H; O2, OSZ, and OST were set to 16.404 like OC; P was set to 15.196 like S; C2 was set to 35.356 like CH2E; C3 was set to 40.947 like CH3E; CA, CB, CS, and CZ were set to 12.403 like C; CE and CF were set to 18.583 like CR1E; CH was set to 12.257 like CH1E; N2 was set to 18.677 like NH2; NA was set to 1.708 like NH1; NB and NC were set to 16.611 like NR; NS was set to 0.0 like N; OH was set to 21.427 like OH1; OZ was set to 14.375 like O; and ZN (zinc ion) was set to 9.854 based on its 1.33 Å Born radius.

Table 5.1 shows the similarity of the 3 "finger" regions of the zinc finger structure with PDB code 1AAY. Table 5.2 shows a sequence alignment of the target DNA and finger 1 of the protein for the 8 zinc finger structures with PDB codes 1AAY, 1A1F, 1A1G, 1A1H, 1A1I, 1A1J, 1A1K, 1A1L. They differ at three DNA positions, at three residues in finger 1 of the protein, and some of them lack a final arginine residue at the end of finger 3.

Table 5.1: Alignment of the three "finger" regions of the zinc finger structure with PDB code 1AAY. The three regions comprise one protein segment, its sequence is simply the three lines given here read in order. Blank spaces are included only to align the sequences. The secondary structure is noted below: two arrows for short beta strands and "===" for alpha helices.

RPYACPVESCDRRFSRSDELTRHIRIHTGQK	finger 1
PFQCRI CMRNFSRSDHLTTHIRTHTGEK	finger 2
PFACDI CGRKFARSDERKRHTKIHLR	finger 3
--> --> ============	secondary structure

Table 5.2: Alignment of the DNA and the "finger 1" protein region for zinc finger structures 1AAY, 1A1F, 1A1G, 1A1H, 1A1I, 1A1J, 1A1K, 1A1L. The differing locations are marked with stars below. The DNA sequence is divided into amino acid codons.

Name	DNA	Protein Finger 1
1AAY:	A GCG TGG GCG T	RPYACPVESCDRRFSRSDELTRHIRIHTGQK
1A1F:	A GCG TGG GAC C	RPYACPVESCDRRFSDSSNLTRHIRIHTGQK
1A1G:	A GCG TGG GCG T	RPYACPVESCDRRFSDSSNLTRHIRIHTGQK
1A1H:	A GCG TGG GCA C	RPYACPVESCDRRFSQSGSLTRHIRIHTGQK
1A1I:	A GCG TGG GCA C	RPYACPVESCDRRFSRSADLTRHIRIHTGQK
1A1J:	A GCG TGG GCG T	RPYACPVESCDRRFSRSADLTRHIRIHTGQK
1A1K:	A GCG TGG GAC C	RPYACPVESCDRRFSRSADLTRHIRIHTGQK
1A1L:	A GCG TGG GCA C	RPYACPVESCDRRFSRSDELTRHIRIHTGQK
	* * *	* * *

5.2.2 Minimizing Group Binding Term Errors

The choice of an error function to minimize depends on the uses to which one plans to put the ACE method after optimizing its parameters. Here, we choose to optimize the parameters for use in ligand design or component analysis of biomolecular binding systems. We seek to minimize errors in all groups' terms of the binding free energy, including the group-pair interactions (I, for short), the total interactions of each group (TI), group solvation terms (S), group contribution terms (C), and group mutation terms (M). The errors over many groups could be combined by taking the root mean squared deviation (rmsd). We also use the "rm4d," root mean fourth-power deviation, in some of our energy functions. The rm4d error between two data sets $\{x_i\}$ and $\{x_i'\}$, which have a one-to-one correspondence between the N members of each, is defined as:

$$\mathrm{rm4d}(\{x_i\},\{x_i'\}) \equiv \left(\frac{1}{N}\sum_{i=1}^{N}|x_i - x_i'|^4\right)^{1/4} \tag{5.1}$$

We used the rm4d function because it assigns more importance to a few large errors than to widespread small errors. For contrast, however, we also used the rmsd and the "rm1d" functions for some of our results. The rm1d could also be called the mean absolute deviation:

$$\mathrm{rm1d}(\{x_i\},\{x_i'\}) \equiv \frac{1}{N}\sum_{i=1}^{N}|x_i - x_i'| \tag{5.2}$$

We took the rmsd, rm1d, and rm4d of the S, TI, C, and M over all groups except those with no charged atoms. We took the rmsd, rm1d, and rm4d of the I term over all group pairs whose FDPB interaction I has magnitude $\geq$ 0.005 kcal mol^{-1}. Finally, in order to optimize rm4d(I), rm4d(S), rm4d(TI), rm4d(C), and rm4d(M) simultaneously, we can let our optimization function be their product or their weighted sum. We use the names "rm1d5", "rmsd5", and "rm4d5" for these products of the 5 error functions:

$$\mathrm{rm1d5} \equiv \mathrm{rm1d(I)} \cdot \mathrm{rm1d(S)} \cdot \mathrm{rm1d(TI)} \cdot \mathrm{rm1d(C)} \cdot \mathrm{rm1d(M)} \tag{5.3}$$

$$\mathrm{rmsd5} \equiv \mathrm{rmsd(I)} \cdot \mathrm{rmsd(S)} \cdot \mathrm{rmsd(TI)} \cdot \mathrm{rmsd(C)} \cdot \mathrm{rmsd(M)} \tag{5.4}$$

$$\mathrm{rm4d5} \equiv \mathrm{rm4d(I)} \cdot \mathrm{rm4d(S)} \cdot \mathrm{rm4d(TI)} \cdot \mathrm{rm4d(C)} \cdot \mathrm{rm4d(M)} \tag{5.5}$$

Using a product rather than a weighted sum has the advantage of fairly weighting each of the terms. For example, when we optimize the ACE parameters to minimize the rm4d5 error function, either a 1% improvement in rm4d(I) or a 1% improvement in rm4d(M) would have the same effect on rm4d5 at any point in the optimization, regardless of how much rm4d(I) and rm4d(M) had decreased in the course of the optimization.

5.2.3 Finite Difference Poisson-Boltzmann Procedure

The FDPB procedure to calculate group terms of the binding free energy is described in Section 4.3.4 of Chapter 4. Using 97 grid points at a final magnification of 368% yields a final grid spacing of 0.16 Å/grid for the cyanovirin homodimer, 0.14 Å/grid for the gp41 ABC:D complex, 0.10 Å/grid for the Arc dimer, and 0.15 Å/grid for the chorismate mutase trimer. For the 8 zinc finger structures, we used 65 grid points and focussing steps with 23%, 92%, and 368% magnification, yielding a final grid spacing of 0.20 Å/grid. Unless otherwise noted, FDPB and ACE results use zero ionic concentration and $\epsilon_{\mathrm{i}} = 4$.

5.2.4 Optimization Procedure

We implemented a procedure to search a many-dimensional parameter space for a parameter set that minimizes any desired error function. A simple example of an error function would be the root mean squared deviation (rmsd), ACE vs. FDPB, for the contribution terms of all atomic groups to the binding free energy. The parameter optimization can be thought of as a search for the minimum height of a surface, where the height is the objective function, and rather than 2 horizontal axes, there are many more "horizontal" axes, one per parameter.

Minimization of a nonlinear function with a many-dimensional parameter space is an inherently difficult problem. We implemented the downhill simplex method of Nelder and Mead [70] with simulated annealing added as by Press *et al.* in *Numerical Recipes* [71]. The N-dimensional parameter space (in this work, $N = 18$ usually) is searched by a simplex, which is a collection of $N + 1$ points. The method attempts to move one simplex point at a time, always the highest point (i.e., the one with the worst function value), by reflecting it through the centroid of the other N points. If this proposed new location is lower than the original highest point, the move is accepted. If the new location is lower than any of the other simplex points, then an "expansion" move, twice as far in the same direction, is attempted. If the original reflection gives a value that is still higher than all the other simplex points, a "contraction" is attempted, trying a point midway between the highest point and the centroid of the other simplex points. If this contraction still gives a point higher than all the others, the entire simplex is shrunk by one half, in the direction of its lowest (best) point. This procedure is vastly better than striking out in random directions, because the simplex changes its shape in a way that makes downhill moves more likely. The expansions stretch the simplex out in the downhill direction, and the contractions shrink the simplex into narrow valleys. A few such moves make the simplex longer along the uphill-downhill direction. This, in turn, makes it even more likely that proposed moves will be downhill.

Simulated annealing is added to the downhill simplex method in the usual way, using a temperature parameter T (with the same units as the function to be minimized) which is started high and lowered according to an annealing schedule. A variant of the Metropolis algorithm [72] is used: whenever comparing two points' heights, $T(-\log(random))$ is added to the current point's height, and $T(-\log(random))$ is subtracted from the proposed new point's height, where "*random*" is a new random number between 0 and 1 each time it is used. Thus there is an incentive to try new points if they are higher than the current highest point by no more than about T. As the temperature is lowered, the range of heights of the simplex points decreases, and so the size of the simplex in the parameter space also decreases.

The ACE parameters which we optimize are α and the effective atom volumes V. For protein systems, there are 17 unique atom types and therefore 17 effective atom volume parameters to minimize. For the zinc finger structures, which include DNA and zinc ions, there are 41 unique atom types. We restrict the parameters to non-negative values.

5.3 Results and Discussion

5.3.1 Low-Temperature Simulated Annealing Chosen

Ideally, simulated annealing should start at a temperature higher than the largest terrain features, and cool very slowly. We found that it would take unpractical amounts of time to do such a rigorous search for the global optimum. In practice, low-temperature optimizations that take hours find lower minima than high-temperature optimizations that take many days. So there are better local optima in the vicinity of the default parameters than can be found in the wider parameter space by many searches in many days of computer time.

All annealing schedules that we used lower the temperature T by the factor f_{cool} every 100 steps. (Every attempted move of the simplex is one step.) Figures 5-3 and 5-4 show the progress of multiple optimization runs by plotting the value found for the objective function $\mathrm{rm4d3} \equiv \mathrm{rm4d(S)} \cdot \mathrm{rm4d(M)} \cdot \mathrm{rm4d(I)}$ (initially $\mathrm{rm4d3}_{\mathrm{init}}$) as the temperature drops throughout the run. The runs which started at an initial temperature of $T_{\mathrm{init}} = 1000 \times \mathrm{rm4d3}_{\mathrm{init}}$ used $f_{\mathrm{cool}} = 0.95$ and took 3.3 days for 24×10^3 steps. The runs which started at $T_{\mathrm{init}} = 100 \times \mathrm{rm4d3}_{\mathrm{init}}$ used f_{cool} values of 0.98, 0.95, and 0.9 (shown as red, light purple, and dark purple in Figure 5-3) and took 6.7, 2.8, and 1.4 days for 49×10^3, 20×10^3, and 10×10^3 steps, respectively. The runs which started at lower temperatures used $f_{\mathrm{cool}} = 0.9$ and took less than a day.

We conclude that running several optimizations with starting temperatures 1/100th the size of the initial value of the optimization function can quickly reach better optima than either higher-temperature or zero-temperature optimizations. So although rigorous global optimization is not computationally feasible, local optima exist that are most easily found by a rather quick annealing schedule. Therefore one can afford to do multiple optimization runs differing only in their random seeds in order to find a set of local optima. This, in turn, allows one to statistically characterize the effect of chance on the optimization.

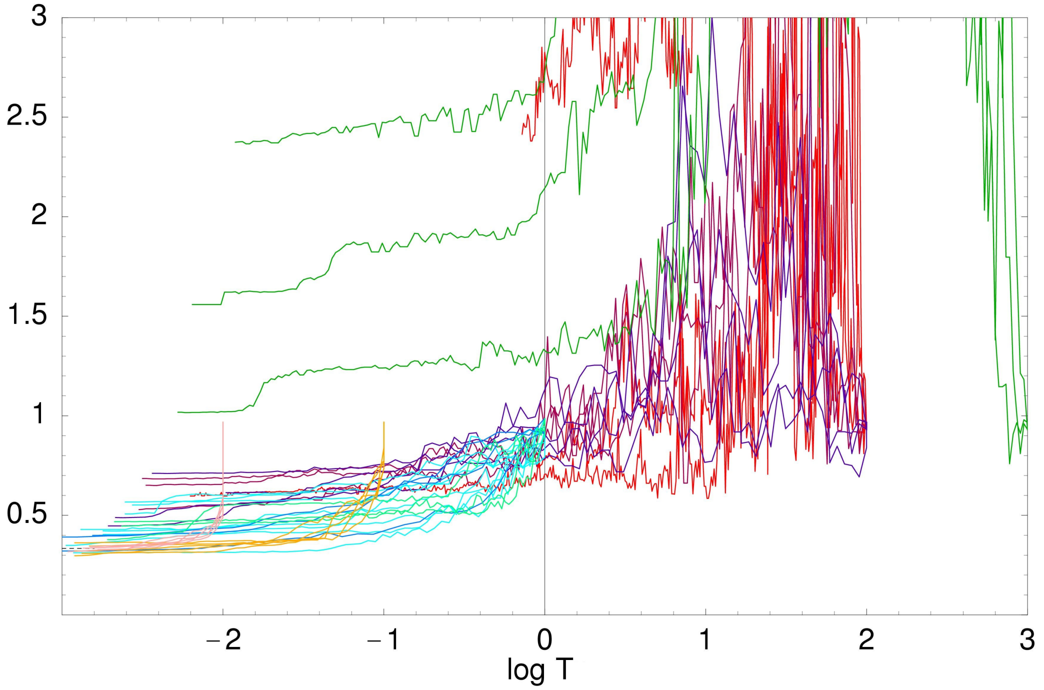

Figure 5-3: History of the error function through the course of each optimization. Each line follows one optimization run from right to left as the temperature drops according to the simulated annealing schedule. The vertical axis is the lowest value of the objective function rm4d3 $\equiv$ rm4d(S) $\cdot$ rm4d(M) $\cdot$ rm4d(I) at any of the simplex's points. The temperature is shown on a logarithmic scale, relative to the objective function value before optimization, $\text{rm4d3}_{\text{init}}$. Optimization runs were begun at initial temperatures T_{init} of 0.01, 0.1, 1, 100, and 1000 times $\text{rm4d3}_{\text{init}}$. The short dotted line at the lower left shows the value of rm4d3 found by a zero-temperature optimization. The higher-temperature runs search more widely in parameter space, but none of them ever find values of the objective function as low as those found by all of the low-temperature runs.

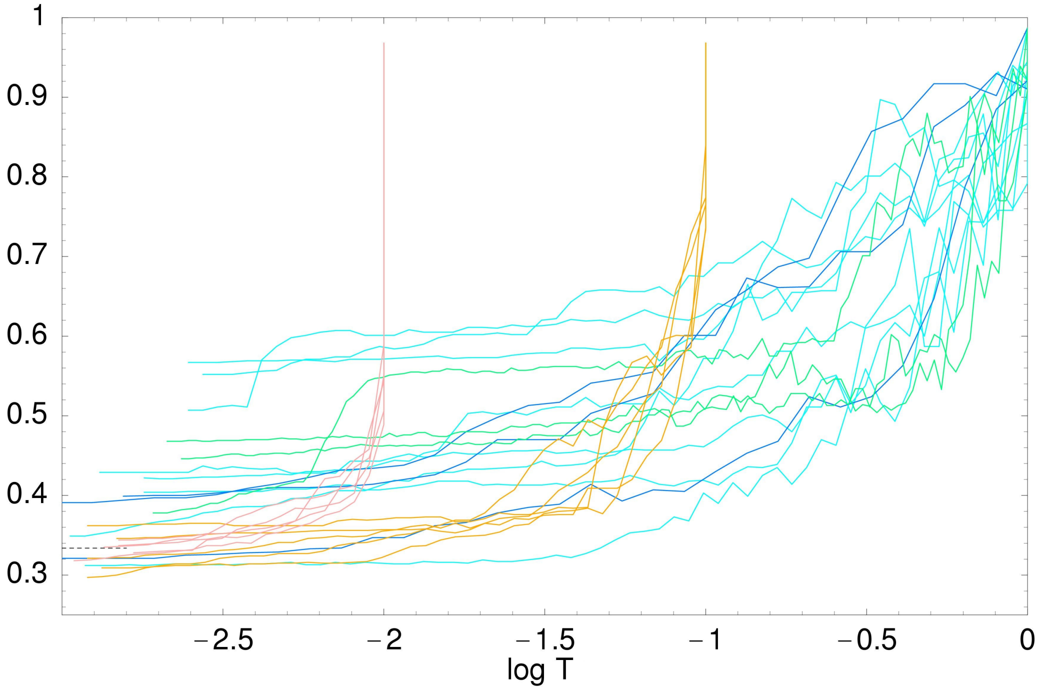

Figure 5-4: Histories of the error function through the course of an optimization. For the optimization runs begun at T_{init} of 0.01, 0.1, and 1 times $\text{rm4d3}_{\text{init}}$, this is a detail view of the bottom left corner of Figure 5-3; see the explanation in that figure's caption. The short dotted line at the lower left shows the value of rm4d5 found by a zero-temperature optimization. About half of the low-temperature optimization runs find lower values than the zero-temperature run.

We decided on the following simulated annealing schedule for all of the following results: The initial temperature T_{init} was set to 0.01 of the value of the error function at the starting position (the "default" parameters). Every 100 attempted moves, the temperature was lowered by a factor of $f_{\text{cool}} = 0.9$. One such ACE parameter optimization takes about 7 processor-hours on a 400MHz Pentium II processor.

5.3.2 Parameters Optimized for Accurate Atomic Solvation Energies

We optimized parameters to minimize the atomic solvation radius error rmsd(R) for the bound cyanovirin dimer. The results are shown in Table 5.3: rmsd(R) and the atomic solvation energy error rmsd(E) are reduced, but the group binding free energy terms rmsd(I), rmsd(TI), and rmsd(M) are increased. The error function rmsd5 increases by +52%, and rm4d5 increases by +396%.

We also optimized parameters to minimize rmsd(E). The results are shown in Table 5.4: again, rmsd(R) and rmsd(E) are reduced, but the group binding free energy terms rmsd(I), rmsd(TI), rmsd(M) are increased. The error function rmsd5 increases by +55%, and rm4d5 increases by +233%.

We conclude that optimizing parameters for more accurate atomic solvation energies does not give more accurate binding free energy terms; in fact, it makes them dramatically less accurate. So, to use ACE to predict binding terms, it is not sufficient to use parameters chosen to give accurate atomic solvation energies. Recall that ACE can be conceptually divided into the prediction of atomic solvation radii, and the use of these solvation radii by the Generalized Born equation to predict screened atomic interactions. The finding that more accurate solvation radii do not necessarily lead to more accurate interaction terms of the binding free energy demonstrates a weakness of the Generalized Born equation. Nevertheless, in order to improve the accuracy of the binding free energy terms, we optimize the ACE parameters. Such optimized parameters will yield ACE solvation radii which compensate for the inaccuracy of the Generalized Born equation in order to produce more accurate binding free energy terms. This is acceptable because the solvation radii are only an intermediate result used to calculate the binding free energy terms.

Table 5.3: Errors before and after optimizing ACE parameters to minimize the error function rmsd(R), the rmsd of the atomic solvation radii in the bound cyanovirin dimer, with $\epsilon_{\mathrm{i}} = 1$. The "with optimized parameters" data are averaged over 8 separate parameter sets optimized in the same way except for different random number seeds. The "1 standard deviation range" is over these 8 data points. The units are kcal mol^{-1} unless otherwise noted.

bound state error function:	rmsd(R)	rmsd(E)
w/ default params :	0.99 Å	2.93
w/ optimized params :	0.65 Å	2.06
fractional change:	-35%	-30%
1 standard dev. range:	±2%	±2%

binding error function:	rmsd(I)	rmsd(TI)	rmsd(S)	rmsd(C)	rmsd(M)
w/ default params:	0.06	0.85	0.89	0.92	1.12
w/ optimized params:	0.08	1.12	0.82	0.96	1.33
fractional change:	+30%	+31%	-7%	+4%	+18%
1 standard dev. range:	±9%	±12%	±6%	±7%	±7%

ACE, using the default parameters, systematically underestimates atomic solvation energies by about 15% for the cyanovirin dimer, as shown in Figure 5-5. We therefore added an optional new parameter, β, to scale all atomic solvation energies before they are used to calculate the interactions. Optimizing only the parameter β to minimize rmsd(E) for the bound cyanovirin dimer results in a value of $\beta = 1.2$ (with standard deviation 0.01 over 8 optimization runs). With the object of predicting group binding terms, however, optimizing β along with the other 18 parameters results in an optimized value of $\beta = 1.0$. Since this value means that the solvation energies are not scaled up, we conclude that the parameter β is superfluous for prediction of group binding free energy terms, and therefore its use is not warranted.

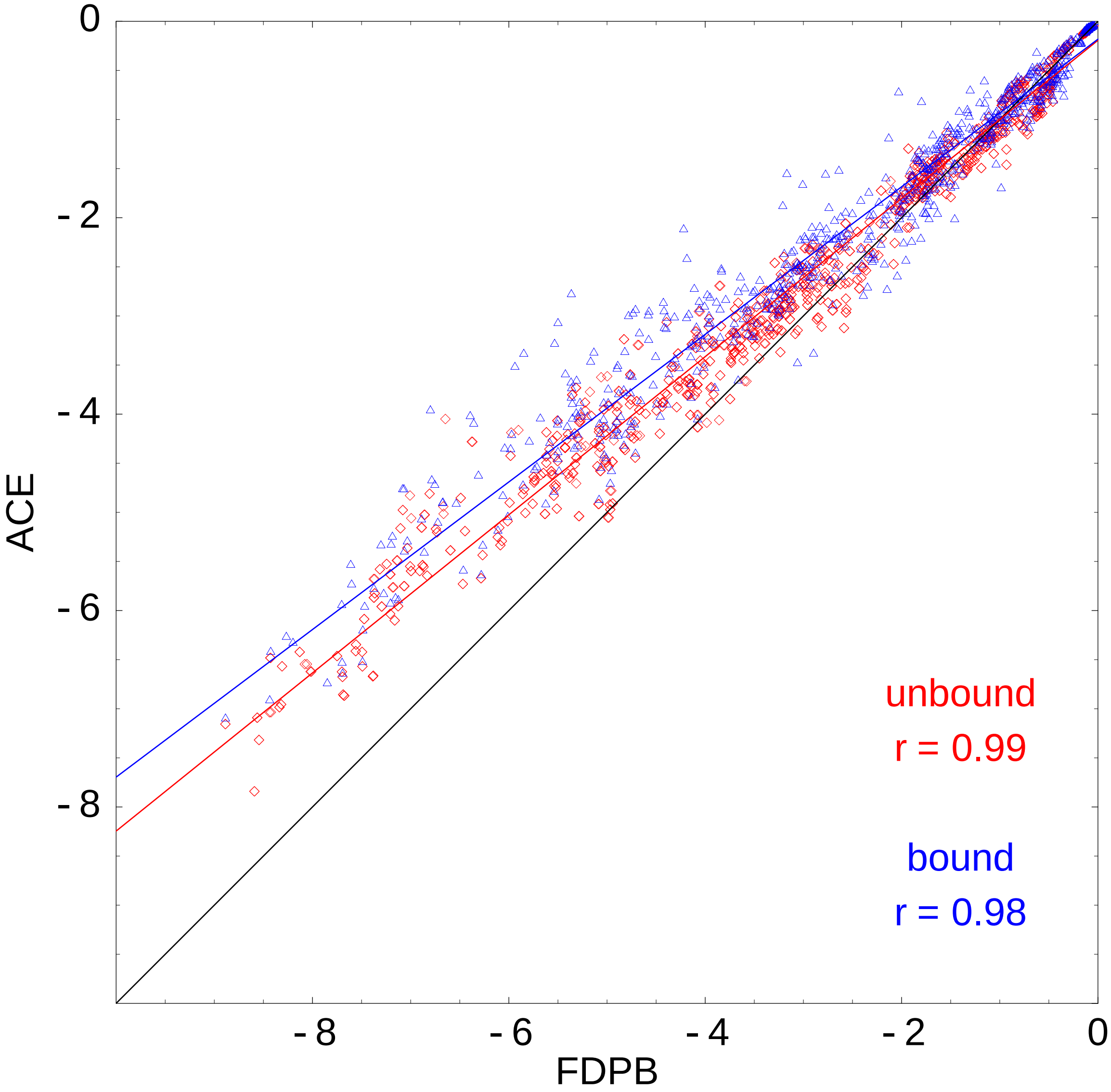

Figure 5-5: Atomic solvation energies (kcal mol^{-1}), ACE with default parameters vs. FDPB with $\epsilon_i = 1$, for all atoms in cyanovirin dimer (shown as blue triangles), and all atoms in the rigidly unbound cyanovirin monomers (shown as red diamonds). Least-square fit lines and correlation coefficients are shown for the bound and unbound data separately.

Table 5.4: Errors before and after optimizing ACE parameters to minimize the error function rmsd(E), the rmsd of the atomic solvation energies in the bound cyanovirin dimer, with $\epsilon_i = 1$. The "with optimized parameters" data are averaged over 8 separate parameter sets optimized in the same way except for different random number seeds. The "1 standard deviation range" is over these 8 data points. The units are kcal mol^{-1} unless otherwise noted.

bound state error function:	rmsd(R)	rmsd(E)
w/ default params :	0.99 Å	2.93
w/ optimized params :	0.88 Å	1.73
fractional change:	-12%	-41%
1 standard dev. range:	±7%	±3%

binding error function:	rmsd(I)	rmsd(TI)	rmsd(S)	rmsd(C)	rmsd(M)
w/ default params:	0.06	0.85	0.89	0.92	1.12
w/ optimized params:	0.08	1.14	0.75	0.86	1.25
fractional change:	+32%	+34%	-16%	-7%	+11%
1 standard dev. range:	±7%	±15%	±7%	±7%	±10%

5.3.3 Minimizing Only a Contribution or Mutation Term

Parameters optimized to minimize the group contribution or mutation binding free energy error terms rm4d(C) or rm4d(M) on the cyanovirin or gp41 binding systems do so by reducing the solvation error term rm4d(S) but increasing the interaction error terms rm4d(I) and rm4d(TI). The results for one of those 4 combinations, the minimization of rm4d(C) on the cyanovirin binding system, are shown in Table 5.5. In the following sections, we minimize the error functions rm4d5, rm1d5, and rmsd5, which we will show has the advantage of reducing both the solvation and interaction error terms rm4d(S), rm4d(I), and rm4d(TI).

5.3.4 Error Reduction for Training System

Optimizing the error function rm4d5 on the cyanovirin system achieves a reduction of -88%, as shown in Table 5.6. Recall from Equation 5.5 that rm4d5 is the product of 5 rm4d errors of group binding terms: group pair interaction I, group solvation S, group total interaction TI, group contribution C, group mutation M.

Optimizing function rm4d5 on the gp41 system acheives a similar reduction of 92% by reducing each of rm4d(I), rm4d(S), rm4d(TI), rm4d(C), and rm4d(M) by at least 22%.

5.3.5 Training and Testing Between Cyanovirin and Gp41 Systems

Optimizing the error function rm4d5 by training on gp41, then using the parameters on cyanovirin (with $\epsilon_i = 1$), achieves a reduction of -57% in rm4d5 versus default parameters on cyanovirin, as detailed in Table 5.7. Optimizing function rm4d5 by training on cyanovirin, then using the parameters on gp41, achieves a lesser reduction of -25% ±32% in rm4d5, mostly by reducing contribution and mutation term errors rm4d(C) and rm4d(M). This demonstrates that it is possible for a parameter set trained on one system to reduce errors on an unrelated system.

5.3.6 Training and Testing Among Four Unrelated Protein Systems

For the 4 protein-protein binding systems — cyanovirin, gp41, Arc, and chorismate mutase — we evaluated whether ACE parameters optimized on one such system will reduce errors when applied to another system. We used each of the 4 systems as a training system, doing multiple optimization runs, identical except for the random seeds, to get a family of parameter sets. Each parameter set was then used on the other 3 protein systems. Each such attempt to optimize a parameter set on a "training" system and transfer it to a different "testing" system could be judged a success if it results in a lower error than the default parameters applied to the testing system. In Figure 5-6, results for all 4×4 combinations of training and testing system are shown. The errors are all shown as a fraction relative to the error for the default parameter set applied to the

Table 5.5: Errors before and after optimizing ACE parameters to minimize the error function rm4d(C), the rm4d of the group contributions to the binding free energy of the cyanovirin dimer, with $\epsilon_i = 1$. The "with optimized parameters" data are averaged over 8 separate parameter sets optimized in the same way except for different random number seeds. The "1 standard deviation range" is over these 8 data points. All error functions have units of kcal mol^{-1}.

error function:	rm4d(I)	rm4d(TI)	rm4d(S)	rm4d(C)	rm4d(M)
w/ default params :	0.24	1.89	1.92	1.90	2.09
w/ optimized params :	0.39	2.33	1.49	1.03	1.65
fractional change:	+62%	+24%	-23%	-46%	-21%
1 standard dev. range:	±18%	±27%	±13%	±3%	±6%

Table 5.6: Reduction of errors by optimizing ACE parameters to minimize the error function rm4d5 $\equiv$ rm4d(I) $\cdot$ rm4d(S) $\cdot$ rm4d(TI) $\cdot$ rm4d(C) $\cdot$ rm4d(M) for cyanovirin dimer binding, with $\epsilon_i = 1$. The "with optimized parameters" data are averaged over 32 separate parameter sets optimized in the same way except for different random number seeds. The "1 standard deviation range" is over these 32 data points. All error functions have units of kcal mol^{-1} except for rm4d5 which has units (kcal mol^{-1})^5 .

error function:	rm4d(I)	rm4d(TI)	rm4d(S)	rm4d(C)	rm4d(M)	rm4d5
w/ default params :	0.24	1.89	1.92	1.90	2.09	
w/ optimized params :	0.23	1.15	1.04	1.12	1.38	
fractional change:	-6%	-39%	-46%	-41%	-34%	-88%
1 standard dev. range:	±4%	±2%	±1%	±2%	±3%	±2%

error function:	rmsd(I)	rmsd(TI)	rmsd(S)	rmsd(C)	rmsd(M)	
w/ default params :	0.06	0.85	0.89	0.92	1.12	
w/ optimized params :	0.06	0.64	0.58	0.63	0.81	
fractional change:	-2%	-25%	-35%	-32%	-28%	
1 standard dev. range:	±4%	±2%	±2%	±3%	±3%	

Table 5.7: Reduction of errors on cyanovirin binding by optimizing ACE parameters to minimize the error function rm4d5 for binding of the gp41 ABC:D system, and then using the parameters on the binding of the cyanovirin homodimer (with $\epsilon_i = 1$). The "with optimized parameters" data are averaged over 32 separate parameter sets optimized in the same way except for different random number seeds. The "1 standard deviation range" is over these 32 data points. All error functions have units of kcal mol^{-1} except for rm4d5 which has units (kcal mol^{-1})^5 .

error function:	rm4d(I)	rm4d(TI)	rm4d(S)	rm4d(C)	rm4d(M)	rm4d5
w/ default params :	0.24	1.89	1.92	1.90	2.09	
w/ optimized params :	0.26	1.34	1.44	1.58	1.89	
fractional change:	+6%	-29%	-24%	-16%	-9%	-57%
1 standard dev. range:	± 7%	± 5%	±4%	±3%	±2%	± 5%

error function:	rmsd(I)	rmsd(TI)	rmsd(S)	rmsd(C)	rmsd(M)	
w/ default params :	0.06	0.85	0.89	0.92	1.12	
w/ optimized params :	0.06	0.73	0.74	0.82	1.03	
fractional change:	-1%	-14%	-17%	-11%	-8%	
1 standard dev. range:	±2%	±4%	±3%	±3%	±2%	

testing system. Therefore, every point to the left of the vertical line could be considered a successful transfer of a parameter set from one system to another. About half of the 4 × 3 combinations of different training and testing systems are generally successful, but half of the transfers are not: between chorismate mutase and Arc, between chorismate mutase and cyanovirin, from gp41 to chorismate mutase, and from gp41 to Arc.

Table 5.8 summarizes the mean rm4d5 errors for the 4 × 4 combinations of training and testing system, as well as the errors for the default parameters used on each system.

The Arc system gets high errors using the default parameters or parameters optimized on gp41. Most of this error is due to a few groups whose energy terms are off by 3 to 13 kcal mol^{-1}. Specifically, by lowering each parameter one at a time, we found that the errors are much reduced by lowering the effective atom volumes $V(\text{HC})$ and $V(\text{NC2})$. These are the atom types of arginine's guanidinium moiety. The HC atom type is also in the NH_3^+ moiety of lysine. The groups with large solvation term errors are near HC and NC2 atoms which are not very solvated. The Arc system is rather unusual in that

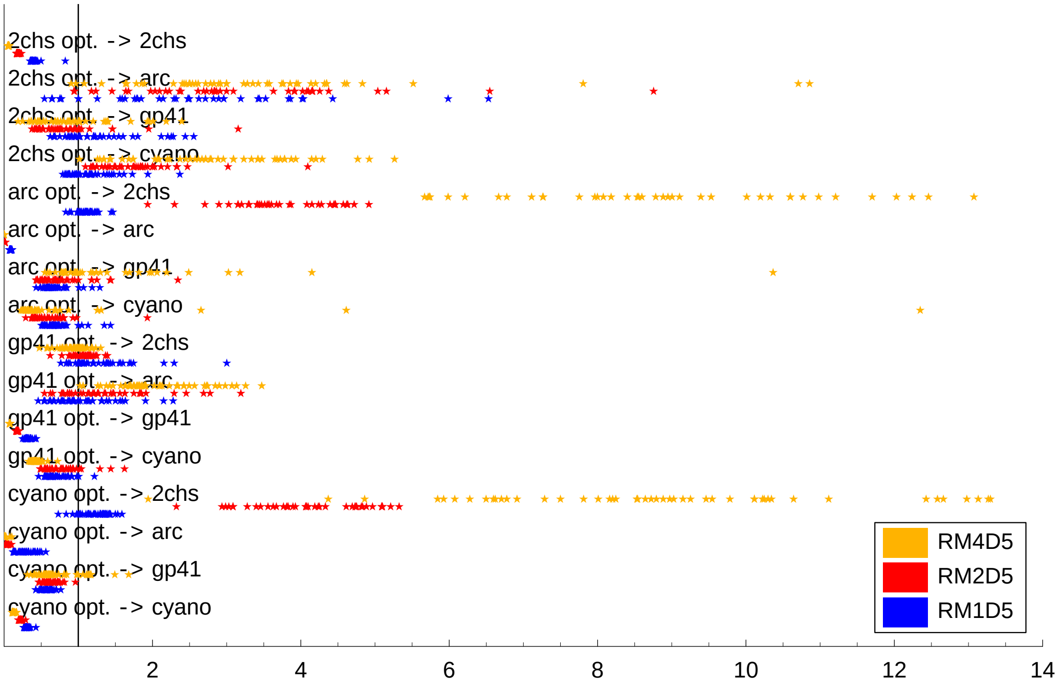

Figure 5-6: Error relative to the default parameter set, for each combination of training system and testing system among the 4 protein binding systems cyanovirin, gp41, Arc, and chorismate mutase (2chs). The 16 rows are for the 4 × 4 combinations of training and testing system. The optimizations and transfers were repeated for each of the error functions rm4d5, rm2d5, and rm1d5, which are shown in that order from top to bottom within each row.

Table 5.8: Error function rm4d5 for each combination of training and testing system. These are the actual rm4d5 values, in units of (kcal mol^{-1})^5, unlike the relative values shown in Figure 5-6.

	rm4d5 error for system:			
	cyanovirin	gp41	Arc	2chs
parameter set:				
default:	3.46	1.50	346.84	6.24
optimized on cyanovirin:	0.42±0.02	1.06±0.29	3.70±0.02	54.80±2.48
optimized on gp41:	1.49±0.06	0.11±0.00	719.77±0.59	5.88±0.17
optimized on Arc:	3.28±2.01	2.34±1.63	0.45±0.00	55.25±2.04
optimized on chor.mut.2chs:	9.78±1.06	1.37±0.53	1195.60±2.00	0.35±0.01

it has Arg and Lys side chains that are somewhat buried. In cyanovirin, gp41, and most proteins, Arg and Lys side chains are exposed to solvent on the surface of the protein. This is why it is possible for an optimization on gp41 to raise $V(\mathrm{HC})$ and $V(\mathrm{NC2})$ without causing errors, because all HC and NC2 atoms are far from other atoms. But when a parameter set optimized on gp41, which has large values of $V(\mathrm{HC})$ and $V(\mathrm{NC2})$, is used on Arc, the HC and NC2 atoms, and other atoms near them, are incorrectly treated as being very deeply buried, and so all of their interactions are made too strong.

5.3.7 Choosing From Among a Family of Optimizations

When we run a set of optimizations, identical except for their random seeds, the resulting collection of parameter sets has a spread of error function values, for both the training system and a different testing system. Using data from the four protein-protein binding systems, we applied various methods of predicting which members of a set of optimizations would perform best when transferred to other systems.

First, there was no significant correlation between the error function for the training and testing systems (among the 4 protein-protein binding systems), so the parameter sets which do best on the training system do not necessarily do better on the testing system.

Second, considering each parameter set as a point in 18-dimensional parameter-space, one might expect that a point near the center of the cluster would do better than average when transferred to another system. However, no significant correlation was found between the distance of a parameter-space point from the center of the cluster, and the error when that parameter set was transferred to another system.

Lastly, when a family of parameter sets are optimized on a training system “1” and then are applied to a second system “2”, some parameter sets do better than others at reducing errors. If the sets that do better on system 2 also do better on system 3, that would be evidence that those sets are more likely to do well on any other system, because they are not over-trained on system 1. In Figure 5-7, we find no such correlation of errors on system 2 with errors on system 3. Some of the individual plots show correlation, but in aggregate the correlation is not significant.

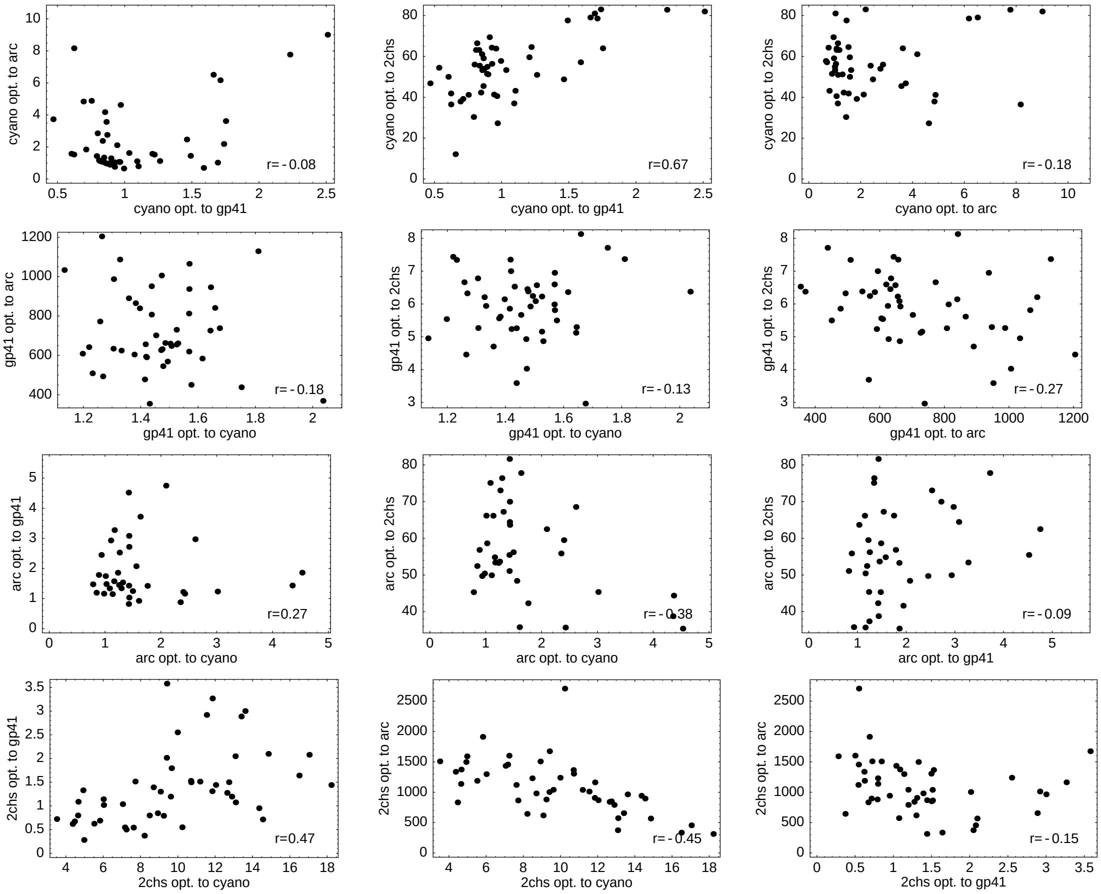

Figure 5-7: For parameter sets optimized on system 1, the error on system 2 vs. the error on system 3. All combinations of systems 1, 2, and 3 from among the our 4 protein-protein binding systems are shown. The error function in all cases is rm4d5. The lack of correlation means that we have not found any parameter sets that can be expected to do better on an arbitrary fourth system.

Ideally, we would like to find a parameter set which does reasonably well on all, or many, protein-protein binding systems. We did not find such a parameter set, nor did we find any promising methods for choosing such a parameter set. We concluded that we were expecting too wide of a range of applicability for the parameter sets, and in Section 5.3.8 we have more success transferring parameter sets between related systems.

5.3.8 Training and Testing Among Eight Zinc Finger Structures

Each of the 8 zinc finger structures is used as the training system for a family of optimization runs, differing only in their random seeds, which minimize the error function rm1d5. These runs generate a family of optimized parameter sets. Each optimized parameter set was applied to each of the 8 structures. The rm1d5 for every combination is shown in Figure 5-8. The average reduction of rm1d5 error for the training structure achieved by optimization is -59% (standard deviation 13%, range -32% to -79%). The average reduction of rm1d5 error achieved by optimizing on any one of the 8 structures and transferring the parameters to any other structure is -25% (standard deviation 31%, range -67% to +33%). For 6 of the 8 structures, 92% of the parameter sets optimized on another structure give lower rm1d5 error than the default parameters. Based on the data for these zinc finger structures, it is generally more advantageous to use a parameter set optimized on another structure rather than using the default parameters.

5.3.9 Optimized Parameter Values

Figure 5-9 shows the values of the ACE parameters, α and the 17 effective atom volumes V. For some parameters (α, V(CH2E), V(OH1), and V(OC)), distinctly different values are obtained by optimization on cyanovirin versus gp41. This is clear evidence that the parameters are being optimized not simply for proteins in general, but for some unique qualities of the training system. Another important observation is that, for these 4 parameters, optimization of either rm1d5 or rm4d5 results in the same range of values. This is also true for all parameters in general, but it is easier to see for these 4 parameters. Since the rm4d5 error function depends so strongly on the few largest error terms, one could reasonably suspect that the rm4d5 error function would be more susceptible to over-training than rm1d5. Where optimization of rm4d5 resulted in distinctly different parameter values for the two training systems, this could be evidence of over-training. But for every such case, optimization of rm1d5 resulted in a similarly distinct difference of parameter values for the two systems. This suggests that rm4d5 does not lead to

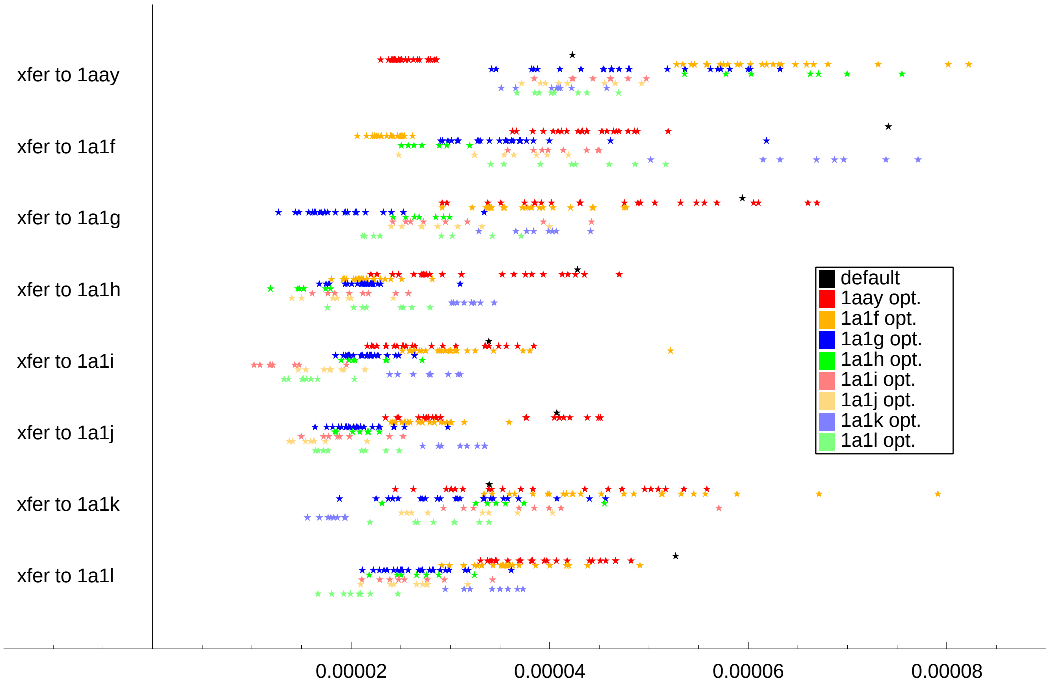

Figure 5-8: ACE vs. FDPB group binding free energy term error function rm1d5 (in (kcal mol^{-1})^5) for 8 zinc finger structures, using the default parameter set and families of parameter sets optimized to minimize rm1d5 on each of the 8 structures. The single black point at the top of each horizontal section is for the default parameter set. For each of the 8 × 8 combinations of training and testing system, there is a row of data points for the family of parameter sets optimized on that training system, and transferred to that testing system.

over-training any more than rm1d5.

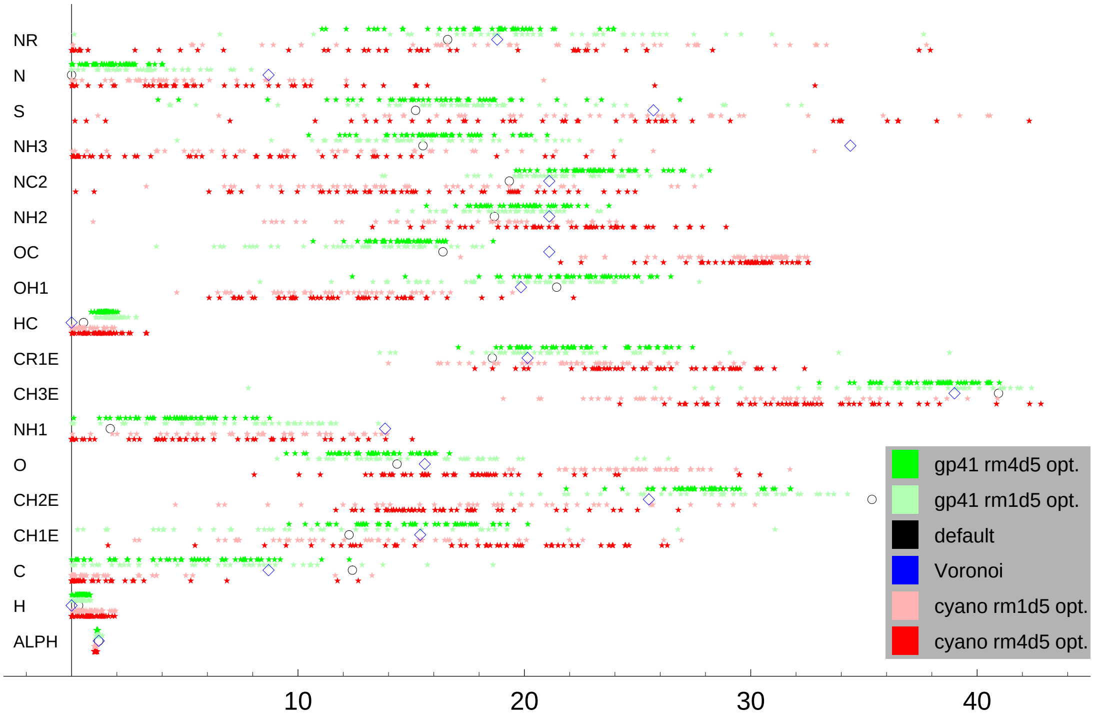

Figure 5-9: ACE parameter values before and after optimization of rm4d5 on the cyanovirin and gp41 systems. The horizontal axis has units of Å^3 for the volume parameters and is unitless for the α parameter. The ACE parameters are listed along the left side. For each parameter, its value before minimization is shown as one black circle. The values from Voronoi polyhedra (adapted from Richards [73] by Schaefer and Karplus [6]) are shown as one blue diamond. Parameters from each of 4 types of optimizations are shown as stars in 4 horizontal rows. For each parameter, the top row, in green, is from optimization of rm4d5 on gp41. The next row down, in light green, is from optimization of rm1d5 on gp41. The bottom row, in red, is from optimization of rm4d5 on cyanovirin. The next row up, in pink, is from optimization of rm1d5 on cyanovirin.

5.3.10 Characterizing an Error Function Surface in Parameter Space

Recall that we conceptualize the parameter space as 18 "horizontal" axes and the objective function as the vertical dimension, so that optimization is a search for low points in parameter space. It is very difficult to characterize a surface with so many dimensions, but we will now describe an experiment and some observations that suggest what the surface is like.

Starting from the default parameter set, we swept each parameter, one at a time, across the full range of reasonable values. At every position in parameter space visited, the rmsd of the solvation radii, rmsd(R), was calculated for the bound cyanovirin dimer with $\epsilon_{\mathrm{i}} = 1$. In Figure 5-10, we show, for each parameter, the error term rmsd(R) versus the parameter's value. Each curve is the profile of the error surface along a different axial direction. The surface of this error function is smooth but steep in some directions. Specifically, when an optimization starts from the default position, the "downhill" direction includes increasing α and decreasing $V(\mathrm{CH2E})$.

We have found that the largest ACE errors are usually caused by a handful of atoms whose atomic solvation energies are incorrectly determined to be too close to zero. Such atoms are then treated as though they were deeply buried in protein, so all of their interactions are too strong because they are not screened enough by solvent. We limit such errors by applying to the atomic solvation radii a flat cutoff at b_a, the adaptive maximum solvation radius described in Chapter 4. But this only limits the errors; it does not fix their root cause: the fact that it is difficult for ACE to accurately calculate an atomic solvation energy with a sum of many small terms for the desolvation caused by the presence of each other solute atom.

For the cyanovirin bound dimer, three particular atoms have this problem of incorrectly high solvation radii. Analysis of the atomic error terms reveals that increasing α or decreasing $V(\mathrm{CH2E})$ specifically alleviates these errors. (Several CH2E atoms are near the three problem atoms.)

This type of error is very sensitive to the parameters. For example, parameter sets optimized on gp41 have large errors when transferred to Arc. But lowering $V(\mathrm{NC2})$ from 19.8 to 15.8 Å^3 lowers the rm4d5 error by a factor of 1/5 to 1/60; and lowering $V(\mathrm{HC})$ as well, from 1.8 to 0.7 Å^3, lowers the rm4d5 error by an additional factor of 1/60 !

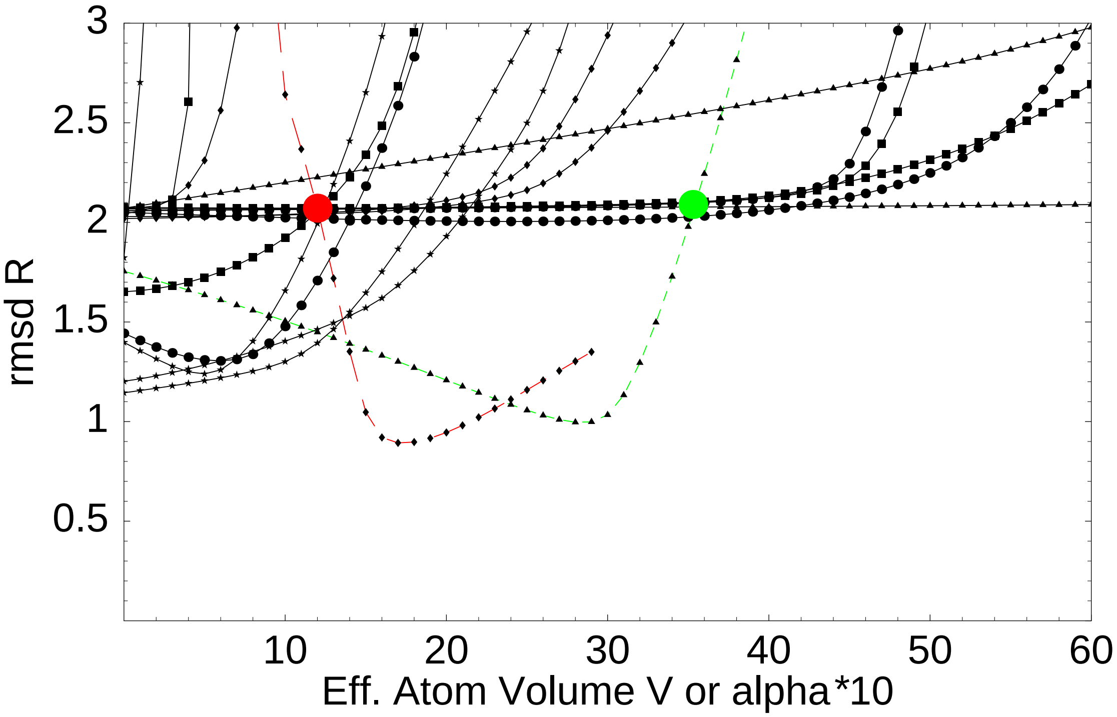

Figure 5-10: Solvation radius rmsd error rmsd(R) for the bound cyanovirin dimer, with $\epsilon_{\mathrm{i}} = 1$, as each parameter is swept, one at a time, over the full range of reasonable values, starting from the default parameter set. The rmsd(R) curves for all parameter sweeps are superimposed. The units of the horizontal axis are Å^3 for the effective atom volume parameters. For visibility, the unitless α parameter is multiplied by 10 before plotting its curve as a red line with long dashing. A large red dot marks the default value $\alpha = 1.2$; the height of the red dot is the rmsd(R) value with the default parameter set; and so the default value of each parameter can be found by seeing where each curve is at the same height as the red dot. The curve for V(CH2E), shown in green with short dashing, is interesting because it, like α, is steep at the default position in parameter space.

And below those approximate V(NC2) and V(HC) values, the error changes much more gradually (like the shape of the green curve in Figure 5-10). So, based on this limited exploration, the error surface seems to be smooth, but it has cliffs that slope up rather abruptly out of the lowlands.

5.3.11 Predicting the Ranking of Group Contribution Terms

From among the list of groups (amino, carbonyl, or side chain) in a protein-protein binding system, consider the 20 groups that have the most favorable contributions to the binding free energy. How many of these does ACE correctly predict to be in the top 20? A similar test can be posed to predict the 20 groups with least favorable contributions to the binding. For brevity, we will lump together 80 such tests: for cyanovirin and gp41, how many of each of their bottom 20 and top 20 does ACE correctly predict to be in the bottom or top 20? With the default parameter set, ACE gets 42 correct out of 80. Optimizing the error function rm4d5 improves this score to 51 correct out of 80. Optimizing function rm4d5 and transfering the parameters to the other protein system still scores better than the default parameters, getting 47 out of 80 correct.

5.3.12 Generalized Born Salt Treatment

We incorporated the approximate treatment of salt effects proposed by Srinivasan *et al.* [47] into ACE. The results in this chapter used zero ionic strength unless otherwise noted, but we investigated whether the optimization of parameters is affected by the presence or absence of salt in the model. Optimizing for the salt-free case means minimizing errors between ACE results (with no salt treatment) and FDPB results (with $I = 0$ M). Optimizing for the salt case means minimizing errors between ACE results (using the Srinivasan salt treatment with $I = 0.145$ M) and FDPB results (with $I = 0.145$ M).

The error rm4d5 in the salt case is lowered almost as much when the 18 parameters are optimized for the salt-free case and then transferred to the salt case, as when the

18 parameters are optimized for the salt case. This is true for function rm4d5 on the training system cyanovirin as well as after training on cyanovirin and transfering the parameters to the testing system gp41. We conclude that there is no need to optimize separate parameter sets for the salt and salt-free cases.

If the parameter k of this salt treatment is optimized as another parameter in addition to the 18 others, it takes a value from 0.3 to 0.5, rather than the 0.73 that Srinivasan *et al.* found best for predicting solvation free energies. (Of course, the other 18 parameters take different values than if only the 18 were optimized.) But the error is only very slightly reduced, so we feel this small error reduction is outweighed by the risk of degrading the prediction of salt effects after transferring the parameters to other systems; and we would recommend keeping their $k = 0.73$ value.

5.4 Conclusion

Although rigorous global optimization is not computationally feasible, local optima exist that are easily found by a low-temperature annealing schedule. Therefore one can afford to do multiple optimization runs differing only in their random seeds in order to find a set of local optima. This, in turn, allows one to statistically characterize the effect of chance on the optimization.

Optimizing parameters for more accurate atomic solvation energies does not give more accurate binding free energy terms; in fact, they worsen dramatically. Parameters optimized to minimize the group contribution or mutation binding free energy term errors do so by reducing the solvation term errors but increasing the interaction term errors. Minimizing an error function which is the product of all the error terms of interest — binding free energy terms for group interactions, solvation, contribution, and mutation — succeeds in reducing all of these group error terms.

Optimized parameters sets often give lower errors on systems unrelated to the training system. This is dependent on particular differences between the training and testing systems' structures; for example, the unusual degree of burial for the Arg and Lys side chains of the Arc repressor dimer means that parameters optimized on most other systems perform poorly when used on Arc. The shape of an error function landscape in parameter space seems to be smooth, but with high cliffs sloping up rather abruptly from the lower regions.

The treatment of salt effects which we added to the ACE model performs well regardless of whether the ACE parameters were optimized with or without a non-zero ionic concentration.

Optimization of the ACE parameters using FDPB data for a zinc finger structure yields parameter sets which still achieve a reduction of -25% (standard deviation 31%) in the rm1d5 error function when the parameters are applied to a different, but related, zinc finger structure. Therefore, our method of optimization for the ACE parameters is valuable for any application in which electrostatic free energies of multiple related structures are needed. Ligand design is one such application.

Chapter 6

The Design of Protein Binding Interfaces: Co-Optimization of Packing and Electrostatic Interactions

6.1 Introduction

Biomolecular binding is essential to many aspects of biology. The functions of proteins, which generally involve molecular association, are dictated by their structures. A protein's structure, in turn, is determined by its sequence. Protein design seeks sequences and corresponding structures with improved binding or stability, or with novel binding abilities.

Discrete search algorithms such as dead-end elimination and A*, which we will introduce and then use in this chapter, allow vast numbers of conformations to be searched systematically. These algorithms have mainly been applied with the goal of increasing stability, for example in the repacking of a protein's hydrophobic core. We extend these search algorithms so that they can be used to optimize both stability and binding affinity.

The treatment of solvation and electrostatics used with discrete search algorithms has been rather crude so far. An accurate treatment of electrostatics, such as the finite-difference solution of the Poisson-Boltzmann equation, is computationally expensive and intrinsically incompatible with these search algorithms because the interaction of an atom pair depends on the location of the other atoms. We introduce a hierarchical treatment of solvation and electrostatics, using three energy functions of increasing accuracy. We show that the correlation of these energy functions allows them to be used as successive screens, narrowing down the list of promising structures.

We also develop a three-stage dead-end elimination/A* procedure which ensures that a wide variety of sequences, as well as a variety of conformations for each sequence, can be passed to the two higher-resolution energy functions. This allows us to overcome the less accurate electrostatic treatment of the low-resolution energy function.

Our protein design method is developed by using it to redesign three residues on the protein barstar to enhance its binding to its partner barnase, while maintaining its folding stability. Our screening protocol is then validated by using it on a redesign of three residues of the HIV-1 glycoprotein gp41. Finally, seven residues of barstar, which have been found to be critical to its binding to barnase, are redesigned. Several mutations predicted to bind more tightly than the wild type, while retaining folding stability, are promising candidates for synthesis. Wild-type barstar and barnase are experimentally known to bind very tightly. So our method's extremely high ranking of the wild-type sequence among all other sequences compares favorably with the lower ranking assigned by the low-resolution energy function used in previous protein design studies.

6.2 Theory: Discrete Conformational Searching Using the Dead-End Elimination and A*Algorithms

To redesign proteins, we assume that we can keep the backbone and most of the residues fixed and only change or move the atoms of a subset of "mobile" residues which we want to redesign. This is perhaps the broadest simplification that we make, but there are very many actual cases in which this has proven to be reasonable: for many pairs of binding partners, a wide variety of mutant versions have been found by X-ray crystallography to adopt very nearly the same bound conformation.

We represent the space of all possible conformations by allowing a discrete set of side chain conformations, called "rotamers", at each of the mobile residues. This makes the problem of finding minimum-energy conformations amenable to systematic search methods. The transition from a continuous to a discrete search space can cause minimum-energy conformations to be overlooked only if the library of rotamers is not fine enough. Search techniques such as dead-end elimination and A*, described in this chapter, have made it possible to identify the global minimum-energy conformation (GMEC) from among more than 10^{40} combinations of protein sequence and conformation [74, 75], a vastly larger number of conformations than continuous methods such as molecular dynamics could ever search. In addition, these discrete search methods allow us to prove that we have not missed the GMEC, unlike all non-discrete methods, and unlike discrete Monte Carlo or simulated annealing methods as well.

Each conformation of the system can be defined by specifying, for each mobile residue i, a rotamer state i_r. This formalism is not limited to protein side chains, although we will call the mobile groups "residues". All atoms in the system which are not part of a mobile residue are "fixed" atoms. To design the sequence as well as the conformation, one simply allows all rotamers of all amino acid types at each mobile residue. Each conformation of each sequence will be called a "rotamer state", or simply a "structure."

The main simplification required to use DEE and A* is that the energy be made pairwise additive:

$$E = E_{\text{fixed}} + \sum_{i=1}^{p} E_{\text{self}}(i_r) + \sum_{i=1}^{p} \sum_{j=i+1}^{p} E_{\text{pair}}(i_r, j_s) \tag{6.1}$$

where p is the number of mobile residues, E_{fixed} is the energy of the fixed atoms, $E_{\text{self}}(i_r)$ is the energy contribution of mobile residue i in rotamer i_r, including its interaction with the fixed atoms and with itself, and $E_{\text{pair}}(i_r, j_s)$ is the interaction of rotamers i_r and j_s at mobile residues i and j. The assumption of a pairwise additive energy means that the interaction of two residues depends only on their positions, but not on the positions of the other mobile residues.

In principle, with discrete rotamers, the finite number of conformations of the whole system could be exhaustively searched. In practice, for p mobile residues and n possible rotamers at each position, the n^p possible conformations of the whole system are usually too numerous to evaluate exhaustively. The search algorithms DEE and A*, described in this chapter, allow the GMEC to be found in a feasible amount of time, without having to evaluate all n^p conformations. Since the number of $E_{\text{self}}(i_r)$ and $E_{\text{pair}}(i_r, j_s)$ terms, np and $\frac{1}{2}p(p-1)n^2$, are usually much less than n^p, it is advantageous to precompute all of them.

6.2.1 Dead-End Elimination

Dead-end elimination (DEE) was originally proposed by Desmet *et al.* [76]. DEE is a way to reduce the size of a very large search space by eliminating rotamers which can not be in the desired rotamer set or sets.

Consider two rotamers of the same residue, i_r and i_s. Rotamer i_r can not be part of the GMEC if its *best* possible energy contribution is still worse than the *worst* possible energy contribution of rotamer i_s:

$$\left[E_{\text{self}}(i_r) + \sum_{j \neq i} \min_t E_{\text{pair}}(i_r, j_t)\right] - \left[E_{\text{self}}(i_s) + \sum_{j \neq i} \max_t E_{\text{pair}}(i_s, j_t)\right] > 0 \tag{6.2}$$

This is the simple DEE criterion [76], illustrated in Figure 6-1. It says that if we let all the other mobile residues interact as well as possible with rotamer i_r, and as poorly as possible with rotamer i_s, and i_r still gets a worse energy than i_s, then i_r is definitely not in the GMEC.

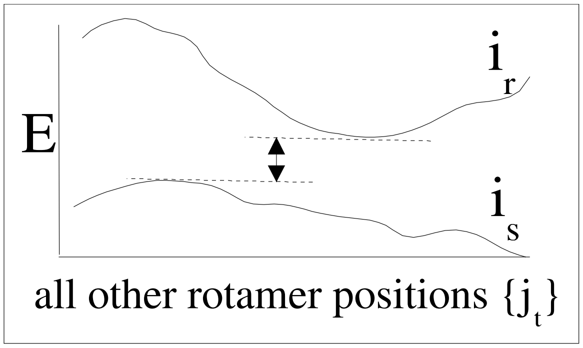

Figure 6-1: Simple DEE criterion, Equation 6.2.

Goldstein [77] improved upon this by eliminating rotamer i_r if it has a worse energy than i_s for all possible conformations of the other mobile residues:

$$E_{\text{self}}(i_r) - E_{\text{self}}(i_s) + \sum_{j \neq i} \min_t (E_{\text{pair}}(i_r, j_t) - E_{\text{pair}}(i_s, j_t)) > 0 \tag{6.3}$$

This is the Goldstein DEE criterion, illustrated in Figure 6-2. The advantage of this method is that it has greater eliminating power than the original criterion (for example, the original method would not eliminate i_r in Figure 6-2); the disadvantage is that it is computationally more expensive, scaling as n^2p rather than np, where p is the number of mobile residues and n is the number of possible rotamers per residue.

Pierce *et al.* [78] and Looger and Hellinga [74] went a step further by partitioning the conformational space by the rotamers k_v of mobile residue k (or k could represent a group of mobile residues). If, for every rotamer k_v, there exists a rotamer i_s such that

$$E_{\text{self}}(i_r) - E_{\text{self}}(i_s) + \sum_{j \neq i \neq k} \min_s [E_{\text{pair}}(i_r, j_t) - E_{\text{pair}}(i_s, j_t)] + [E_{\text{pair}}(i_r, k_v) - E_{\text{pair}}(i_s, k_v)] > 0 \tag{6.4}$$

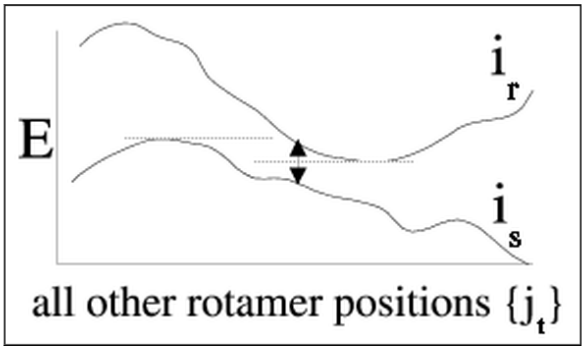

Figure 6-2: Goldstein DEE criterion, Equation 6.3.

then rotamer i_r can be eliminated. The eliminating rotamer i_s may be different in different partitions k_v. For example, for some arrangements of the rest of the system, rotamer i_{s1} is always better than i_r; and for other arrangements of the rest of the system, rotamer i_{s2} is always better than i_r. For every arrangement of the rest of the system, there is a better alternative to i_r, and so i_r is eliminated. This type of approach is often referred to as “divide-and-conquer.”

Any of the criteria above can be extended to eliminate pairs of rotamers, which is sometimes necessary when elimination over individual residues fails to reduce the search space sufficiently. The simple DEE criterion applied to pairs is

$$\left[E_{\text{self}}([i_r j_s]) + \sum_{k \neq j \neq i} \min_t E_{\text{pair}}([i_r j_s], k_t)\right] - \left[E_{\text{self}}([i_u j_v]) + \sum_{k \neq j \neq i} \max_t E_{\text{pair}}([i_u j_v], k_t)\right] > 0 \tag{6.5}$$

Here we essentially define $[i_r j_s]$ as a single effective rotamer. So its self term includes the interactions of rotamers i_r and j_s with themselves, with each other, and with the fixed atoms. Likewise, $E_{\text{pair}}([i_r j_s], k_t)$ is the interaction between the rotamer pair $[i_r j_s]$ and the single rotamer k_t. If the criterion above holds, then the pair of rotamers i_r and j_s can not be together in the GMEC, but either one or the other alone may still be in the GMEC.

The Goldstein criterion applied to pairs is

$$E_{\text{self}}([i_r j_s]) - E_{\text{self}}([i_u j_v]) + \sum_{k \neq j \neq i} \min_t [E_{\text{pair}}([i_r j_s], k_t) - E_{\text{pair}}([i_u j_v], k_t)] > 0 \tag{6.6}$$

DEE can also be used to help search for structures with energy within a given distance ΔE_{cut} from the minimum energy. Each of the criteria above can be modified so that a rotamer or rotamer pair is eliminated only if it can not be part of any conformation with energy less than ΔE_{cut} above the minimum energy. For example, Equation 6.3 becomes:

$$E_{\text{self}}(i_r) - E_{\text{self}}(i_s) + \sum_{j \neq i} \min_t (E_{\text{pair}}(i_r, j_t) - E_{\text{pair}}(i_s, j_t)) > E_{\text{cut}} \tag{6.7}$$

However, this weakens the technique considerably. We will discuss how we use DEE to reduce the search space, and then use A* to search within the reduced space.

As an example of the computational cost of DEE, consider the single-residue Goldstein criterion in Equation 6.3: the equation contains 2 arbitrary mobile residues, i and j, and 3 arbitrary rotamers r, s, and t. Assuming each of the p mobile residues has the same number of possible rotamers n, the computational cost of applying the single-residue Goldstein criterion in every possible permutation is $O(n^3p^2)$ ("of order" n^3p^2). The scaling of the cost of the other possible DEE criteria are summarized in Table 6.2.1. Of course, they are all vastly smaller than the $O(n^p)$ cost scaling of a systematic search of all conformations.

Table 6.1: Dependence of DEE computational cost on the number of mobile residues (p) and on the number of rotamers per residue (n).

Criterion	# of min. calculated	time to find a min.	Total time
Simple DEE (Eqn. 6.2)	np	np	n^2p^2
Goldstein DEE (Eqn. 6.3)	n^2p	np	n^3p^2
Split DEE (Eqn. 6.4)	n^2p	np	n^3p^2
Simple Pair DEE (Eqn. 6.5)	n^2p^2	np	n^3p^3
Goldstein Pair DEE (Eqn. 6.6)	n^4p^2	np	n^5p^3

Our implementation of DEE follows a schedule including Goldstein singles and pairs, and split singles:

1. Goldstein singles DEE (Equation 6.3) at every mobile position, repeated until no more rotamers can be eliminated.
2. Split singles DEE (Equation 6.4) at every mobile position, repeated until no more rotamers can be eliminated.
3. Goldstein pairs DEE (Equation 6.6) for every rotamer pair, repeated until no more rotamer pairs can be eliminated. Whenever the pairs criterion eliminates all the pairs that a rotamer i_r can make with some residue j, then i_r itself can be eliminated [79].
4. If pairs DEE eliminated any single rotamers, then try repeating the whole procedure from the top.

This procedure stops when no more rotamers can be eliminated. In general, a large number of system conformations will remain.

6.2.2 A* Search (Branch and Bound)

The search algorithm called A* can now be used to search the system conformations remaining after DEE [80, 81]. Imagine a tree representing partial or complete sets of rotamers for the system, as shown in Figure 6-3. The mobile residues are kept in a fixed order, and their rotameric states are decided upon in that order. The root node (at the top) means that nothing has been decided yet. The row of n_1 nodes below that represent all possible rotamers that can be placed at the first mobile residue. Each node in that row has n_2 branches, representing all the possible rotamers that can be placed at the second mobile residue. And so on, until the bottom row of the tree has $\prod_{i=1}^{p} n_i$ nodes, each of which is a "goal" node, or a complete conformation of the system. Throughout our treatment of the theory, we set all n_i to the same value, n, and so there are n^p goal nodes; for actual applications, the n_i can be arbitrarily distributed.

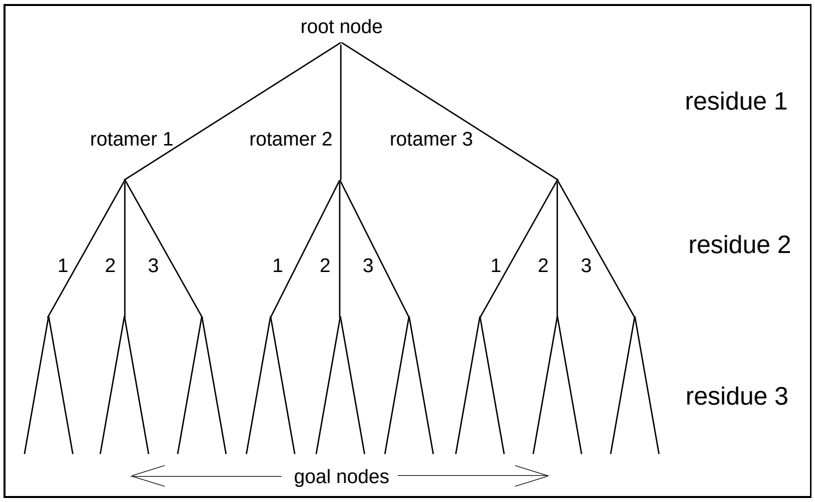

Figure 6-3: A conformational search represented as a tree. The first branching, from the top, or "root" node, represents placing each possible rotamer at mobile residue 1. The next set of branchings represent placing each possible rotamer at mobile residues 2, and so on. Each node at the bottom of the tree is a "goal" node, representing one of the possible conformations of the whole system. The figure is adapted from reference [81] and taken with permission from reference [69].

The A* algorithm is a method for finding the optimal path from the root node to a goal node of a search tree. In this problem, there is only one direct path from the root node to each node in the bottom row, and the optimal path represents the GMEC of the system. The A* algorithm scores each node that it visits with a function $\mathbf{f}^*$, which is a sum of $\mathbf{g}^*$, the known cost to get there from the root node, and $\mathbf{h}^*$, a heuristic lower-bound estimate of the cost to get from there to a goal node.

$$\mathbf{f}^* = \mathbf{g}^* + \mathbf{h}^* \tag{6.8}$$

So $\mathbf{f}^*$ for a given node is an estimate of the minimum total energy of the system, given the rotamers that have been placed so far at that node. Number the levels of the tree by p_f, the number of mobile residues placed so far at that level. At a given node, mobile residues numbered 1 to p_{f} have been placed, and mobile residues numbered $p_{\mathrm{f}}+1$ to p have not yet been placed. With a pairwise additive energy function, the cost $\mathbf{g}^*$ to get to the given node is

$$\mathbf{g}^* = E_{\mathrm{fixed}} + \sum_{i=1}^{p_{\mathrm{f}}} E_{\mathrm{self}}(i_r) + \sum_{i=1}^{p_{\mathrm{f}}} \sum_{j=i+1}^{p_{\mathrm{f}}} E_{\mathrm{pair}}(i_r, j_s) \tag{6.9}$$

The $\mathbf{h}^*$ function is an estimate of the cost of reaching a goal node, but the algorithm requires that $\mathbf{h}^*$ always underestimate this cost. So $\mathbf{h}^*$ is a lower-bound estimate of the cost to get from the current node to any goal node. At any goal node, where all residues have been placed, $\mathbf{g}^*$ is equal to the total energy of the conformation, and $\mathbf{h}^*$ is zero.

In the basic version of A*, the data structure used to implement the search is a list of nodes sorted by $\mathbf{f}^*$. The algorithm repeatedly finds the node with the minimum value of $\mathbf{f}^*$ and expands it, finding the values of $\mathbf{f}^*$ for all of the nodes immediately below that node. These nodes are all added to the list of nodes, and the process repeats by expanding the node that now has the minimum $\mathbf{f}^*$. The process continues until the node with minimum $\mathbf{f}^*$ is a goal node. Note that the nodes in the list can be at a variety of levels in the tree. Because $\mathbf{f}^*$ of a node is a lower-bound estimate of the $\mathbf{f}^*$ of any goal node beneath it, a goal node with lower $\mathbf{f}^*$ than the other nodes on the list must also have lower $\mathbf{f}^*$ than all other goal nodes, and therefore must be the GMEC.

The computational cost of the A* algorithm depends on a case-by-case basis. In the worst case, all $O(n^p)$ nodes will have to be visited. In the best case, the algorithm will follow a direct path down one branch of the tree, visiting $O(np)$ nodes. The efficiency of the algorithm also depends on

1. the quality of $\mathbf{h}^*$, the lower-bound estimate of the cost to get from the given node to any goal node, and

2. the order in which the mobile residues are expanded as one goes down the tree.

It is also important that $\mathbf{h}^*$ be calculated quickly, since this calculation will be the most computationally difficult step.

We will now introduce two different expressions for $\mathbf{h}^*$, and then use the second because it is a higher, and therefore better, lower-bound estimate of the cost to get from the given node to any goal node. The first, proposed by Leach and Lemon [81], is straightforward:

$$\mathbf{h}^* = \sum_{j=p_\mathrm{f}+1}^{p} \min_s \left[E_\mathrm{self}(j_s) + \sum_{i=1}^{p_\mathrm{f}} E_\mathrm{pair}(i_r, j_s) + \sum_{k=j+1}^{p} \min_t E_\mathrm{pair}(j_s, k_t) \right] \tag{6.10}$$

where mobile residues 1 to p_f are those that have been placed so far at the current node. The first sum is over the mobile residues which have not been placed yet. The first and second terms inside the square brackets are the self term of j_s and the interactions of j_s with the mobile residues already placed. The third and final term is a lower bound for the interactions of j_s with the residues not yet placed. Since the last term inside the brackets can be computed and stored for each j_s, the calculation of $\mathbf{h}^*$ for each node scales as $O(np)$.

A second expression for $\mathbf{h}^*$, used by Gordon and Mayo [82] with an algorithm similar to A*, defines new energy terms E'_self and E'_pair by dividing each rotamer's self term among its interaction terms:

$$\begin{aligned} E'_\mathrm{self}(i_r) &= 0 && (6.11) \\ E'_\mathrm{pair}(i_r, j_s) &= \frac{E_\mathrm{self}(i_r) + E_\mathrm{self}(j_s)}{p-1} + E_\mathrm{pair}(i_r, j_s) && (6.12) \end{aligned}$$

Putting E'_self and E'_pair into Equation 6.10 in place of E_self and E_pair,

$$\mathbf{h}^* = \sum_{j=p_\mathrm{f}+1}^{p} \min_s \left[\sum_{i=1}^{p_\mathrm{f}} E'_\mathrm{pair}(i_r, j_s) + \sum_{k=j+1}^{p} \min_t E'_\mathrm{pair}(j_s, k_t) \right] \tag{6.13}$$

Because part of the self terms are now inside the last min operator with the pair terms, this definition of $\mathbf{h}^*$ in Equation 6.13 almost always gives a higher, and therefore better, lower bound than Equation 6.10. Tests performed with the two bounds have shown that Equation 6.13 results in a faster search [69], and so this is the one we use.

The order in which the mobile residues are placed does affect the values of $\mathbf{h}^*$ and the speed of the search. Leach and Lemon [81] used a heuristic method to choose the order wisely. For each rotamer, this quantity is calculated:

$$V(i_r) = E_{\text{self}}(i_r) + \sum_{j \neq i}^{p} \min_s E_{\text{pair}}(i_r, j_s) \tag{6.14}$$

For each mobile residue position i, the difference of the two lowest values of $V(i_r)$ is computed. The residue with the largest difference is expanded first in the tree, followed by the residue with the second largest difference, and so on.

6.2.3 Depth-First A* Search

After the A* algorithm finds the GMEC, the goal node with minimum energy $E = E_{\text{min}}$, it could continue running in order to find the next-lowest energy goal node, and so on. In practice, however, the list of nodes which the method maintains grows too large for available computer memory when the algorithm is used this way. A different type of search, depth-first A*, is better suited to finding additional low-energy conformations; specifically, all conformations with energy within ΔE_{cut} of the minimum. A full depth-first search could begin traversing the tree by going all the way down the left-most branch of the tree, then taking one step up to go back down to the next leaf, and so on until the whole tree has been traversed. A full traversal of the tree is not feasible in our case, so we use the depth-first A* search, which can skip over large parts of the tree by evaluating $\mathbf{f}^*$ at each node it visits. As the depth-first search proceeds, each node's $\mathbf{f}^*$ is calculated. If $\mathbf{f}^* > E_{\text{min}} + \Delta E_{\text{cut}}$, then all goal nodes beneath that node have energy $E > E_{\text{min}} + \Delta E_{\text{cut}}$, so the search travels back up from the node rather than exploring the fruitless subtree beneath it.

6.2.4 Search Procedure

Here is our complete DEE/A* search procedure:

1. Find the rotamer state with minimum energy E, the GMEC:

   (a) DEE to narrow down the set of rotamers that may by in the GMEC. Following the schedule at the end of Section 6.2.1, we eliminate rotamers using Goldstein singles, then split singles, then Goldstein pairs, then repeating until no more rotamers can be eliminated.

   (b) A* to find the GMEC among all combinations of the remaining rotamers. Call its energy E_{min}.

2. If, in addition to the GMEC, one wants all rotamer states with E within ΔE_{cut} of the GMEC energy E_{min}, then:

   (a) DEE, starting over with all rotamers allowed, to narrow down the set of rotamers that may be in any structure within ΔE_{cut} of the GMEC energy E_{min}.

   (b) Depth-first A* to find all rotamer states with $E \leq E_{\text{min}} + \Delta E_{\text{cut}}$.

6.2.5 Rotamer Library

The rotamer library is a list of side chain conformations. Since bond lengths and angles are fairly constant, each conformation can be defined by a list of dihedral angles χ. Most χ angles of rotatable bonds have local minima in the vicinity of −60°, +60°, and +180°. We begin with the rotamer library of Dunbrack and Karplus [83], which has 3 such values for most χ angles, whose precise values were determined separately for each amino acid type based on a statistical analysis of many protein structures.

We allow 19 natural amino acid types, all but proline, because its backbone is different. We allow two forms of histidine, singly protonated on ND1 or NE2. We also double the number of histidine rotamers by allowing χ_2 to flip the ring by 180°, because PDB X-ray crystal structures (upon which the Dunbrack and Karplus library is based) can not be trusted to have histidine rings flipped correctly. We do not allow protonated forms of

Asp, Glu, and His, but our method can be extended to allow this if care is taken to include the correct free energy cost of protonation.

Using discrete rotamers can have the disadvantage of being too coarsely sampled. Especially since the van der Waals energy function is so sharp, a rotamer or pair of rotamers may not fit, even though small adjustments in their positions would allow them to. We make up for this in two ways: We expand the rotamer library to include adjustments of ±10° to the χ_1 and χ_2 dihedrals. This increases the number of rotamers by a factor of 9 for most amino acid types. Expanding the rotamer library so much puts us in danger of going too far by sampling too finely: the search required of DEE and A* is much harder, not just because there are more rotamers, but also because the minima are not as sharp; many of the rotamer states have similar, low energies.

6.2.6 Fleximers

Our solution is to employ the flexible rotamer model suggested by Mendes *et al.* [84]. Each Dunbrack and Karplus rotamer is grouped with the 8 rotamers related to it by χ_1 and χ_2 adjustments of ±10°. These groups of rotamers $\{i_r\}$ are called flexible rotamers, or “fleximers”, and denoted with symbols like $i_{\boldsymbol{\mathcal{R}}}$. The individual rotamers within a fleximer are called its subrotamers, or rigid rotamers, or simply rotamers. Throughout our method, we will use 3 levels of description: the sequence (the set of amino acids at the mobile residues), the fleximer state (the set of fleximers), and the rotamer state (the set of rotamers). Only the rotamer state defines an exact conformation, but we would like to define an approximate energy F for each fleximer state, so that we can apply DEE and A* to finding low-energy fleximer states in their smaller search space.

We require that F be pairwise additive, like the energy E:

$$F = E_{\text{fixed}} + \sum_i F_{\text{self}}(i_{\boldsymbol{\mathcal{R}}}) + \sum_i \sum_{j>i} F_{\text{pair}}(i_{\boldsymbol{\mathcal{R}}}, j_{\boldsymbol{\mathcal{S}}}) \tag{6.15}$$

where $F_{\text{self}}(i_{\boldsymbol{\mathcal{R}}})$ is the contribution of a single fleximer to the fleximer energy F of the system, and $F_{\text{pair}}(i_{\mathcal{R}}, j_{\mathcal{S}})$ is the contribution of a pair of fleximers to F. Our definitions for F_{self} and F_{pair} are the simplest forms used by Mendes *et al.* [84]:

$$\begin{aligned} F_{\text{self}}(i_{\mathcal{R}}) &= \min_{r\in\mathcal{R}} E(i_r) && (6.16) \\ F_{\text{pair}}(i_{\mathcal{R}}, j_{\mathcal{S}}) &= \min_{\substack{r\in\mathcal{R}\\ s\in\mathcal{S}}} [E(i_r) + E(j_s) + E(i_r, j_s)] - F_{\text{self}}(i_{\mathcal{R}}) - F_{\text{self}}(i_{\mathcal{S}}) && (6.17) \end{aligned}$$

The value of F for a given fleximer set is meant to approximate the lowest E of any of its rotamer sets. But, because we approximated it as pairwise additive, F will not necessarily equal the E of any particular rotamer set [69]. The meaning of the min operator in the Mendes *et al.* [84] definition of F_{pair} in Equation 6.16 is illustrated in Figure 6-4. A fleximer $i_{\mathcal{R}}$ is allowed to get credit for its best possible interaction with a neighboring fleximer $j_{\mathcal{S}}$ *and* for its best possible interaction with another neighboring fleximer $k_{\mathcal{T}}$, even if both interactions are not possible for any set of rotamers i_r, j_s, and k_t. In this case, F underestimates E. It is also possible for F to overestimate E, because of the way the self terms do not cancel out in the F_{pair} definition. Unfortunately, this means that we may not find the true GMEC, even though this was previously guaranteed by the DEE and A* methods. Alternative definitions of F_{self} and F_{pair} have been developed to restore the guarantee that $F \leq E$ and therefore no low energy structures will be missed [69]. But such definitions make F underestimate E by so much that many more fleximer states with low F must be examined to find any rotamer states with low true energy E. So we use the Mendes *et al.* [84] definitions in Equation 6.16.

So, we do DEE/A* as in Section 6.2.4 using the fleximer energy F to get a list of all fleximer states with F within 30 kcal mol^{-1} of the minimum. Then, for each fleximer state on the list, we do DEE/A* again to find the minimum energy E rotamer state that it contains. This is a rather small search space, so the DEE/A* proceeds rapidly, and a rotamer state representing each of thousands of fleximer states can be found in a reasonable amount of time. When this has been done for every fleximer state within 30 kcal mol^{-1} of the minimum F, we are done. The final product is a list of rotamer states, each representing a different fleximer state, sorted by energy E.

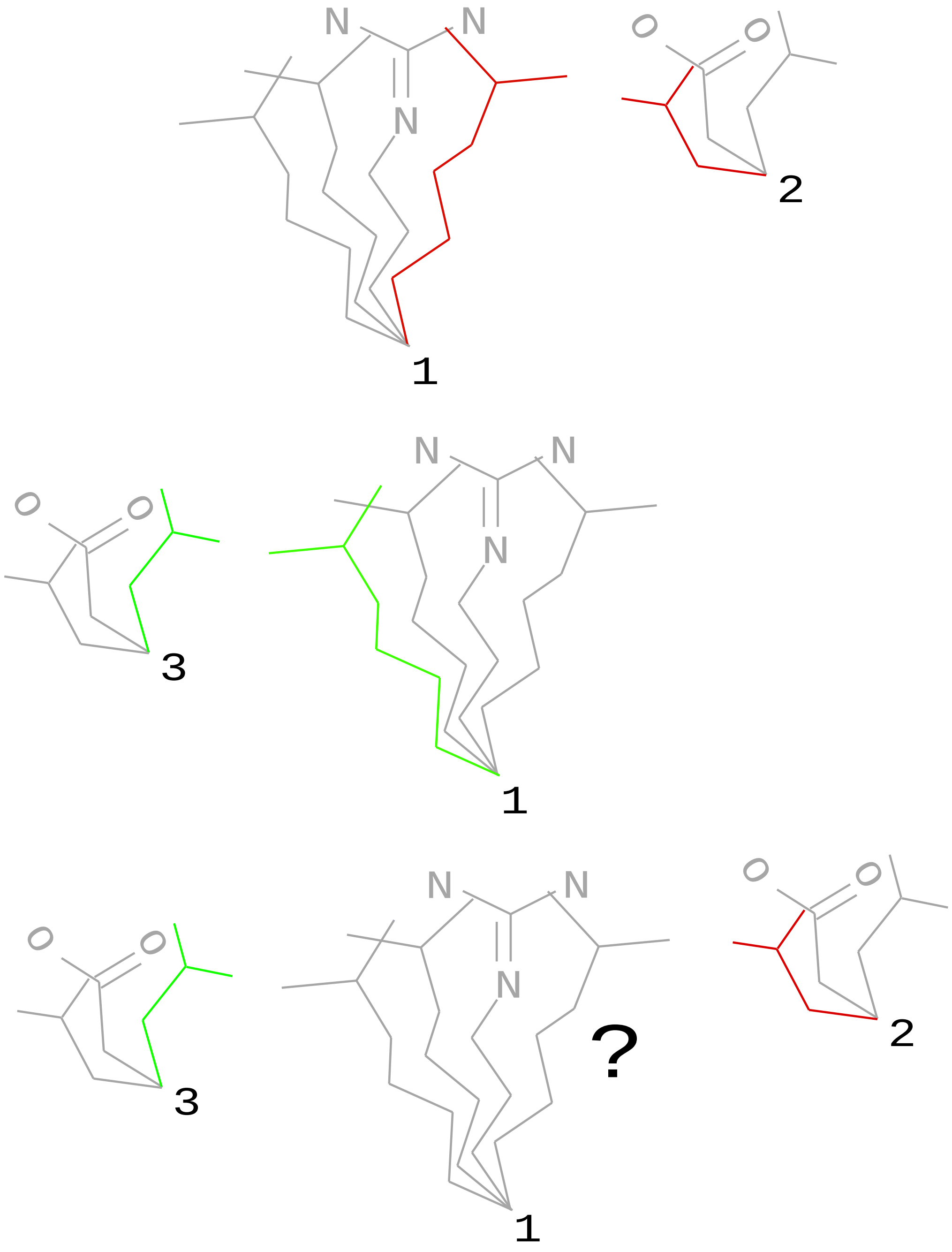

Figure 6-4: Illustration of the weakness of approximating the fleximer energy F as pairwise additive. Residues 1 and 2 are given credit for their best possible interaction, using the rigid rotamers shown in red. Similarly, residues 1 and 3 are given credit for their best possible interaction, using the rigid rotamers shown in green. Since residue 1 can not be in two places at once, no set of rigid rotamers can fully realize both favorable interactions. The figure is taken with permission from reference [69].

The method we have just described — searching for fleximer states, then for rotamer states — was used by Mendes *et al.* [84] and Caravella [69]. In Section 6.3.3, we describe how we use 3 levels of description rather than 2, searching for low-energy sequences, then fleximer states, then rotamer states.

6.3 Methods

6.3.1 Designing for Tight Binding and Stable Folding

The usual goal of ligand design is to redesign a molecule in order to bind as tightly as possible to its binding partner. In the case of proteins, one would also require the redesigned protein to remain folded in the unbound state. So that is exactly what we will do: find redesigned molecules with minimum binding free energy, and with folding free energy below a cutoff to ensure stability of the unbound protein.

Previous attempts at computational ligand design [85, 86, 69] have aimed to maximize the stability of the bound complex. Such a ligand could achieve this either by improved interactions with the binding partner, or by improved intramolecular interactions, or a combination of the two. There is no guarantee that such a redesigned ligand will have better binding free energy than the wild type; nor is there a guarantee that its folded form will be stable.

The other major shortcoming of previous protein design methods has been their use of a poor approximation of electrostatic free energy. For computational ease, they have used Coulombic or distance-dependent Coulombic energy, neither of which includes the important terms for the desolvation of charges and the screening of intramolecular interactions upon binding. We have used the finite-difference solution of the Poisson-Boltzmann equation (FDPB), which is a reasonable, accurate, and computationally tractable estimate using the continuum solvent model, for the final evaluation of structures. But FDPB is still computationally expensive, so we have developed a procedure which uses 3 energy functions, with electrostatic terms of low, medium, and high accuracy (and corresponding computational cost) as successive screens. The "low-resolution" energy uses distance-dependent Coulombic electrostatic energy, which is pairwise additive (the atom-pair interactions are independent of the other atoms), and therefore can be used with the dead-end elimination (DEE) and A* algorithms. Structures which look promising based on their low-resolution energy are passed on to the "medium-resolution" energy function, which uses the ACE electrostatic approximation of Schaefer and Karplus [6]. Finally, structures which look promising based on their medium-resolution energy are passed on to the "high-resolution" energy function, which uses FDPB electrostatics. The key requirement of this approach is that the energy function used at one stage is sufficient to eliminate poor structures yet does not discard structures that would score well in subsequent stages.

6.3.2 Brief Overview of Entire Design Procedure

Here is a brief overview of the entire design procedure. Some terms used here are not defined until later.

1. Begin with the atomic structure of 2 binding partners.

2. Choose a set of side chains to redesign (the "mobile residues").

3. Choose a set of amino acid types to allow at each position (e.g., all but proline).

4. Choose a rotamer library, which has all discrete positions for each amino acid type.

5. Carry out DEE/A* to rank all sequences with a low-resolution pairwise additive binding free energy within some cutoff (ΔE_{cut}) of the minimum. Here $\Delta E_{\text{cut}} =$ 30 kcal mol^{-1} was used. It was advantageous to represent the results at this stage as a list of sequences, each represented by 10 structures (each defined by its set of rotamers).

6. After DEE/A*, follow a strategy to:

   (a) Evaluate the medium-resolution free energy function only for structures with promising low-resolution free energy;

   (b) Evaluate the high-resolution free energy function only for structures with promising medium-resolution free energy.

7. Reject all structures with a high-resolution folding free energy more than 1 kcal mol^{-1} worse than the wild type.

8. Keep only one structure to represent each sequence — the one with the best high-resolution binding free energy.

9. Final result: a list of sequences and their high-resolution binding and folding free energies.

6.3.3 Three Stages of DEE/A*: Amino Acids to Fleximers to Rotamers

Naive Two-Stage Method: Fleximers to Rotamers

Rather than just running DEE/A* to rank the possible sets of rotamers at the mobile residue positions, we described in Section 6.2.6 how rotamers can be separated into groups called fleximers. Then DEE/A* can be run in two stages: first to rank the fleximer states, and then, for as many of the top-ranked fleximer states as desired, to rank the rotamer states within each fleximer set, or to find only the one best rotamer set for each fleximer set. This two-stage DEE/A* method was used by Mendes *et al.* [84] and Caravella [69]. The method is pictured in Figure 6-5.

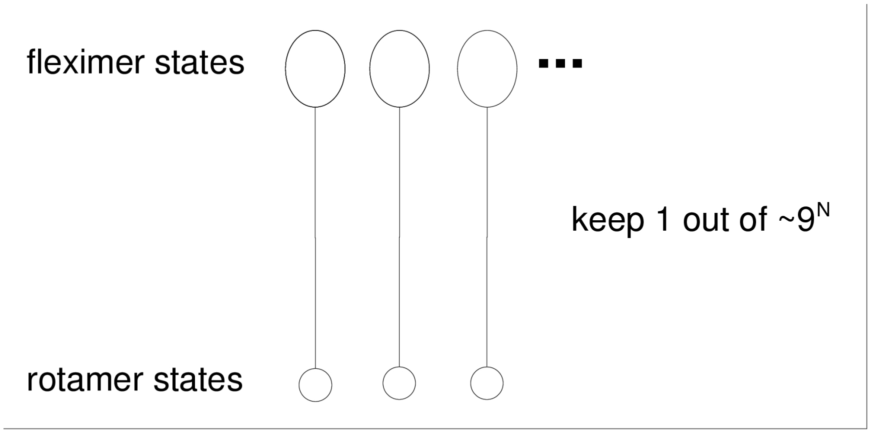

Figure 6-5: Naive Two-Stage Method: Fleximers to Rotamers

This method works well when DEE/A* is used to re-pack side chains without changing their amino acid types. With the rotamer library we used, the amino acid types have an average of 155 rotamers each. These are grouped into about 17 fleximers each. Each fleximer has about 9 rotamers. (Some amino acids have far more conformational freedom than others. Those averages are based on from 1 to 729 rotamers per amino acid type, grouped into 1 to 81 fleximers, each containing from 1 to 27 rotamers.)

When amino acid type is allowed to vary, however, each position has 3098 possible rotamers grouped into 338 fleximers. The first stage produces a ranked list of fleximer states which has too little diversity of amino acid sequence. The low-resolution energy function will typically favor some sequences over others, and each sequence will be represented by vast numbers of fleximer states on the list. Sequences which the low-resolution energy function does not favor (some of which *would* be favored by the high-resolution energy function) appear only a few kcal mol^{-1} higher in energy than the minimum, but such impossibly huge numbers of other fleximer states rank higher that the list could never practically be made long enough to find these sequences.

So the problem with this method is that each sequence is represented by far too many fleximer states, and so we can not get a large enough variety of sequences.

Naive Two-Stage Method: Amino Acids to Rotamers

In order to get a larger variety of sequences, let us have the first stage of DEE/A* rank sequences (i.e., amino acid states). This stage produces a ranked list of sequences fairly rapidly. Then, for each sequence on the list, a second stage of DEE/A* can be done to rank rotamer states, perhaps with the intention of keeping only the best few rotamer states to represent each sequence. This method is pictured in Figure 6-6. But the second stage, because it must consider about 155^N possible rotamer states, is too slow (5 minutes per sequence on a 1 GHz Pentium III), considering that we want to sort through $\sim 10^5$ sequences.

Three-Stage Method: Amino Acids to Fleximers to Rotamers

To get a large number of possible sequences in a reasonable amount of time, we have decided on a three-stage method, pictured in Figure 6-7.

1. Do DEE/A* on amino acid types, to get a ranked list of sequences.

2. Then, for each sequence, do DEE/A* to get 10 best fleximer states.

3. Then, for each fleximer state, do DEE/A* to get 1 best rotamer state.

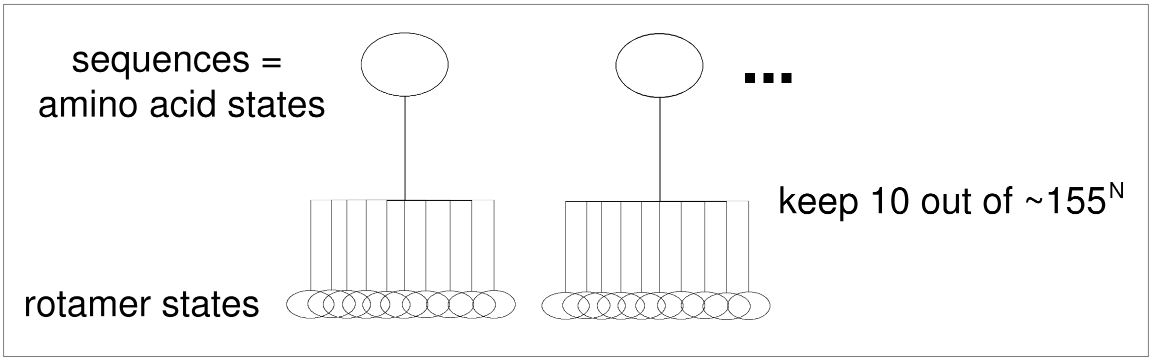

Figure 6-6: Naive Two-Stage Method: Amino Acids to Rotamers

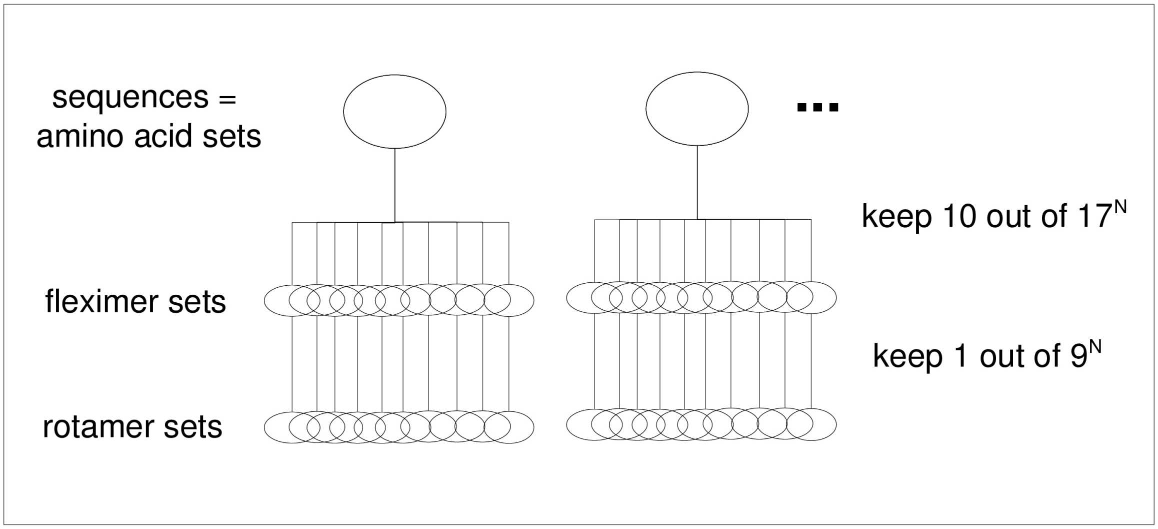

Figure 6-7: Three-Stage Method: Amino Acids to Fleximers to Rotamers

The second step is the slowest step, but it is feasible at 20 seconds per sequence. Despite the fact that the low- and high-resolution energy functions can disagree greatly about which sequences they favor, we have found that keeping only 10 structures per sequence has worked well, presumably because (1) the choice of rotamers for a given sequence is most constrained by the need to avoid van der Waals clashes, and (2) the low- and high-resolution electrostatic energy functions agree more about close-range interactions.

6.3.4 Test Systems

Barnase and Barstar

The proteins barnase and barstar bind extremely tightly (dissociation constant $K_d \approx 10^{-14}$M) [87, 88]. Barstar has been shown by Lee and Tidor [89] to be electrostatically optimized for tight binding to barnase when their shapes are held fixed but barstar partial atomic charges are allowed to vary. That study found a set of seven side chains that are especially important for binding, in that the binding free energy is particularly sensitive to the charge for those seven residues. Interestingly, the optimum side chain charges of all 7 match the actual wild-type charges. The side chains in this set of seven, which we call the “Lee 7” residues (Asn33, Asp35, Trp38, Asp39, Thr42, Val73, Glu76), make up about half of barstar’s binding interface, as shown in Figure 6-8.

We chose to apply our ligand design method to the “Lee 7” barstar residues, primarily as a validation of the method. Since barnase and wild-type barstar bind so tightly, we expect a good design method to predict, out of all possible sequences and conformations, that the wild-type sequence and conformation will be among the very best. Our method does successfully make this prediction, in contrast with other approaches. Surprisingly, it also suggests a few mutations (Val 73 → Gln or His) predicted to make binding even tighter. Before redesigning the “Lee 7” residues, we first redesigned two smaller systems in order to develop and verify our method’s use of a hierarchy of three energy functions. The first smaller system we redesigned is a subset of the “Lee 7” barstar residues: three

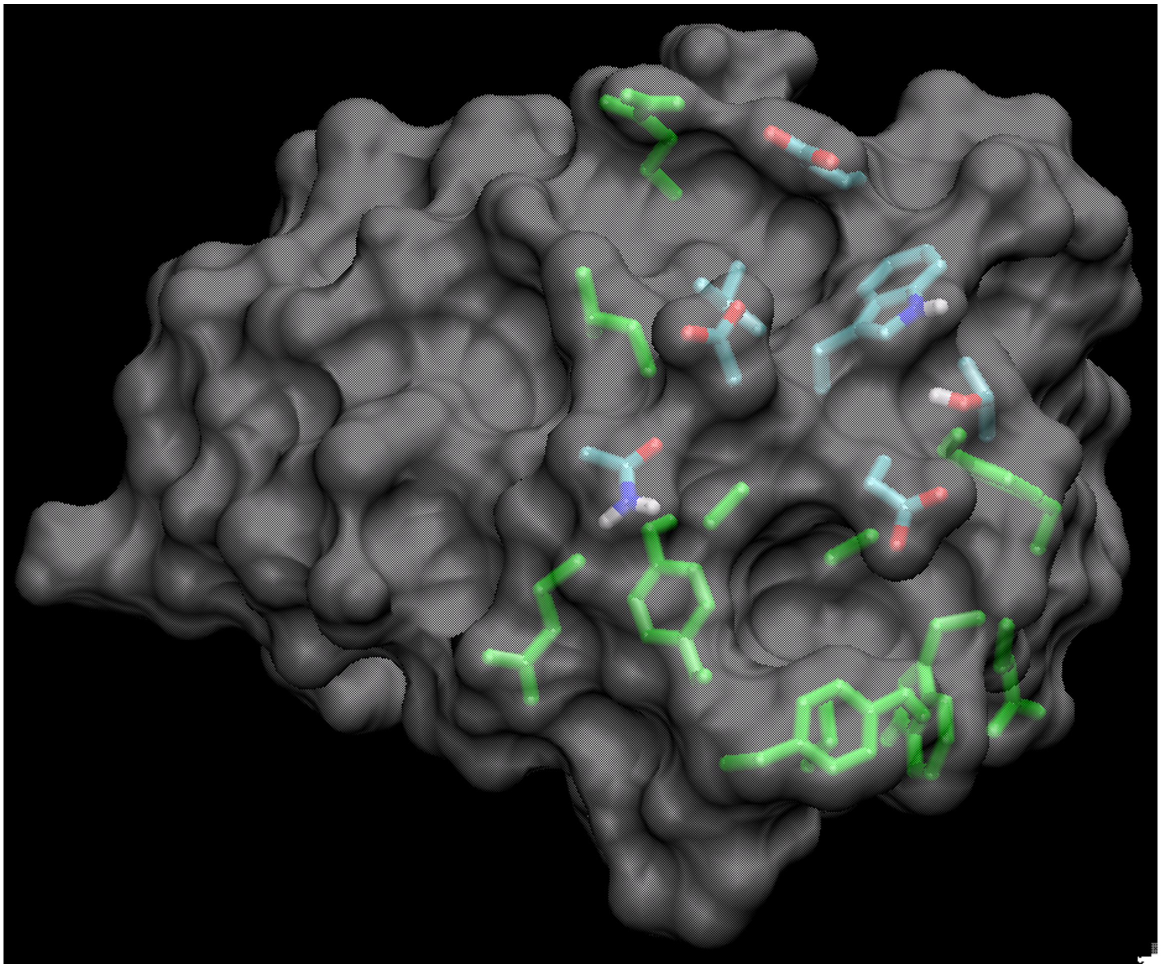

Figure 6-8: The “Lee 7” residues of barstar (Asn33, Asp35, Trp38, Asp39, Thr42, Val73, Glu76). All barstar side chains that bury solvent-accessible surface area (SASA) upon binding to barnase, shown as licorice, can be seen through the translucent gray surface of barstar. The “Lee 7” side chains are shown as element-colored licorice. The other barstar side chains that bury solvent-accessible surface area upon binding to barnase are shown as green licorice. Barnase is not shown; it would be in front of, and to the right of, barstar in this view. (Figures 6-8, 6-21, 6-23, 6-32, and 6-33 were made with the molecular graphics program VMD [36].)

barstar residues in the center of the binding interface, Asp35, Trp38, Val73, which we will call the “center 3” residues. The second system we redesigned is a set of three residues on the glycoprotein gp41, which we will introduce in the next section.

In redesigning barstar side chains on the wild-type backbone, one faces an interesting stability problem. The experimental free energy of folding for barstar in water is -5.28 kcal mol^{-1}, and slightly more stable, -5.88 kcal mol^{-1}, in 300 mM NaCl [34]. Proteins typically have folding free energies in the range -5 to -15 kcal mol^{-1}; since barstar is on the less stable side of that range, care must be taken to maintain stability in the unbound state for any redesigned barstar while still optimizing binding affinity. In the current scheme, this is achieved by effectively carrying out a constrained optimization on the computed binding free energy with the constraint being that the computed folding free energy be no more than one kcal mol^{-1} worse than wild type. One source of the marginal stability observed for barstar appears to be the concentration of negatively charged side chains at the binding interface that complement a positive patch on barnase. The enhanced stability in higher ionic strength is consistent with this view through ionic shielding of the charges.

We started with the 2.0-Å X-ray crystal structure (PDB entry 1BRS) [33] of the barnase/barstar complex. Twelve interfacial water molecules from the crystal structure have well-defined positions in the crystal structure. Of course, when barnase is unbound, these water molecules could have less well-defined positions, but whatever effect this has on the free energy of binding is a constant for all redesigned structures of barstar, and therefore we may safely treat the water molecules as a rigid part of barnase for purposes of redesigning barstar. In Appendix A, we develop a method to give conformational flexibility, and the option of removal, to interfacial water molecules by treating them just like protein side chains in a DEE/A* search.

Selected protein segments A and D from the X-ray crystal structures were used. Some atoms with missing density in the crystal structure are omitted; they are far from the binding interface. Polar hydrogens were built onto the crystal structure using the HBUILD facility [90] in CHARMM [11]. We use the PARAM19 partial atomic charge and atomic radius (van der Waals radius) parameters, including in the FDPB and ACE models [10].

gp41

Our second test system is part of the protease-resistant core of the HIV-1 glycoprotein gp41, the 3 inner helices (ABC) binding to one of the outer helices (D) (pictured in Figure 4-8 of Chapter 4) [64]. This rigid binding system is a simple model of the assumed last stage of folding, in which the outer helices dock onto the inner helices.

We started with the 2.0-Å X-ray crystal structure 1AIK [64]. Polar hydrogens were built onto the crystal structure using the HBUILD facility [90] in CHARMM [11]. We use the PARAM19 partial atomic charge and atomic radius parameters, including in the FDPB and ACE models [10].

6.3.5 Energy Functions

Self and Pair Terms

The dead-end elimination and A* algorithms rely on having an energy function that is pairwise additive with respect to the rotamers; i.e. the energy function has no three-body terms. Rather, it can be expressed as a sum of terms that depend on single rotamers, terms that depend on pairs of rotamers, and constant terms, as shown in Equation 6.1. Some of the energy terms that one would like to use are not pairwise additive, including a term to account for the hydrophobic effect, or more sophisticated treatments of solvation and electrostatics. The other terms — covalent, van der Waals, and Coulombic electrostatics — are pairwise additive with respect to the rotamers because they are also pairwise additive with respect to the atoms. That is, these energy terms can be expressed as a sum of terms which depend only on each single atom, and terms which depend only on each pair of atoms.

For a pairwise additive energy term, the self-energy of all the fixed atoms is a constant for all mutant conformations. Therefore, since we only care about the relative energies of the conformations, such constant terms are omitted from the calculated energies.

Covalent, van der Waals, Electrostatic, and Hydrophobic Terms

We calculate the free energy as a sum of covalent, van der Waals, electrostatic, and hydrophobic terms:

$$G = G_{\mathrm{cov}} + G_{\mathrm{vdW}} + G_{\mathrm{es}} + G_{\mathrm{hydr}} \tag{6.18}$$

We have not included a term for side chain entropy: a side chain dihedral angle that can occupy three minimum-energy positions in the unfolded state but is sterically constrained upon folding or binding could incur a free energy penalty of up to $k_{\mathrm{B}}T \ln 3$ [91]. Inclusion of this term would require a method of assessing the rotational freedom of each side chain or each dihedral angle. The free energy of translational and rotational entropy can be neglected because it depends only on the overall size of the molecules, so it will not vary significantly among our mutant conformations. We assume that vibrational and electronic free energy terms are not affected by folding or binding. Our model assumes rigid binding, so we are not considering the possibility of conformational change upon binding. This is justified by the fact that our test system (barnase and barstar) and many biomolecular complexes have been found to bind nearly rigidly, so variants of such a system that are designed to have a more favorable rigid binding energy than the wild type will surely bind rigidly as well.

Covalent Terms

The covalent free energy is a the sum of bond, angle, dihedral, and improper dihedral terms, with the commonly-used CHARMM PARAM19 parameter set [10]. This term accounts for intramolecular strain. The covalent free energy is pairwise additive.

$$G_{\mathrm{cov}} = G_{\mathrm{bond}} + G_{\mathrm{angl}} + G_{\mathrm{dihe}} + G_{\mathrm{impr}} \tag{6.19}$$

van der Waals Term

We calculate the van der Waals energy term using the standard Lennard–Jones 6-12 potential. The van der Waals term is omitted between atom pairs which are bonded (a "1-2" pair) or connected via 2 bonds ("1-3"), as intended for the CHARMM PARAM19 parameter set, using the `NBXMOD 5` CHARMM energy parameter.

The CHARMM PARAM19 parameter set specifies atomic van der Waals radii and van der Waals energy well depths for all atom types. The van der Waals term is sharp; i.e., putting atoms just a little too close to each other can make the van der Waals energy very unfavorable. In Section 6.2.6, we described how we use fleximers so that our rotamers can sample the conformations more finely. Another measure we take to address this problem of coarse sampling with a sharp energy function is to scale all of the atomic van der Waals radii by 90% . Any possible conformation of mobile side chains that is a relatively compact or snug fit would require rather tightly constrained side chain geometry. If the rotamer library is not fine enough, the library rotamers which are most similar to such a "possible conformation" may have atoms that "bump" or "clash", and therefore have a very unfavorable van der Waals energy. This is an unfortunate consequence of using a discrete rotamer library. We compensate for this by scaling all van der Waals radii by 90%, so that conformations that would have slight van der Waals clashes are not penalized for it, because in general there exists a slightly different conformation that would not have a van der Waals clash, but which is not in the rotamer library.

Specifically, when DEE/A* is run to give the "Lee 7" barstar residues conformational freedom, but with the wild-type amino acid types fixed, the conformation most similar to the crystal structure has van der Waals clashes involving atoms in 2 of the 7 residues, Asp39 and Glu76, penalizing it by about 18 kcal mol^{-1}. Varying the factor by which the van der Waals radii are scaled, we found that 0.9 was low enough to relieve the clashes in this particular case. Dahiyat and Mayo [92] used DEE to repack hydrophobic amino acids into a protein core; trying several van der Waals scale factors with DEE, they found that a value of 0.9 resulted in "a well-packed native-like protein".

The van der Waals term is pairwise additive. Therefore, the van der Waals interactions between all pairs of fixed atoms are not calculated, because they are a constant for all mutant conformations.

Electrostatic Term

The electrostatic free energy is inherently non-pairwise-additive: the presence of every solute atom, even an uncharged atom, screens the interaction between every other pair of charged atoms. To take advantage of the eliminating power of dead-end elimination and A*searches, we use an approximate electrostatic energy term that is pairwise additive but not entirely accurate. In subsequent calculations, when the search space is much smaller, more accurate treatments of electrostatics and solvation are incorporated which are computationally more demanding.

We will now introduce three electrostatic energy functions of successively higher accuracy and computational cost. In Section 6.4.2, we will discuss how these successive levels of accuracy are used to narrow down vast numbers of structures to the few with the best binding and folding energies calculated with our most accurate energy function.

Low-resolution Electrostatics

We desire an approximation to the electrostatic free energy which is pairwise additive. Perhaps the simplest such approximation would be the Coulombic energy. The Coulombic energy is simply the sum of the interactions of all atom pairs from Coulomb's law, scaled with an effective dielectric constant ϵ:

$$G_{\text{ES, Coul}} = \sum_{i \neq j} \frac{q_i q_j}{\epsilon \; r_{ij}} \tag{6.20}$$

The conversion factor $1 = 332.0716$ kcal mol^{-1} Å e^{-2}, where e is the magnitude of the electron charge, allows us to use our preferred units.

The distance-dependent Coulombic energy is usually a better approximation; it uses an effective dielectric constant that depends linearly on distance, as a rough approximation of the trend that the interactions of distant atom pairs are more screened by the high-dielectric solvent than near atom pairs:

$$G_{\text{ES, dd}} = \sum_{i \neq j} \frac{q_i q_j}{\epsilon_{\text{dd}}(r_{ij}) r_{ij}} \tag{6.21}$$

where

$$\epsilon_{\mathrm{dd}}(r) = \epsilon \frac{r}{1\mathring{A}} \tag{6.22}$$

For low-resolution electrostatic energy, we use the distance-dependent Coulombic energy with $\epsilon = 4$, a choice which we shall call “4r dielectric”. This is a common choice when using CHARMM PARAM19 parameters. The distance-dependent Coulombic electrostatic term is omitted between atom pairs which are bonded (a “1-2” pair) or connected via 2 bonds (“1-3”), and is scaled by 0.4 for atom pairs which are connected via 3 bonds (a “1-4” pair), as intended for the CHARMM PARAM19 parameter set. This set of rules for omitting and reducing the electrostatic term is set with the `NBXMOD 5 E14FAC 0.4` CHARMM energy parameters.

Medium-resolution Electrostatics

The medium-resolution electrostatic free energy function is the ACE analytical approximation of Schaefer and Karplus [6], modified as described in Chapter 4 to reduce its error. The CHARMM PARAM19 topology and parameter sets are used. We used values of the ACE parameters, $\alpha = 1.2$ and effective atom volumes V as given in reference [7] and distributed with CHARMM version 27a2; they were obtained by minimizing the solute volume fluctuations for a set of 12 protein structures (i.e. making the sum of the Gaussian distributions representing each atom’s desolvation as similar as possible to the step function with value 1 inside the solvent-accessible surface and 0 outside) [55]. We extended the volumes to similar atom types as described in Chapter 4. Internal and external dielectric constants of $\epsilon_{\mathrm{i}} = 4$ and $\epsilon_{\mathrm{s}} = 80$ were used. No salt was used, although it would be trivial to include salt effects in the future. The medium-resolution electrostatic energy is not pairwise additive.

High-resolution Electrostatics

The high-resolution electrostatic energy is calculated using FDPB, calculated by the program DELPHI [3, 4, 5], as described in Chapter 3. Internal and external dielectric constants of $\epsilon_i = 4$ and $\epsilon_s = 80$ were used. As with medium-resolution electrostatics, no salt was used, although it would be trivial to include salt effects in the future. The high-resolution electrostatic energy is not pairwise additive.

For the purposes of the current work, the FDPB continuum electrostatic treatment is the best model considered. If another model is preferred, it can be substituted with the overall scheme remaining essentially unchanged. For the work reported here, $\Delta\Delta G^{\text{bind}}_{\text{ES, FDPB}}$ values converged at a relatively coarse grid (65x65x65 grid points, or 0.8 Å/grid), which was used throughout.

We chose the finite-difference grid spacing by finding the coarsest grid spacing at which $\Delta\Delta G^{\text{bind}}_{\text{ES, FDPB}}$, the differences of two states' electrostatic binding free energies, is converged. This was not done exhaustively; we tried cubes with 33, 65, or 129 grid points on a side, and calculated the FDPB electrostatic binding free energy for 4629 structures found by DEE/A* mutating the "center 3" barstar residues (Asp35, Trp38, Val73). In Figure 6-9, we see that plotting the $\Delta G^{\text{bind}}_{\text{ES, FDPB}}$ values for 33 vs. 129 grids/side gives a standard deviation of 0.840 kcal mol^{-1} from the best-fit line, which does not have slope 1, so we judge that the values of $\Delta\Delta G^{\text{bind}}_{\text{ES, FDPB}}$ are not converged at 33 grids/side. The figure also shows a similar comparison for 65 vs. 129 grids/side; the standard deviation is 0.304 kcal mol^{-1} from the best-fit line, which has slope 1.01, so we judge that the values of $\Delta\Delta G^{\text{bind}}_{\text{ES, FDPB}}$ are converged at 65 grids/side. Therefore, we use a cube with 65x65x65 grid points.

To ensure that the FDPB program DELPHI calculates all structures on the exact same grid, we placed 2 dummy atoms (zero charge, zero radius) at opposite corners of a cube containing all possible library rotamers at all mobile residues. The grid spacing was about 0.8 Å/grid.

Hydrophobic Term

The hydrophobic term is taken to be proportional to the solvent-accessible surface area (SASA) of the solute molecule(s).

$$G_{\text{hydr}} = (\gamma)(SASA) \tag{6.23}$$

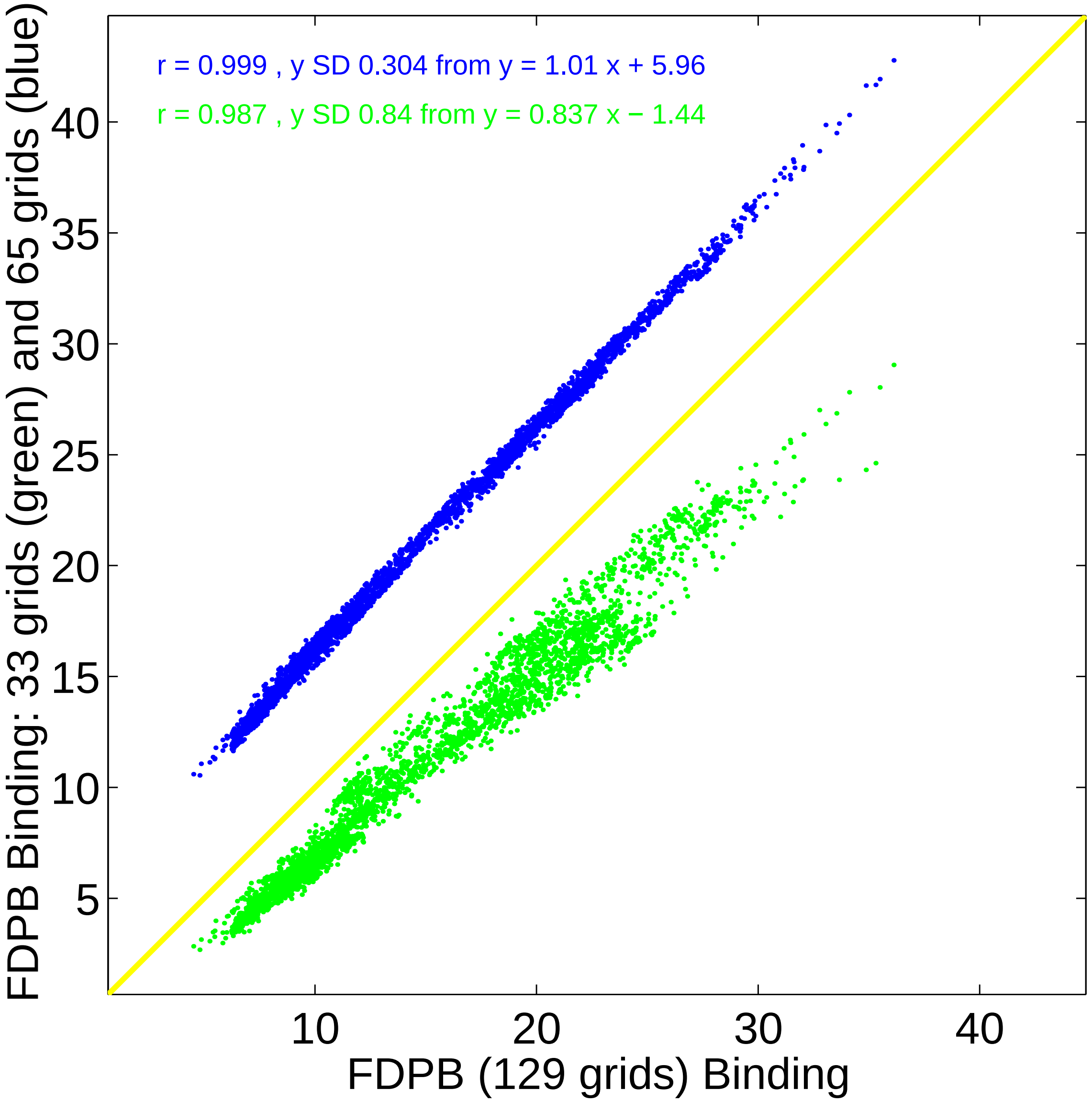

Figure 6-9: In green, $\Delta G^{\text{bind}}_{\text{ES, FDPB}}$ for 33 grids/side (1.6 Å/grid) vs. 129 grids/side (0.4 Å/grid) shows a standard deviation of 0.840 kcal mol^{-1} from the best-fit line, which does not have slope 1. In blue, $\Delta G^{\text{bind}}_{\text{ES, FDPB}}$ for 65 grids/side (0.8 Å/grid) vs. 129 grids/side (0.4 Å/grid) shows a standard deviation of 0.304 kcal mol^{-1} from the best-fit line, which has slope 1.01 . This demonstrates convergence of FDPB electrostatic binding free energy differences $\Delta\Delta G^{\text{bind}}_{\text{ES, FDPB}}$ at 65 grids/side, but not at 33 grids/side. The points of each color are for the same 4629 mutant structures from DEE/A* on the "center 3" residues (Asp35, Trp38, Val73). Both dimensions are in kcal mol^{-1}.

We used the value $\gamma = 5.0$ cal mol^{-1} Å^{-2}, as used by Sitkoff *et al.* [93], except where stated otherwise. We used CHARMM [10] to analytically calculate the solvent-accessible surface area of the bound and unbound states of every structure, and of the unfolded model compound for each amino acid type. A probe radius of 1.4 Å, representing a water molecule, was used. (The program `msms`, with the `--all_components` option, can usually get identical results, but we found it to be somewhat unreliable, crashing or infinitely looping for some combinations of conformation and probe radius.)

The hydrophobic term is not pairwise additive, so it must be omitted from the low-resolution energy function for use with DEE/A*; but it is included in the medium- and high-resolution energy functions. It should be noted that approximate solvent exposure calculations that are pairwise additive have been constructed and could be added here [94].

The factor γ can be considered as the sum of two terms, the cavity term and the solvent van der Waals term, which are both taken to be proportional to the solvent-accessible surface area:

$$\gamma = \gamma_{\text{cav}} + \gamma_{\text{vdW}} \tag{6.24}$$

The cost of creating an uncharged, hydrophobic cavity in the solvent is represented by the cavity term $\gamma_{\text{cav}} \approx +47$ cal mol^{-1} Å^{-2}, (for example, but this value is far from certain) [12]. The cavity term corresponds most closely to experiments involving partitioning between oil and water. The van der Waals interaction of the solute atoms with the solvent are represented by the solvent van der Waals term $\gamma_{\text{vdW}} \approx -40$ cal mol^{-1} Å^{-2}) [12]. Of course, a solute molecule immersed in either water or oil will have about the same van der Waals interaction with either. But, since there are no explicit solvent atoms in our model, we use a favorable term proportional to the solvent-accessible surface area, $\gamma_{\text{vdW}} \cdot (SASA)$, for the van der Waals between the solute and solvent. The values cited above for the cavity and solvent van der Waals terms combine to give +7 cal mol^{-1} Å^{-2}, very approximately, for γ.

There is still confusion and debate on what terms are actually causing the experimentally observed hydrophobic effect, and on its role in binding and other processes [12]. Some assume the cavity term to be dominated by the entropy change of creating a cavity in the solvent with the shape of the solute [13]. Others [14] explain it in terms of hydrogen bonding with water. Water molecules next to a hydrophobic group have higher free energy than they would if the hydrophobic group were replaced by more water, with which the waters could make a good hydrogen bond network. So part of the hydrophobic term is for the solvent's hydrogen bonds.

We used $\gamma = 5$ cal mol^{-1} Å^{-2}, as used by Sitkoff *et al.* [93], for most of the results in this work. But the best value of γ is still not clear. Some fits to experimental data get values around 5 cal mol^{-1} Å^{-2}, such as 7.2 cal mol^{-1} Å^{-2} from the non-polar solvation free energy of neutral small molecules [46], $\approx$ 8 cal mol^{-1} Å^{-2} from the solvation free energies of small apolar molecules [95], and the 7 cal mol^{-1} Å^{-2} cited above [12]. Other experiments obtain values around 25 cal mol^{-1} Å^{-2}, the "canonical value" from the transfer of alkanes from alkane solvent to water [15], which we believe is better understood as measuring the cavity term. For example, 22.8 ± 0.8 cal mol^{-1} Å^{-2} from partitioning of a family of host-guest pentapeptides (Ac-WLXLL) between water and n-octanol [16] and 34 cal mol^{-1} Å^{-2} from oil-water partitioning, based on water-cyclohexane transfers of alkanes [17]. The maximum reasonable value for γ would be 75 cal mol^{-1} Å^{-2}, the macroscopic surface tension of water.

In Section 6.4.1, we assess the effects of γ values of 5, 22.8, and 75 cal mol^{-1} Å^{-2} on the results of a protein redesign.

Definitions of low-, medium-, High-resolution Energy Functions

The low-resolution energy function is pairwise additive so that it can be used with DEE/A*.

$$G_{\text{low}} \equiv G_{\text{cov}} + G_{\text{vdW}} + G_{\text{ES, dd}} \tag{6.25}$$

The medium- and high-resolution energies are not pairwise additive because of their electrostatic and hydrophobic terms. The electrostatic terms are the computationally expensive part (FDPB being far more expensive than ACE).

$$G_{\text{med}} \equiv G_{\text{cov}} + G_{\text{vdW}} + G_{\text{ES, ACE}} + G_{\text{hydr}} \tag{6.26}$$

$$G_{\text{high}} \equiv G_{\text{cov}} + G_{\text{vdW}} + G_{\text{ES, FDPB}} + G_{\text{hydr}} \tag{6.27}$$

Note that the van der Waals term G_{vdW} uses van der Waals radii scaled by 90%; the "4r dielectric" distance-dependent Coulombic electrostatic term $G_{\text{ES, dd}}$ uses $\epsilon = 4$, and the hydrophobic term uses $\gamma = 5$ cal mol^{-1} Å^2 unless otherwise specified.

Binding and Folding Terms

In the preceding sections, we have discussed the terms of the total free energy G. Previously published design methods have only used a total free energy in the bound state, G^{bound}, to rank structures [85, 86]. In our method, however, we always calculate terms of the binding free energy ΔG^{bind} and the folding free energy ΔG^{fold}.

$$\Delta G^{\text{bind}} \equiv G^{\text{bound}} - G^{\text{unbound}} \tag{6.28}$$

$$\Delta G^{\text{fold}} \equiv G^{\text{unbound}} - G^{\text{unfolded}} \tag{6.29}$$

Minimizing G^{bound} will not necessarily minimize ΔG^{bind} or ΔG^{fold}. By calculating both ΔG^{bind} and ΔG^{fold} correctly, we have predictions not only of the binding free energy, but of the stability of the molecules when unbound. This is very valuable because one need not bother to synthesize a ligand predicted to be unstable.

Binding Free Energy

We must use one energy function with which to rank rotamer states in DEE/A*, and we choose it to be the binding free energy $\Delta G^{\text{bind}}_{\text{low}}$. Our procedure is capable of ranking on the bound, binding, or folding free energies, or any linear combinations thereof. Later in the design method, we use the more accurate functions $\Delta G_{\text{med}}^{\text{bind}}$ and $\Delta G_{\text{high}}^{\text{bind}}$ to rank the promising rotamer states.

We assume rigid binding; that is, our model of the unbound state separates the two binding partners by rigidly translating one of the binding partners away from the other (far enough so they have no interaction). Therefore, the covalent term $\Delta G_{\text{cov}}^{\text{bind}}$ cancels. For the non-pairwise-additive energy terms $G_{\text{ES, ACE}}$, $G_{\text{ES, FDPB}}$, and G_{hydr}, the binding free energy terms must be calculated for each conformation, but the calculation is a straightforward subtraction of the bound state free energy term and the unbound state free energy term.

Self and Pair Terms of the Binding Free Energy for Pairwise Additive Terms

The pairwise additive energy terms G_{vdW} and $G_{\text{ES, dd}}$ break up into self and pair terms which can be further simplified by the assumption of rigid binding. If all mobile residues are on the same binding partner, then since the low-resolution energy is pairwise additive, the binding pair terms $\Delta G_{\text{pair}}^{\text{bind}}$ are all zero, and DEE/A* is actually more powerful than is required to rank the rotamer states by binding free energy! Recall that the low-resolution energy is calculated as a sum of constant, self, and pair terms, involving 0, 1, and 2 mobile residues, respectively (Equation 6.1). The constant subterms of $G_{\text{vdW}}^{\text{bind}}$ and $G_{\text{ES, dd}}^{\text{bind}}$ — the van der Waals and electrostatic interaction of all the fixed atoms of one binding partner with all the fixed atoms of the other — is not calculated because it would not alter the relative ranking of the mutant conformations. The binding self term for mobile residue i in binding partner "A" involves only the interaction of residue i with the fixed atoms of the other binding partner "B":

$$\Delta G_{\text{low}}^{\text{bind}}(i) = G_{\text{vdW}}(i : B) + G_{\text{ES, dd}}(i : B) \tag{6.30}$$

$$\Delta G_{\text{low}}^{\text{bind}}(i,j) = \begin{cases} G_{\text{vdW}}(i:j) + G_{\text{ES, dd}}(i:j) & \text{if } i,j \text{ are on opposite binding partners} \\ 0 & \text{if } i,j \text{ are on the same binding partner} \end{cases} \tag{6.31}$$

Folding Free Energy

We use the low-resolution folding free energy $\Delta G_{\text{low}}^{\text{fold}}$ in DEE/A* to discard rotamers and rotamer pairs with very high contributions to the folding free energy (above a cutoff of 25 kcal mol^{-1}). Later in the design method, we use the more accurate functions $\Delta G_{\text{med}}^{\text{fold}}$ and $\Delta G_{\text{high}}^{\text{fold}}$ to assess the stability of every promising structure.

Unfolded State Model

The unfolded state is modelled as a collection of model compounds totally separated in solvent. Since only the mobile residues vary amongst all rotamer states considered, and we discard constant energy terms, we only include the mobile residues in the unfolded state. For example, if a rotamer state has Trp, Asp, and Asp side chains at the 3 mobile residues, then its unfolded state free energy is the sum of the Trp unfolded model compound free energy, plus twice the Asp unfolded model compound free energy.

The unfolded model compound for each amino acid type "R" is created as an N-acetyl-"R" methylamide (CH_3-(CO)-(NH)-(C_α"R")-(CO)-(NH)-CH_3), with the backbone held in an extended conformation, and the side chain atoms beyond C_β minimized to completion using the low-resolution energy function G_{low} but with full van der Waals radii.

To demonstrate why it is vital to compare different mutants using the difference $\Delta G^{\text{fold}} \equiv G^{\text{unbound}} - G^{\text{unfolded}}$ rather than simply G^{bound} or G^{unbound}, Table 6.2 shows terms of $G_{\text{cov}}^{\text{unfolded}}$ for each amino acid type. Some amino acids, namely tryptophan and histidine, have significant intra-side chain strain energy with this parameter set. All tryptophan and histidine side chain conformations in the rotamer library have similar strain energy. So, when we take the differences $\Delta G^{\text{bind}} \equiv G^{\text{bound}} - G^{\text{unbound}}$ or $\Delta G^{\text{fold}} \equiv G^{\text{unbound}} - G^{\text{unfolded}}$, such terms cancel out. But if one naively ranked sequences by G^{bound}, tryptophan and histidine side chains would be unfairly penalized.

Table 6.2: Covalent Energy Terms of Unfolded Model Compounds, broken into intra-side chain and side chain – backbone interaction terms. (Note: In the CHARMM PARAM19 topology library, "His" is histidine protonated on $N_{\delta 1}$; whereas "Hsd" is histidine protonated on $N_{\epsilon 2}$.)

amino acid	intra-side chain	side chain – backbone interaction
Ala	0.00	0.00
Arg	0.36	1.68
Asn	0.09	1.95
Asp	0.01	1.08
Cys	0.00	0.27
Gln	0.10	1.77
Glu	0.06	1.55
Gly	0.00	0.00
His	1.76	1.58
Hsd	5.62	1.09
Ile	0.28	1.25
Leu	0.13	0.95
Lys	0.01	1.68
Met	0.03	1.63
Phe	0.17	1.16
Ser	0.01	0.20
Thr	0.05	0.35
Trp	12.98	1.04
Tyr	0.21	1.16
Val	0.03	0.51

The way that the unfolded model compounds are built does not have a very large effect on the free energies. We tried using side chains only (the C_β atom and beyond) as the model compounds, and minimizing as above. Only a few of the intra-side chain energies changed significantly, compared to the N-acetyl-"R" methylamide model compounds: Gln by $+0.69$ kcal mol^{-1}, Ile by -0.33 kcal mol^{-1}, and all other side chains by less than $|0.12 \text{ kcal mol}^{-1}|$.

The minimized N-acetyl-"R" methylamide model compounds look reasonable to the human eye, but we wanted to make sure that the minimization is not being led astray by the strength of the "4r dielectric" distance-dependent Coulombic electrostatics, which is probably too strong for such a small molecule in solvent. So we did minimizations using only covalent and van der Waals terms, with full van der Waals radii. Only a few of the intra-side chain energies changed significantly, compared to the minimization which also included $G_{\mathrm{ES,\ dd}}$: Asp by +0.44 kcal mol^{-1}, Hsd by +0.56 kcal mol^{-1}, and all other side chains by less than $|0.1\ \ \mathrm{kcal\ mol}^{-1}|$.

Folding Thermodynamic Cycle

Figure 6-10 illustrates the definition of the folding free energy. The actual unfolded state is a poorly-defined ensemble of conformations. In general, the binding partners may both have mobile residues, in which case we are calculating the folding free energy of both binding partners.

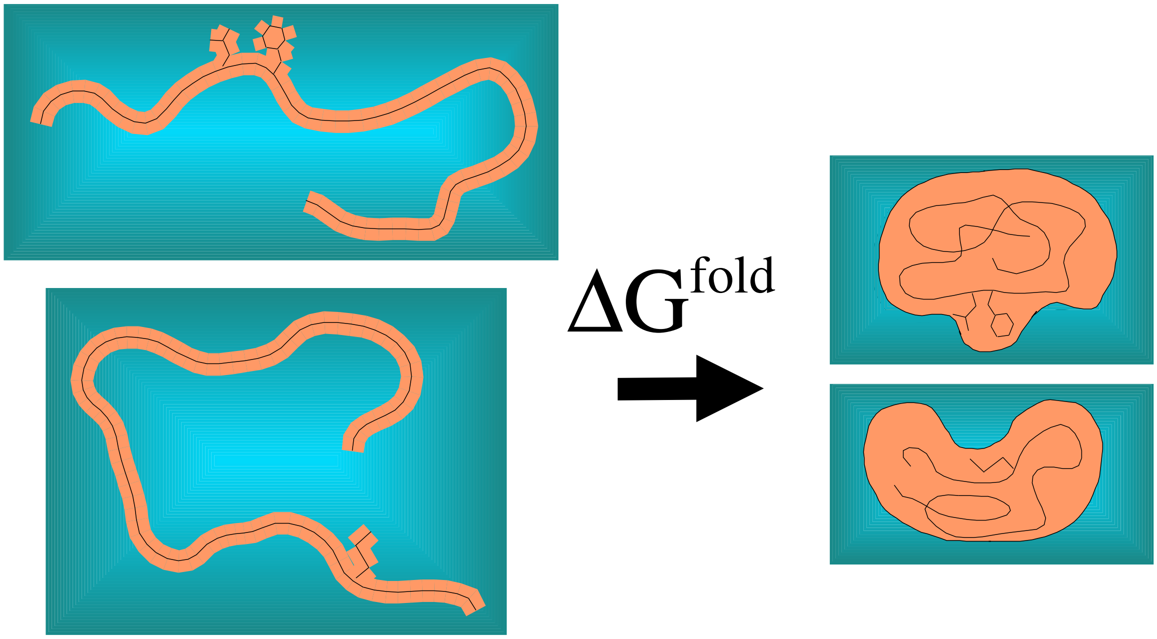

Figure 6-10: Folding Cartoon. Of course, the unfolded state is actually an ensemble of states. Blue represents water solvent ($\epsilon_{\mathrm{s}} = 80$); orange represents protein dielectric ($\epsilon_{\mathrm{i}} = 4$).

We assume that, to within a constant energy term, we can represent the unfolded state by a collection of model compounds for the mobile residues. Therefore, we use the thermodynamic cycle in Figure 6-11 for the folding of a rotamer state with N mobile residues. The thermodynamic cycle connects the unfolded state to the folded state by desolvating the unfolded state, bringing the atoms to their folded configuration, and then re-solvating.

$$\Delta G^{\text{fold}} = -\Delta G^{\text{solv}}(\text{unfolded}) + \Delta G^{\text{fold, desolvated}} + \Delta G^{\text{solv}}(\text{unbound}) \tag{6.32}$$

The cycle assumes that, for both the solvated and desolvated state free energies:

$$G(\text{unfolded}) \cong \sum_{i=1}^{N} G(\text{model compound } i) + \textit{constant} \tag{6.33}$$

All of the terms except the electrostatics are, to within a neglected constant, a straightforward free energy difference of the folded unbound system and the unfolded model compounds:

$$\begin{aligned} \Delta G^{\text{fold}}_{\text{cov}} \;=\; & +G_{\text{cov}}(\text{unbound}) \\ & - \textstyle\sum_{i=1}^{N} G_{\text{cov}}(\text{model compound } i) \\ & +\textit{constant} \end{aligned} \tag{6.34}$$

$$\begin{aligned} \Delta G^{\text{fold}}_{\text{vdW}} \;=\; & +G_{\text{vdW}}(\text{unbound}) \\ & - \textstyle\sum_{i=1}^{N} G_{\text{vdW}}(\text{model compound } i) \\ & +\textit{constant} \end{aligned} \tag{6.35}$$

$$\begin{aligned} \Delta G^{\text{fold}}_{\text{hydr}} \;=\; & +G_{\text{hydr}}(\text{unbound}) \\ & - \textstyle\sum_{i=1}^{N} G_{\text{hydr}}(\text{model compound } i) \\ & +\textit{constant} \end{aligned} \tag{6.36}$$

The medium-resolution electrostatic (ACE) folding term can be calculated with ACE calculations to get the $-\Delta G^{\text{solv}}$(unfolded) and ΔG^{solv}(unbound) solvation terms, plus

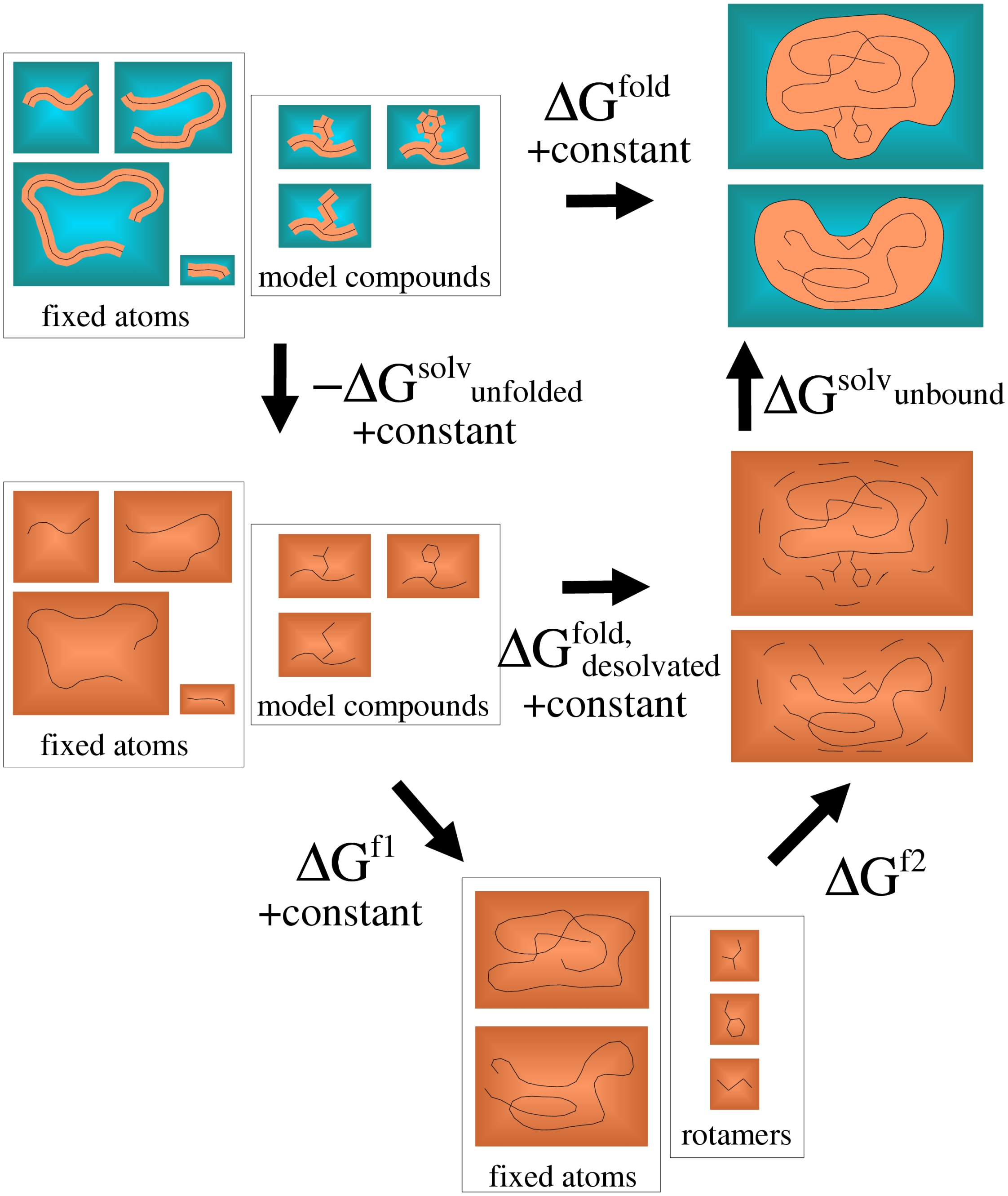

Figure 6-11: Folding Thermodynamic Cycle. Our unfolded state model assumes that, for both the solvated and desolvated state free energies, $G(\text{unfolded}) = \sum_{i=1}^{N} G\ (\text{model compound } i) + \textit{constant}$. Therefore, the solvated and desolvated unfolded states are replaced in this representation by the set of unfolded model compounds, plus the fixed atoms (whose energy is constant over all mutants). Blue represents water solvent ($\epsilon_{\text{s}} = 80$); orange represents protein dielectric ($\epsilon_{\text{i}} = 4$). The fixed side chains are not shown in this cartoon, but they are always attached to the backbones.

Coulombic electrostatics with $\epsilon_{\mathrm{i}} = \epsilon_{\mathrm{s}} = 4$ to get the $\Delta G^{\text{fold, desolvated}}$ term:

$$\begin{aligned}
\Delta G^{\text{fold}}_{\text{ES, ACE}} \quad = \quad & -\textstyle\sum_{i=1}^{N} \Delta G^{\text{solv}}_{\text{ES, ACE}}(\text{model compound } i) \\
& +\Delta G^{\text{solv}}_{\text{ES, ACE}}(\text{folded}) \\
& +G_{\text{ES, Coul}}(\text{folded}) \\
& -\textstyle\sum_{i=1}^{N} G_{\text{ES, Coul}}(\text{model compound } i) \\
& +\mathit{constant}
\end{aligned} \tag{6.37}$$

Note that the Coulombic terms must include all atom pairs (by setting the `NBXMOD 0 E14FAC 1` CHARMM energy parameters), as FDPB and ACE do, as opposed to the `NBXMOD 5 E14FAC 0.4` behavior described above for the van der Waals and distance-dependent Coulombic terms.

For the FDPB electrostatic term only, it is convenient to split up

$$\Delta G^{\text{fold, desolvated}} = \Delta G^{\text{f1}} + \Delta G^{\text{f2}} + \mathit{constant} \tag{6.38}$$

as shown in Figure 6-11. The term ΔG^{f1} is the free energy of removing the mobile side chains from their model compound backbones and changing their conformation to that in the folded state, all while desolvated. This can be calculated simply with Coulombic electrostatics with $\epsilon_{\mathrm{i}} = \epsilon_{\mathrm{s}} = 4$.

$$\begin{aligned}
\Delta G^{\text{f1}} + \mathit{constant} \quad = \quad & +\textstyle\sum_{i=1}^{N} G^{\text{desolvated}}_{\text{ES, Coul}}(\text{rotamer } i) \\
& -\textstyle\sum_{i=1}^{N} G^{\text{desolvated}}_{\text{ES, Coul}}(\text{model compound } i) \\
& +\mathit{constant}
\end{aligned} \tag{6.39}$$

By "rotamer i", we mean the atoms of mobile side chain number i.

FDPB calculations can obtain the $-\Delta G^{\text{solv}}(\text{unfolded})$ solvation term:

$$-\Delta G^{\text{solv}}(\text{unfolded}) \quad = \quad -\textstyle\sum_{i=1}^{N} \Delta G^{\text{solvation}}_{\text{ES, FDPB}}(\text{model compound } i) \tag{6.40}$$

ΔG^{f2} is the unscreened interaction of the mobile side chains with the fixed atoms. FDPB calculations can obtain $\Delta G^{\mathrm{f2}} + \Delta G^{\mathrm{solv}}(\text{unbound})$:

$$\begin{aligned}\Delta G^{\mathrm{f2}} + \Delta G^{\mathrm{solv}}(\text{unbound}) \quad = \quad & +G_{\mathrm{ES}}^{\mathrm{solvated}}(\text{folded}) \\ & -G_{\mathrm{ES}}^{\mathrm{desolvated}}(\text{fixed atoms}) \\ & -\textstyle\sum_{i=1}^{N} G_{\mathrm{ES}}^{\mathrm{desolvated}}(\text{rotamer } i)\end{aligned} \tag{6.41}$$

Putting these terms together (we will simplify this expression below):

$$\begin{aligned}\Delta G_{\mathrm{ES,\ FDPB}}^{\mathrm{fold}} \quad = \quad & -\Delta G^{\mathrm{solv}}(\text{unfolded}) \\ & +\Delta G^{\mathrm{f1}} + \mathit{constant} \\ & +\Delta G^{\mathrm{f2}} + \Delta G^{\mathrm{solv}}(\text{unbound}) \\ \\ = \quad & -\textstyle\sum_{i=1}^{N} \Delta G_{\mathrm{ES,\ FDPB}}^{\mathrm{solvation}}(\text{model compound } i) \\ & +G_{\mathrm{ES}}^{\mathrm{solvated}}(\text{folded}) \\ & -G_{\mathrm{ES}}^{\mathrm{desolvated}}(\text{fixed atoms}) \\ & -\textstyle\sum_{i=1}^{N} G_{\mathrm{ES}}^{\mathrm{desolvated}}(\text{rotamer } i) \\ & +\textstyle\sum_{i=1}^{N} G_{\mathrm{ES,\ Coul}}^{\mathrm{desolvated}}(\text{rotamer } i) \\ & -\textstyle\sum_{i=1}^{N} G_{\mathrm{ES,\ Coul}}^{\mathrm{desolvated}}(\text{model compound } i) \\ & +\mathit{constant}\end{aligned} \tag{6.42}$$

$\Delta G_{\mathrm{ES,\ FDPB}}^{\mathrm{solvation}}(\text{model compound } i)$ is obtained, for each amino acid type used by any rotamer i, from FDPB; the free energy difference of the solvated state (internal, external dielectric constants $\epsilon_{\mathrm{i}} = 4$, $\epsilon_{\mathrm{s}} = 80$) and the desolvated state ($\epsilon_{\mathrm{i}} = 4$, $\epsilon_{\mathrm{s}} = 4$).

$G_{\mathrm{ES}}^{\mathrm{solvated}}(\text{folded})$ is obtained by FDPB ($\epsilon_{\mathrm{i}} = 4$, $\epsilon_{\mathrm{s}} = 80$). A FDPB result such as this one, which is not a difference of two runs with the charges in the exact same locations, contains the true electrostatic free energy, plus a fictitious "grid energy":

$$G_{\mathrm{ES,\ FDPB}}^{\mathrm{solvated}}(\text{folded}) = G_{\mathrm{ES}}^{\mathrm{solvated}}(\text{folded}) + G_{\mathrm{grid}}(\text{folded}) \tag{6.43}$$

The grid energy is decomposable into atomic contributions, each of which depends only on the atom's position, the dielectric constant at its location, and the placement of the finite-difference grid:

$$G_{\mathrm{grid}}(\mathrm{folded}) = G_{\mathrm{grid}}(\mathrm{fixed\ atoms}) + \sum_{i=1}^{N} G_{\mathrm{grid}}(\mathrm{rotamer}\ i) \tag{6.44}$$

Since we use the same finite-difference grid for all rotamer states, $G_{\mathrm{grid}}(\mathrm{fixed\ atoms})$ is a constant, and so is ignored. The $\sum_{i=1}^{N} G_{\mathrm{grid}}(\mathrm{rotamer}\ i)$ terms will be cancelled out by the same terms in the $G_{\mathrm{ES}}^{\mathrm{desolvated}}(\mathrm{rotamer}\ i)$ terms.

$G_{\mathrm{ES}}^{\mathrm{desolvated}}(\mathrm{fixed\ atoms})$ is a constant, and so is ignored.

$G_{\mathrm{ES}}^{\mathrm{desolvated}}(\mathrm{rotamer}\ i)$ is obtained by FDPB ($\epsilon_{\mathrm{i}} = 4$, $\epsilon_{\mathrm{s}} = 4$). The FDPB result also contains a grid energy which will cancel out in the end:

$$G_{\mathrm{ES,\ FDPB}}^{\mathrm{desolvated}}(\mathrm{rotamer}\ i) = G_{\mathrm{ES}}^{\mathrm{desolvated}}(\mathrm{rotamer}\ i) + G^{\mathrm{grid}}(\mathrm{rotamer}\ i) \tag{6.45}$$

$G_{\mathrm{ES,\ Coul}}^{\mathrm{desolvated}}(\mathrm{folded})$ and $G_{\mathrm{ES,\ Coul}}^{\mathrm{desolvated}}(\mathrm{model\ compound}\ i)$ are obtained by a simple calculation of Coulombic electrostatics, because $\epsilon_{\mathrm{i}} = \epsilon_{\mathrm{s}} = 4$ in all space. Note that this Coulombic term must include all atom pairs (by setting the `NBXMOD 0 E14FAC 1` CHARMM energy parameters), as FDPB and ACE do, as opposed to the `NBXMOD 5 E14FAC 0.4` behavior described above for the van der Waals and distance-dependent Coulombic terms.

Simplifying, we get:

$$\begin{aligned} \Delta G_{\mathrm{ES,\ FDPB}}^{\mathrm{fold}} = & -\textstyle\sum_{i=1}^{N} \Delta G_{\mathrm{ES,\ FDPB}}^{\mathrm{solvation}}(\mathrm{model\ compound}\ i) \\ & + G_{\mathrm{ES,\ FDPB}}^{\mathrm{solvated}}(\mathrm{folded}) \\ & -\textstyle\sum_{i=1}^{N} G_{\mathrm{ES,\ FDPB}}^{\mathrm{desolvated}}(\mathrm{rotamer}\ i) \\ & +\textstyle\sum_{i=1}^{N} G_{\mathrm{ES,\ Coul}}^{\mathrm{desolvated}}(\mathrm{rotamer}\ i) \\ & -\textstyle\sum_{i=1}^{N} G_{\mathrm{ES,\ Coul}}^{\mathrm{desolvated}}(\mathrm{model\ compound}\ i) \\ & + constant \end{aligned} \tag{6.46}$$

where every term is obtainable by either FDPB or Coulomb's law.

Self and Pair Terms of the Folding Free Energy for Pairwise Additive Terms

The pairwise additive terms $\Delta G^{\text{fold}}_{\text{cov}}$, $\Delta G^{\text{fold}}_{\text{vdW}}$, and $\Delta G^{\text{fold}}_{\text{ES, dd}}$ are calculated as sums of self and pair terms. The unbound self energy term for mobile residue i in binding partner "A" involves the interaction of residue i with itself and with the fixed atoms of its own binding partner "A". Likewise for the unfolded self energy term, except that the fixed atoms in the unfolded state consist of only the model compound backbone ("*bb*" in the equations).

$$\Delta G^{\text{fold}}_{\text{low}}(i) = G^{\text{unbound}}_{\text{low}}(i) + G^{\text{unbound}}_{\text{low}}(i : A) - (G^{\text{unfolded}}_{\text{low}}(i) + G^{\text{unfolded}}_{\text{low}}(i : bb)) \qquad (6.47)$$

The covalent term $\Delta G^{\text{fold}}_{\text{cov}}$ appears only in the self terms, not the pair terms, since no 2 mobile residues are within 3 bonds of each other.

$$\Delta G^{\text{fold}}_{\text{low}}(i,j) = \begin{cases} 0 & \text{if } i,j \text{ are on opposite binding partners} \\ G_{\text{vdW}}(i : j) + G_{\text{ES, dd}}(i : j) & \text{if } i,j \text{ are on the same binding partner} \end{cases} \qquad (6.48)$$

6.4 Results and Discussion

6.4.1 Results for Barnase/Barstar "Center 3" Redesign

We used the methods described above to redesign the "center 3" residues of barstar (Asp35, Trp38, Val73) for optimal binding to barnase. Three-stage DEE/A* (see Section 6.3.3) was run to rank sequences, and to rank up to 10 structures for each sequence, by low-resolution binding energy. Before DEE elimination of single rotamers and rotamer pairs, rotamers and rotamer pairs were eliminated by cutoffs of 25 kcal mol^{-1} applied to the self and pair terms of the low-resolution binding and folding energies.

Low-Resolution Energy Function

DEE/A* found 481 sequences with low-resolution binding energy within 30 kcal mol^{-1} of the minimum. (There are $20^3 = 8000$ total possible sequences.) Each sequence was represented by up to 10 structures, so there were 4629 total structures. The amino acids that these 481 sequences have at the 3 mobile residue positions is shown in Table 6.3. Of these 3 positions, only residue 73 has much solvent exposure in the bound state, but all 3 positions sample many amino acid types. This diversity of amino acid types is made possible by our three-stage DEE/A* procedure, because the first stage can produce a long list of sequences, and the second stage keeps the total number of structures under control by only keeping 10 fleximer states per sequence.

Table 6.3: Amino Acid Frequency in the 481 sequences within 30 kcal mol^{-1} of the minimum binding energy found by DEE/A* for the "center 3" barstar residues (Asp35, Trp38, Val73). The wild-type amino acids are shown in boldface.

Amino Acid	barstar 35	barstar 38	barstar 73
ALA	7	18	23
ARG		17	24
ASN	76	19	27
ASP	**340**	52	41
CYS	12	18	23
GLN		22	33
GLU		47	76
GLY	3	17	22
HIS		25	26
HSD		23	22
ILE		18	23
LEU		19	24
LYS		18	23
MET		19	25
PHE		26	
SER	13	18	23
THR	27	18	23
TRP		**42**	
TYR		28	
VAL	3	17	**23**

The top of the list of sequences, as ranked by low-resolution binding energy, is shown in Table 6.4. The low-resolution binding energy function favors sequences with as much negative charge on barstar as possible. Such sequences get a favorable direct interaction with the net positive charge of barnase, but the low-resolution energy function has no desolvation penalty for the shielding of the charged groups from solvent by the binding partner, nor does it have a penalty for the enhancement of the negative charges' interactions with each other upon binding.

Medium-Resolution Energy Function

Medium-resolution binding and folding energies were then calculated for all 4629 structures. The beginning of the list of sequences sorted by medium-resolution binding energy is shown in Figure 6.5. Using this energy function which includes the desolvation penalty and indirect interaction binding terms, sequences with maximum negative charge are no longer given the best binding energies, in contrast to the low-resolution energy function.

The medium- vs. low-resolution binding energies are plotted in Figure 6-12. They are strongly correlated, but the color-coding, by total charge of the mobile residues, shows that the low-resolution binding energy incorrectly favors some structures with -3 and -2 charge. These could be called "false positives," because the low-resolution energy function scored them favorably, but the medium-resolution energy function found them to be unfavorable. On the other hand, there are no "false negatives" (these would appear as points in the lower right of Figure 6-12). A lack of false negatives is a very good feature in a low-resolution energy function, because it makes it possible to use the low-resolution energy function as a screen to determine which structures to pass along to the better energy functions. A false positive only costs a little extra computation before the better energy function finds it unfavorable, at which point it can be discarded. A false negative, on the other hand, is a desirable structure which may never be found if one can not

Table 6.4: Barstar "center 3" sequences sorted by low-resolution binding energy relative to the wild type $\Delta\Delta G^{\text{bind}}_{\text{low}} = \Delta G^{\text{bind}}_{\text{low}} - \Delta G^{\text{bind}}_{\text{low}}(\text{wildtype})$. The wild-type sequence is shown in boldface; it appears in the list, but is also shown at the beginning of the table for reference. Note that "H" and "h" are HIS and HSD, the PARAM19 forms of histidine protonated on the $N_{\delta 1}$ and the $N_{\epsilon 2}$ atoms, respectively. For clarity, every position which matches the wild type is shown as a blank space.

$\Delta\Delta G^{\text{bind}}_{\text{low}}$	Sequence		
(kcal mol^{-1})	35	38	73
Wild type:			
0.	**D**	**W**	**V**
In order of low-res. binding:			
-12.35		D	E
-11.72		E	E
-10.63			E
-8.89		Y	E
-7.73		F	E
-7.62		D	D
-7.39		H	E
-7.03		h	E
-6.99		E	D
-5.91			D
-5.64		Q	E
-5.36		D	Q
-4.73		E	Q
-4.24		D	N
-4.16		Y	D
-4.05		N	E
-3.98		M	E
-3.69		L	E
-3.65			Q
-3.61		E	N
-3.48		D	H
-3.15		D	M
-3.01		F	D
-2.85		E	H
-2.78		D	L
-2.66		H	D
-2.62		D	R
-2.54		T	E
-2.54		D	I
-2.53			N
-2.52		E	M
-2.35		S	E
-2.31		h	D
↪			
-2.15		E	L
-1.99		E	R
-1.90		E	I
-1.90		Y	Q
-1.85		D	T
-1.76			H
-1.71		D	
-1.69		D	S
-1.45		D	K
-1.44			M
-1.43		D	C
-1.22		E	T
-1.11		D	A
-1.08		E	
-1.07			L
-1.06		E	S
-0.93		D	G
-0.91		Q	D
-0.91			R
-0.82		E	K
-0.82			I
-0.80		E	C
-0.78		Y	N
-0.78		K	E
-0.75		F	Q
-0.58		C	E
-0.49		D	h
-0.48		E	A
-0.40		H	Q
-0.30		E	G
-0.27		I	E
-0.14			T
-0.05		h	Q
-0.02		Y	H
0.	**D**	**W**	**V**
...			

Table 6.5: Barstar "center 3" sequences sorted by medium-resolution binding energy relative to the wild type $\Delta\Delta G^{\text{bind}}_{\text{medium}} = \Delta G^{\text{bind}}_{\text{medium}} - \Delta G^{\text{bind}}_{\text{medium}}(\text{wildtype})$. The wild-type sequence is shown in boldface; it appears in the list, but is also shown at the beginning of the table for reference. Note that "H" and "h" are HIS and HSD, the PARAM19 forms of histidine protonated on the $N_{\delta 1}$ and the $N_{\epsilon 2}$ atoms, respectively. For clarity, every position which matches the wild type is shown as a blank space.

$\Delta\Delta G^{\text{bind}}_{\text{medium}}$	Sequence		
(kcal mol^{-1})	35	38	73
Wild type:			
0.	**D**	**W**	**V**
In order of medium-res. binding:			
-9.76			E
-8.10		Y	E
-6.82		F	E
-4.18			D
-3.84		h	E
-3.39			Q
-3.30		H	E
-2.98		Y	D
-2.53		Q	E
-2.49			M
-2.02		M	E
-1.95			N
-1.95			L
-1.71		F	D
-1.67			H
-1.60			I
-1.51		Y	Q
-0.67		L	E
-0.64		Y	M
-0.51		F	Q
-0.25			T
-0.13		Y	N
-0.08			S
0.	**D**	**W**	**V**
...			

afford to consider many more of the structures that the low-resolution energy function finds unfavorable.

High-Resolution Energy Function

High-resolution binding and folding energies were then calculated for all 4629 structures. The beginning of the list of sequences sorted by high-resolution binding energy is shown in Figure 6.6.

The high- vs. medium-resolution binding energies, plotted in Figure 6-13, are strongly correlated, and the error does not correlate with the total charge as it did for the medium- vs. high-resolution data (the lines on the plot are discussed in Section 6.4.2, and may be ignored for now).

Electrostatic Binding Energy

The low-, medium-, and high-resolution energy functions all have the same covalent and van der Waals terms, and differ only in their hydrophobic and electrostatic terms. To clarify the source of their differences, we compare their electrostatic terms only. In Figure 6-14, the color-coding of FDPB vs. "4r dielectric" binding electrostatics data shows that, for each total charge, the correlation is fairly good; the largest source of error is a function only of the total charge. As stated earlier, this error is caused by the low-resolution "4r dielectric" electrostatic energy's lack of desolvation penalties and indirect interaction binding terms. The FDPB vs. ACE binding electrostatics data, shown in Figure 6-15, have much better correlation.

Folding Energy

The low-, medium-, and high-resolution folding energies are shown in Figures 6-16 and 6-17 (the lines on the latter figure are discussed in Section 6.4.2, and may be ignored for now). The medium- vs. high-resolution folding energies are very well correlated. Again,

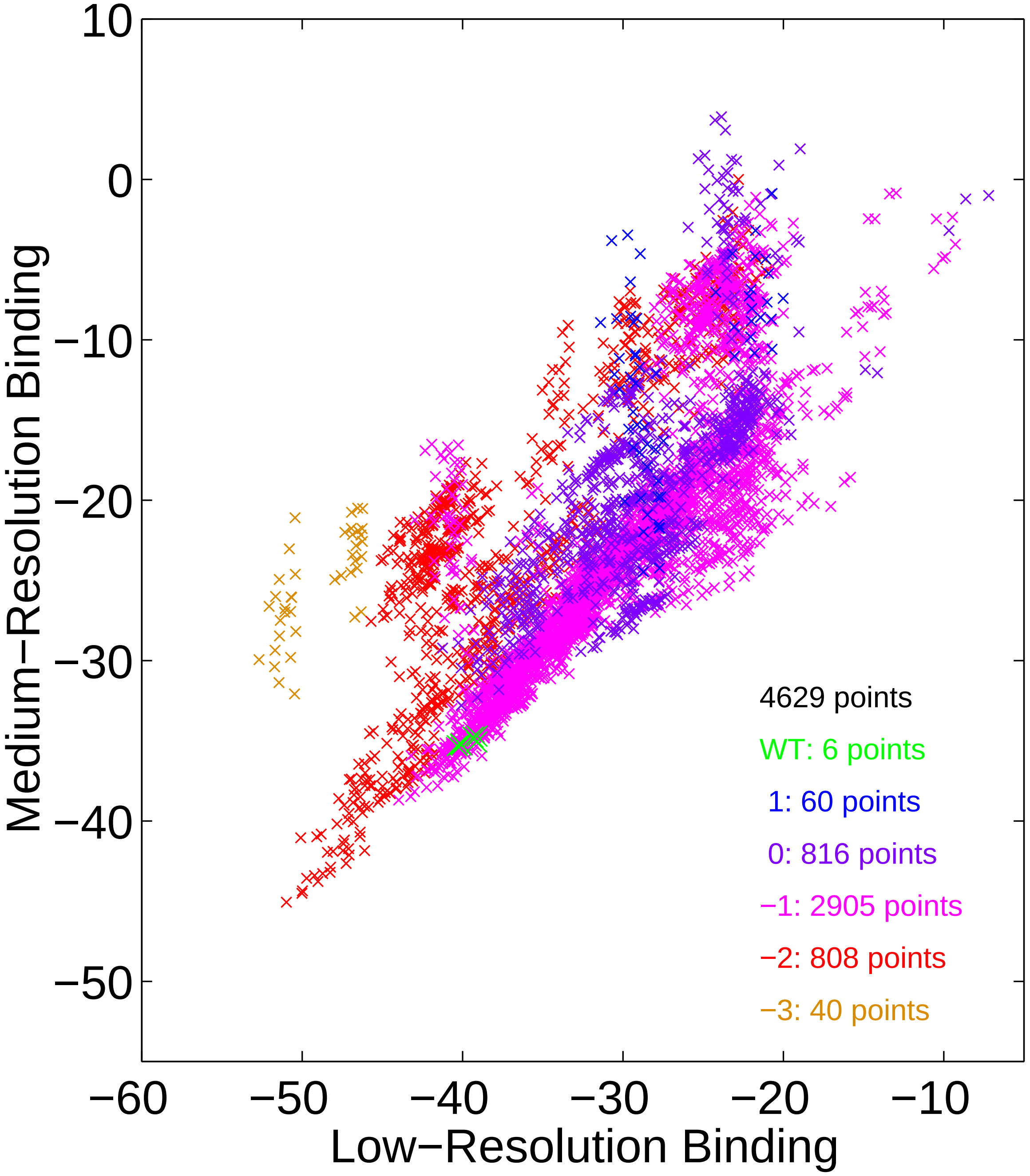

Figure 6-12: Medium-resolution $\Delta G^{\text{bind}}_{\text{med}}$ vs. low-resolution binding $\Delta G^{\text{bind}}_{\text{low}}$ energy for 4629 mutant structures from DEE/A* on the "center 3" residues (Asp35, Trp38, Val73). Structures are colored by the charge of these three residues. Structures with the wild-type sequence are shown as green "X"s. The energy on each axis has an arbitrary constant term.

Table 6.6: Barstar “center 3” sequences sorted by high-resolution binding energy relative to the wild type $\Delta\Delta G^{\text{bind}}_{\text{high}} = \Delta G^{\text{bind}}_{\text{high}} - \Delta G^{\text{bind}}_{\text{high}}(\text{wildtype})$. The wild-type sequence is shown in boldface; it appears in the list, but is also shown at the beginning of the table for reference. Note that “H” and “h” are HIS and HSD, the PARAM19 forms of histidine protonated on the $N_{\delta 1}$ and the $N_{\epsilon 2}$ atoms, respectively. For clarity, every position which matches the wild type is shown as a blank space.

$\Delta\Delta G^{\text{bind}}_{\text{high}}$	Sequence		
(kcal mol^{-1})	35	38	73
Wild type:			
0.	**D**	**W**	**V**
In order of high-res. binding:			
-4.05			E
-2.20			Q
-2.19		Y	E
-2.13		F	E
-1.62		H	E
-1.51			D
-1.08			L
-1.01			N
-0.91			H
-0.88			I
-0.47			M
-0.43		h	E
-0.23		F	Q
-0.11		Y	Q
0.	**D**	**W**	**V**
...			

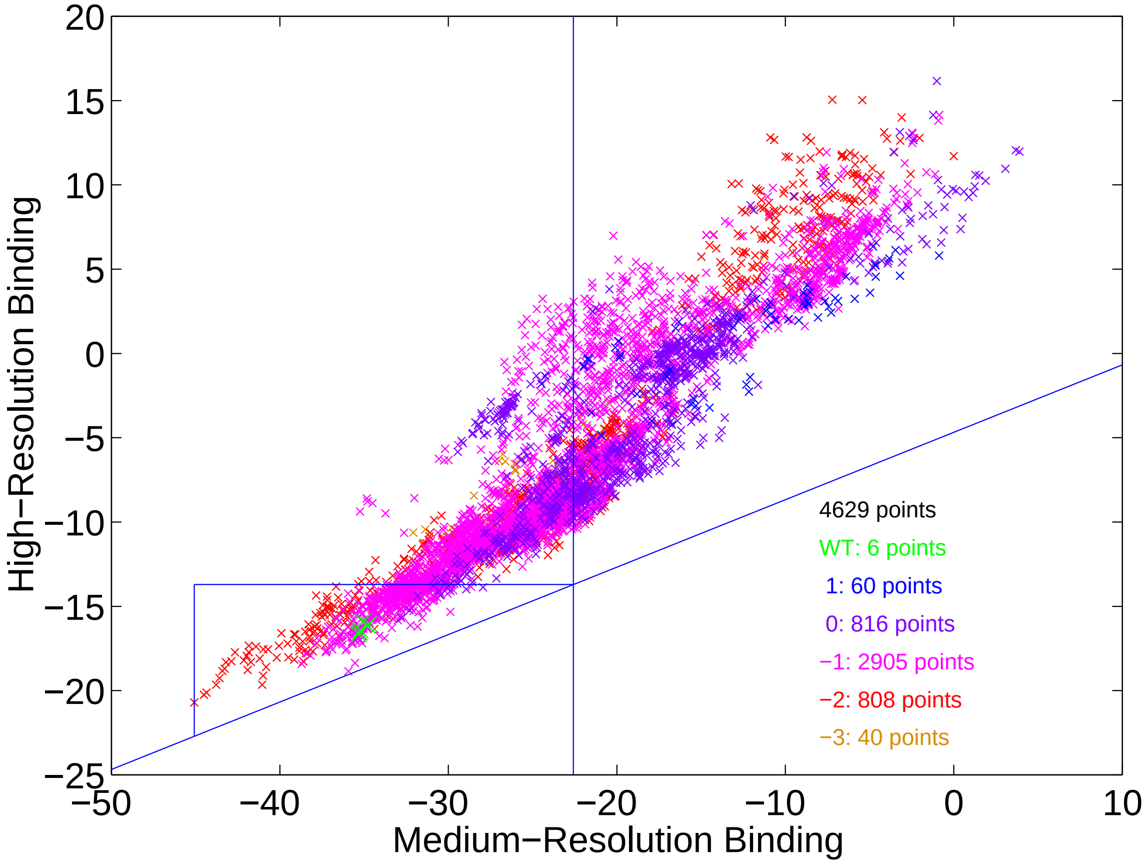

Figure 6-13: High-resolution $\Delta G^{\text{bind}}_{\text{high}}$ vs. medium-resolution binding $\Delta G^{\text{bind}}_{\text{med}}$ energy for 4629 mutant structures from DEE/A* on the "center 3" residues (Asp35, Trp38, Val73). Structures are colored by the charge of these three residues. Structures with the wild-type sequence are shown as green "X"s. The lines illustrate the protocol in Section 6.4.2. The energy on each axis has an arbitrary constant term.

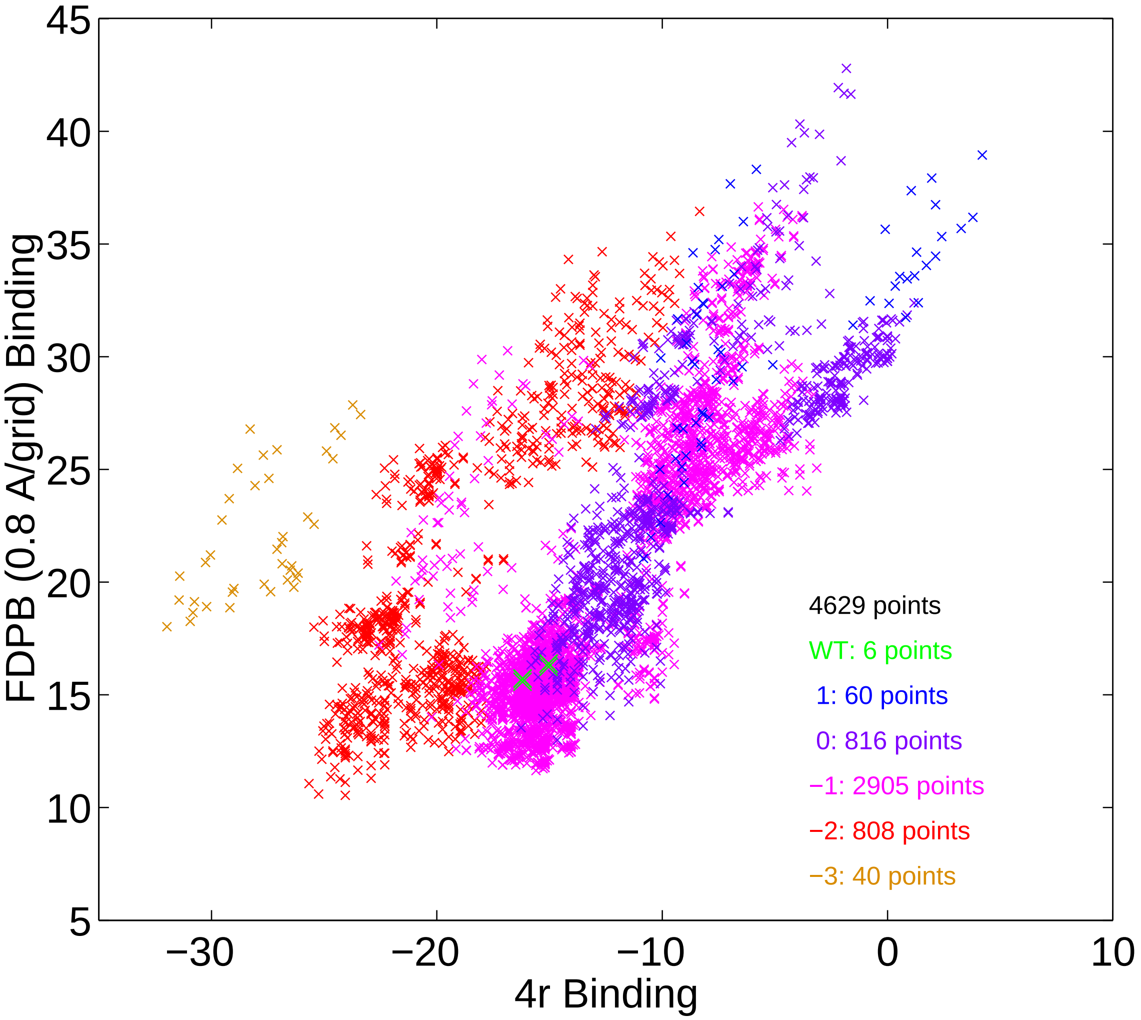

Figure 6-14: FDPB $\Delta G^{\text{bind}}_{\text{ES, FDPB}}$ vs. "4r dielectric" distance-dependent Coulombic $\Delta G^{\text{bind}}_{\text{ES, dd}}$ electrostatic binding free energy for 4629 mutant structures from DEE/A* on the "center 3" residues (Asp35, Trp38, Val73). Structures are colored by the charge of these three residues. The color-coding shows that, for each total charge, the correlation is fairly good; the largest source of error is a function only of the total charge. Structures with the wild-type sequence are shown as green "X"s. The FDPB electrostatic term uses 0.8 Å/grid. Both dimensions are in kcal mol^{-1}.

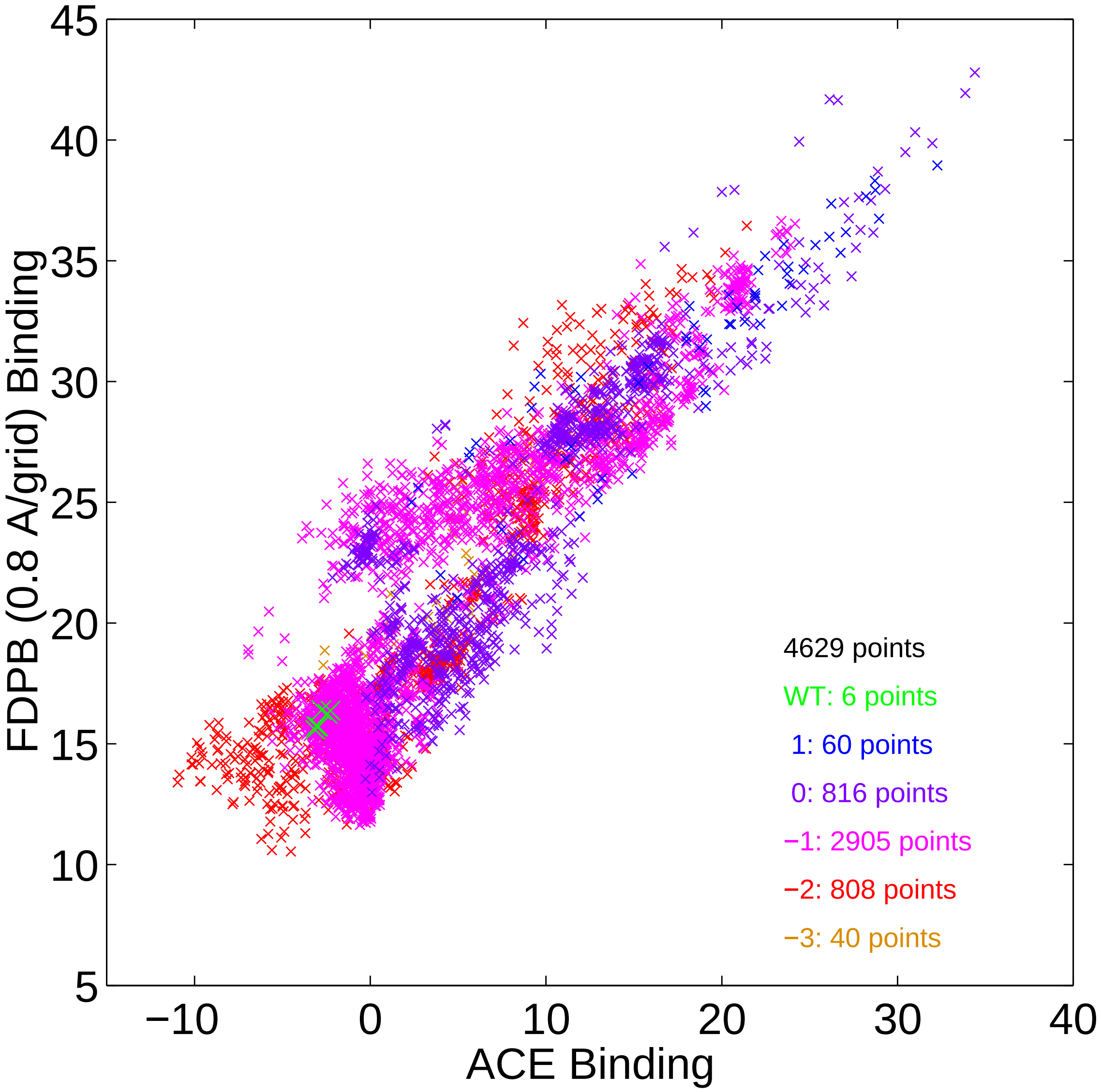

Figure 6-15: FDPB $\Delta G^{\text{bind}}_{\text{ES, FDPB}}$ vs. ACE $\Delta G^{\text{bind}}_{\text{ES, ACE}}$ electrostatic binding free energy for 4629 mutant structures from DEE/A* on the "center 3" residues (Asp35, Trp38, Val73). Structures with the wild-type sequence are shown as red diamonds. The FDPB electrostatic term uses 0.8 Å/grid. The correlation coefficient is $r = 0.902$. Both dimensions are in kcal mol^{-1}.

there are no false negatives (which would appear as points in the lower right of these plots).

It is helpful to consider the distribution of the structures in the two dimensions of binding and folding energy, as shown in Figures 6-18, 6-19, and 6-20 for the low-, medium-, and high-resolution energy functions. On such a plot, the promising structures are in the lower left corner. That is, they are not much above the wild type in folding energy, and they bind as tightly as possible (are as far left as possible). The most striking thing about Figures 6-19 and 6-20 is that 2 structures with the wild-type sequence are in the extreme lower left corner; i.e., the wild-type sequence is predicted to be a very tight binder and folder.

Use of Wild-Type Folding Stability

Our energy functions were developed to compare binding and folding free energies of different structures, not to calculate absolute binding or folding free energy values. In order to make use of our calculated folding free energies, we take the known experimental folding stability of wild-type barnase and barstar as a reference. The experimental free energy of folding for barstar in water is -5.28 kcal mol^{-1}, and slightly more stable, -5.88 kcal mol^{-1}, in 300 mM NaCl [34]. Proteins typically have folding free energies in the range -5 to -15 kcal mol^{-1}; since barstar is on the less stable side of that range, redesigned structures that are predicted to be less stable than the wild type are not promising candidates. Therefore, we have chosen to screen the structures with the following criterion: the high-resolution folding free energy must be no more than one kcal mol^{-1} worse than wild type.

Of the six structures representing the wild-type sequence (shown in Figure 6-21), two of them have much better folding energy than the others. One of these two has the best binding energy by a small margin, and this is the structure we chose to represent the wild-type sequence. Its structure is very similar to the crystal structure.

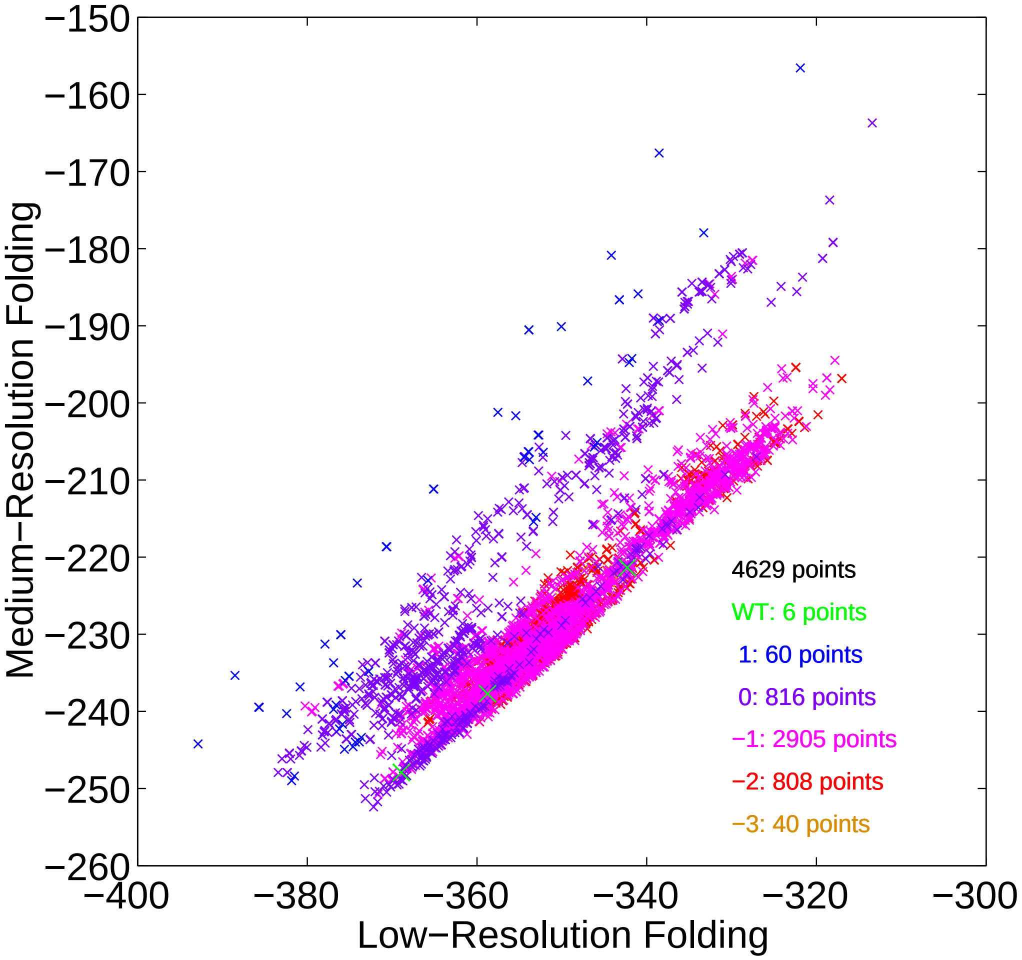

Figure 6-16: Medium-resolution $\Delta G^{\text{fold}}_{\text{med}}$ vs. low-resolution folding $\Delta G^{\text{fold}}_{\text{low}}$ energy for 4629 mutant structures from DEE/A* on the "center 3" residues (Asp35, Trp38, Val73). Structures are colored by the charge of these three residues. Structures with the wildtype sequence are shown as green "X"s. The energy on each axis has an arbitrary constant term.

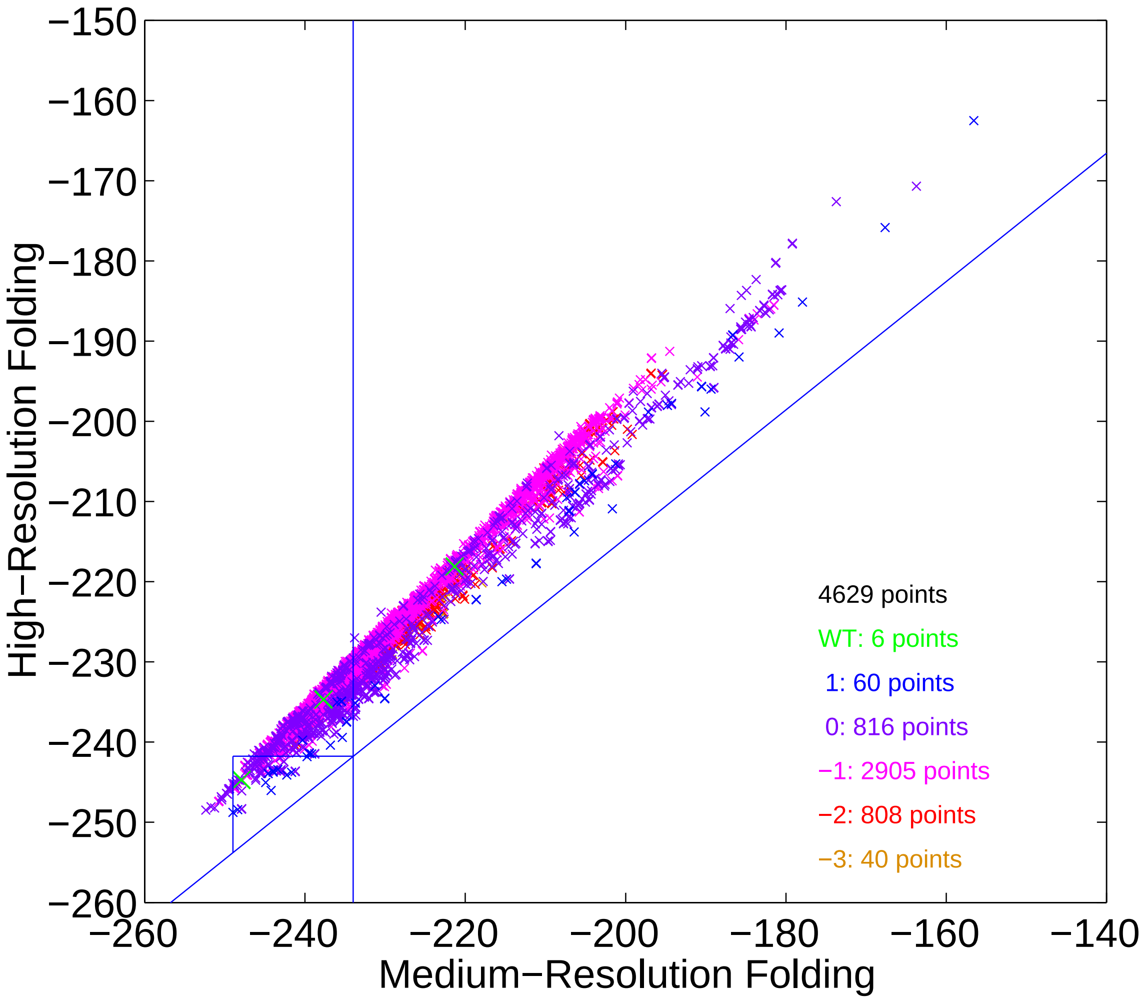

Figure 6-17: High-resolution $\Delta G^{\text{fold}}_{\text{high}}$ vs. medium-resolution folding $\Delta G^{\text{fold}}_{\text{med}}$ energy for 4629 mutant structures from DEE/A* on the "center 3" residues (Asp35, Trp38, Val73). Structures are colored by the charge of these three residues. Structures with the wild-type sequence are shown as green "X"s. The lines illustrate the protocol in Section 6.4.2: The slanted line is an empirical lower bounding line; there are no points below the line, and therefore no "false negatives". To get every structure whose $\Delta G^{\text{fold}}_{\text{high}}$ is within $c^{\text{fold}} = 7$ kcal mol^{-1} of the minimum (below the horizontal line segment), only structures with $\Delta G^{\text{fold}}_{\text{med}}$ below a cutoff (to the left of the vertical line) need to have their $\Delta G^{\text{fold}}_{\text{high}}$ calculated. The energy on each axis has an arbitrary constant term.

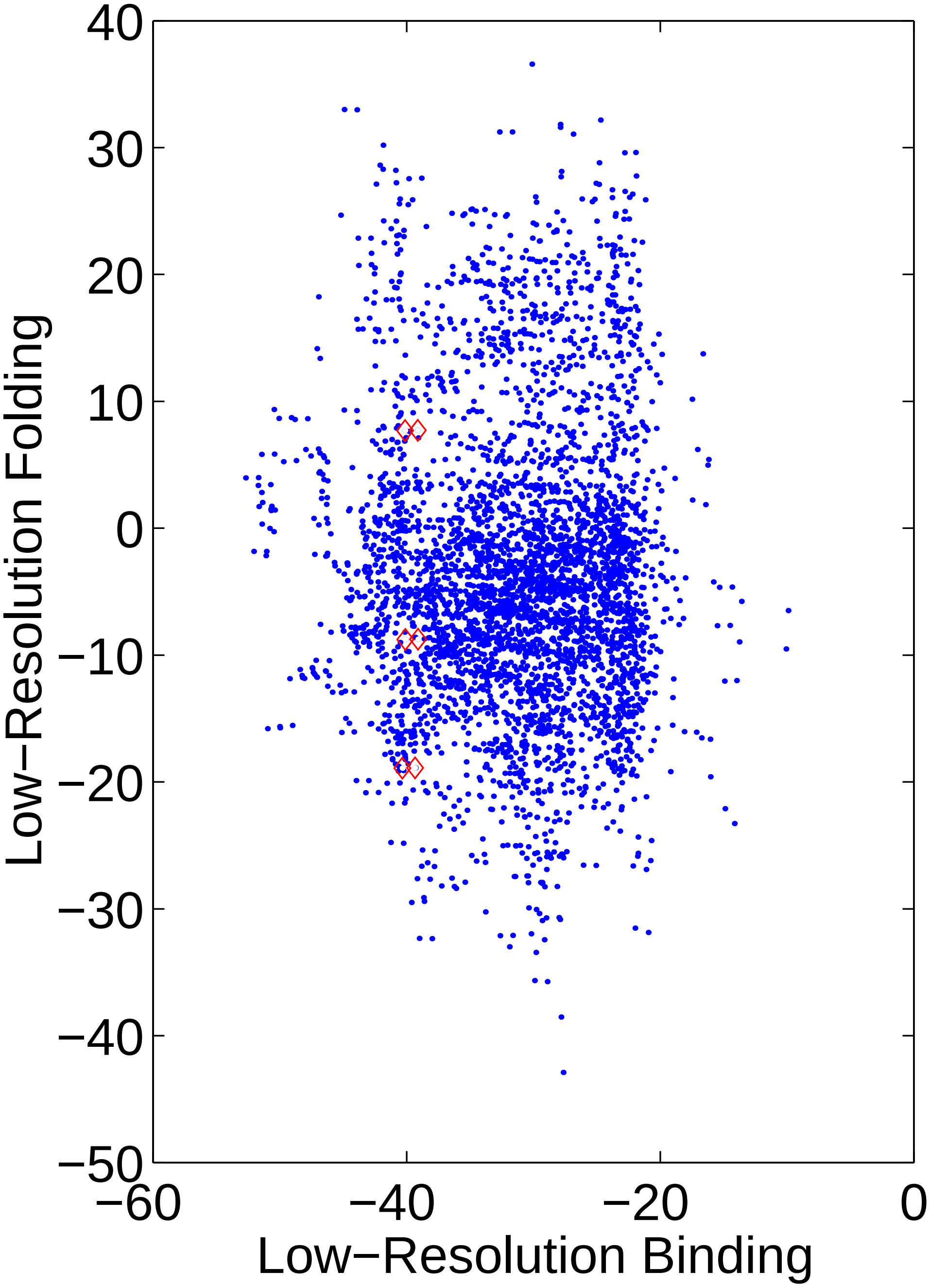

Figure 6-18: Low-resolution folding $\Delta G^{\text{fold}}_{\text{low}}$ vs. binding $\Delta G^{\text{bind}}_{\text{low}}$ energy for 4629 mutant structures from DEE/A* on the "center 3" residues (Asp35, Trp38, Val73). Structures with the wild-type sequence are shown as red diamonds. The energy on each axis has an arbitrary constant term. Both dimensions are in kcal mol^{-1}.

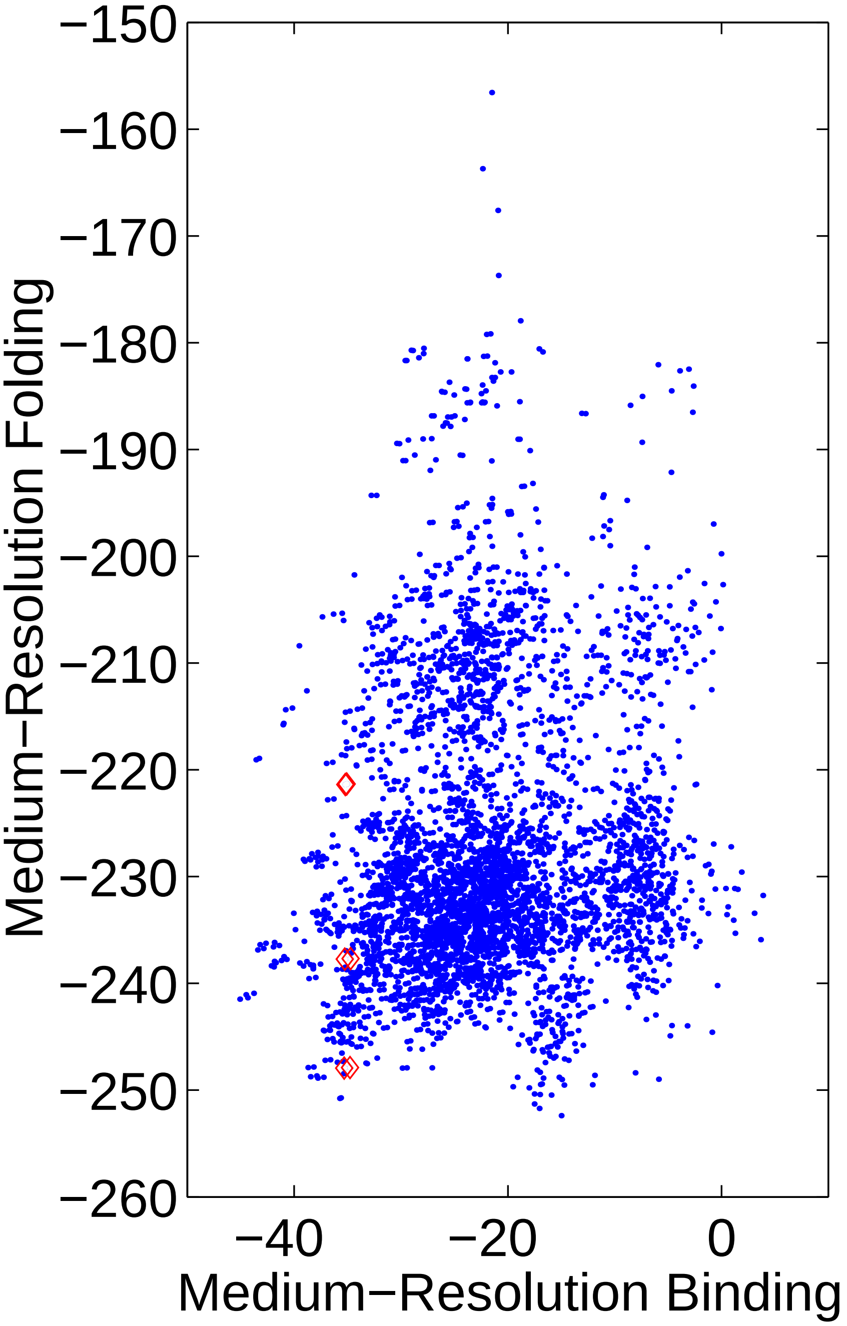

Figure 6-19: Medium-resolution folding $\Delta G^{\text{fold}}_{\text{med}}$ vs. binding $\Delta G^{\text{bind}}_{\text{med}}$ energy for 4629 mutant structures from DEE/A* on the "center 3" residues (Asp35, Trp38, Val73). Structures with the wild-type sequence are shown as red diamonds. The energy on each axis has an arbitrary constant term. Both dimensions are in kcal mol^{-1}.

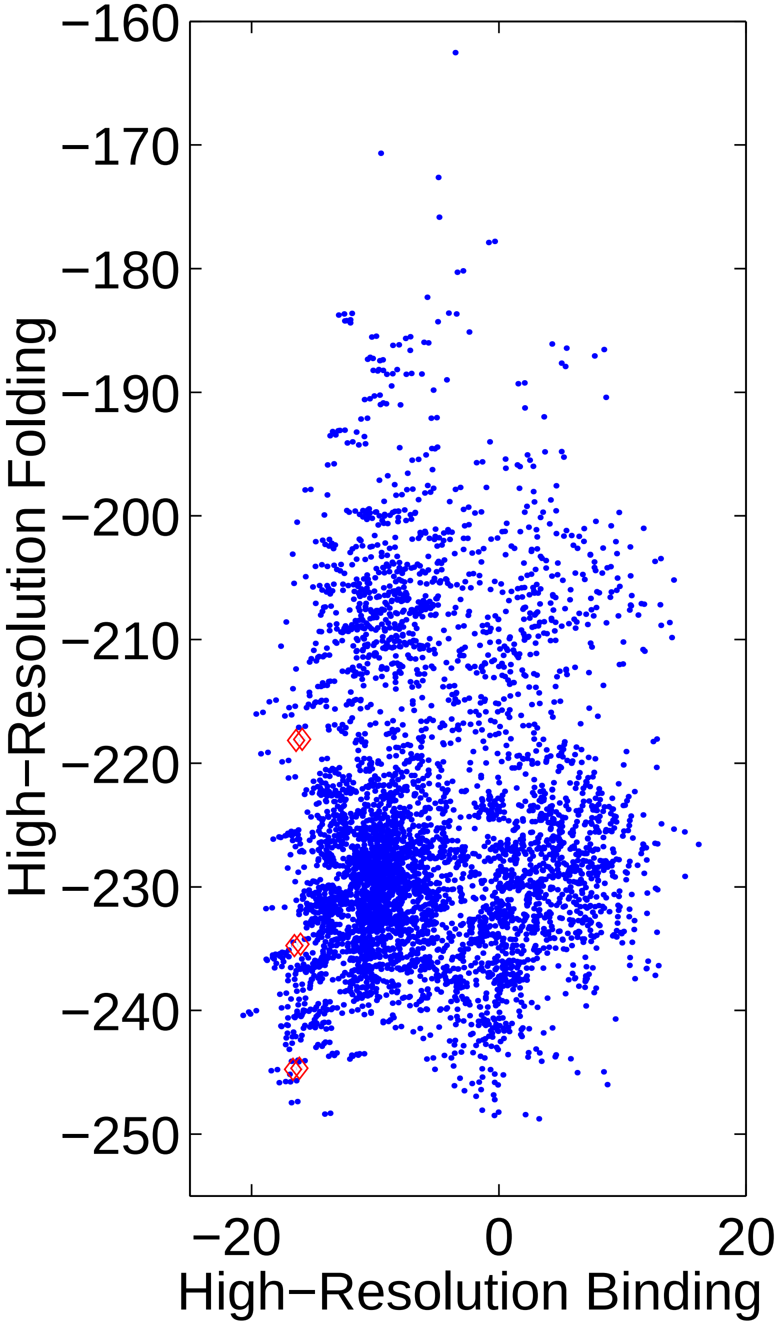

Figure 6-20: High-resolution folding $\Delta G^{\text{fold}}_{\text{high}}$ vs. binding $\Delta G^{\text{bind}}_{\text{high}}$ energy for 4629 mutant structures from DEE/A* on the "center 3" residues (Asp35, Trp38, Val73). Structures with the wild-type sequence are shown as red diamonds. The energy on each axis has an arbitrary constant term. Both dimensions are in kcal mol^{-1}.

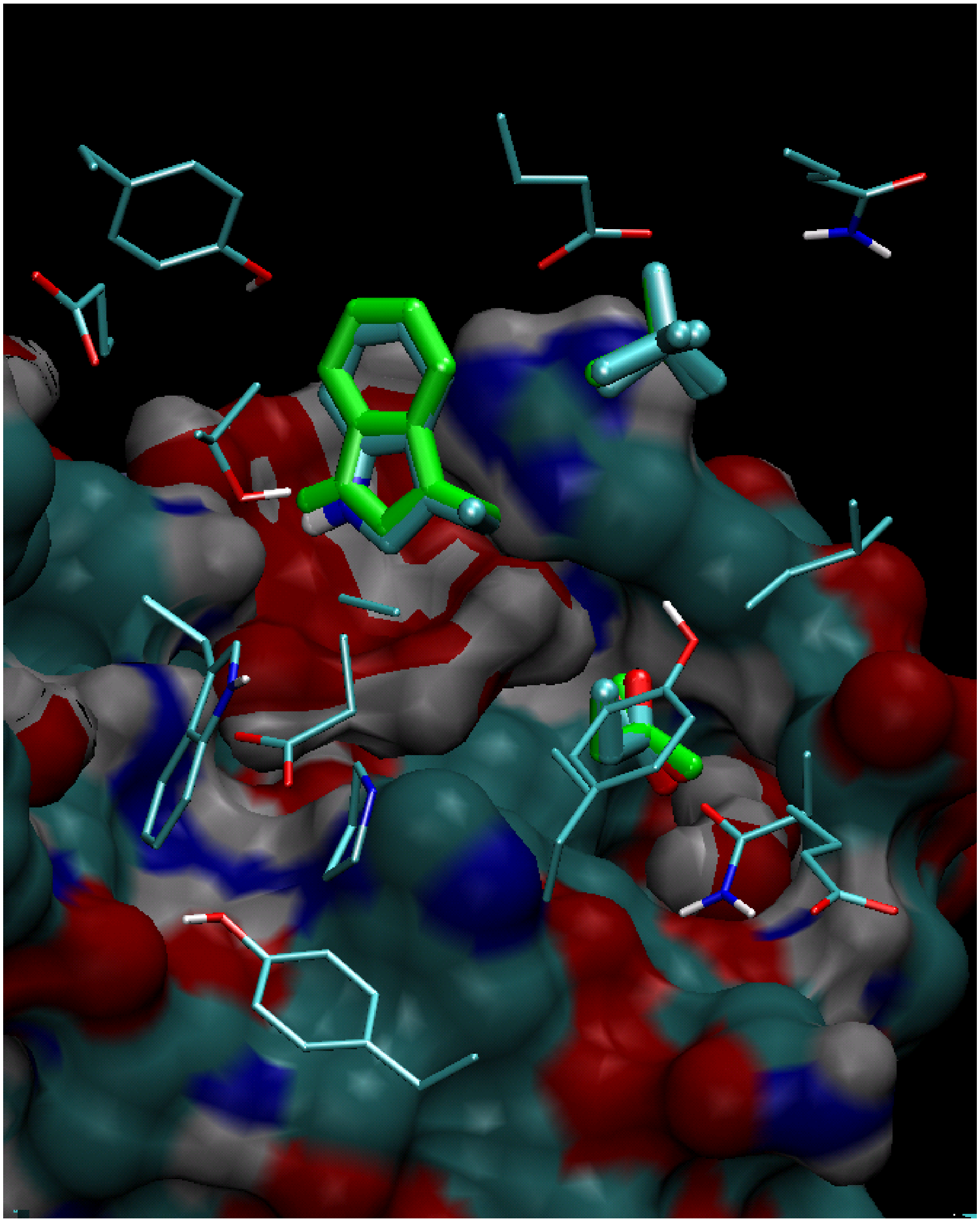

Figure 6-21: DEE/A* Redesign of “center 3” Residues: Structures with Wild-type Sequence DWV. Barnase and the interfacial waters are shown as surfaces colored by element. We are looking through barnase, which is invisible except for side chains that bury solvent-accessible surface area upon binding. The “center 3” barstar residues (Asp35, Trp38, Val73) are shown in fat licorice. The crystal structure positions are shown in green. The 6 DEE/A*-designed structures with the wild-type sequence are colored by element. They are very much like the crystal structure, except that some of them rotate Val73.

Prediction of Stable Tight-Binding Structures

To represent each sequence with a single set of values for the binding and folding free energies, we use the structure with the best high-resolution binding free energy from among from among those whose high-resolution folding free energy is no more than one kcal mol^{-1} worse than wild type. Out of the 4629 structures shown as points on Figure 6-20, this leaves only the six points on Figure 6-22. The final result of our design procedure is this small collection of promising sequences. The information in Figure 6-22 is also shown in Table 6.7.

Table 6.7: Barstar "center 3" sequences found with good high-resolution binding and folding. (For each sequence, the structure with lowest high-resolution binding free energy, which also has high-resolution folding free energy no more than one kcal mol^{-1} worse than wild type, is used.) They are sorted by high-resolution binding free energy relative to the wild type $\Delta\Delta G^{\text{bind}}_{\text{high}} = \Delta G^{\text{bind}}_{\text{high}} - \Delta G^{\text{bind}}_{\text{high}}(\text{wildtype})$. The high-resolution folding free energy is also shown relative to the wild type. The wild-type sequence is shown in boldface; it appears in the list, but is also shown at the beginning of the table for reference. For clarity, every position which matches the wild type is shown as a blank space.

$\Delta\Delta G^{\text{bind}}_{\text{high}}$	$\Delta\Delta G^{\text{fold}}_{\text{high}}$	Sequence		
(kcal mol^{-1})		35	38	73
Wild type:				
0.	**0.**	**D**	**W**	**V**
Good folders in order of binding:				
-1.77	-0.10			Q
-0.28	-0.40			I
-0.24	-0.99			M
-0.13	0.67			L
0.	**0.**	**D**	**W**	**V**
2.56	-3.61			R

Wild-type Sequence Asp35, Trp38, Val73

The wild-type sequence does very well, ranking #5 out of 8000 possible sequences (20^3). Furthermore, as stated above, the best structure with the wild-type sequence is very similar to the crystal structure. So, at the end of our procedure, we have a structure very much like the crystal structure ranked 5th out of 3×10^{10} possible structures (3098^3)!

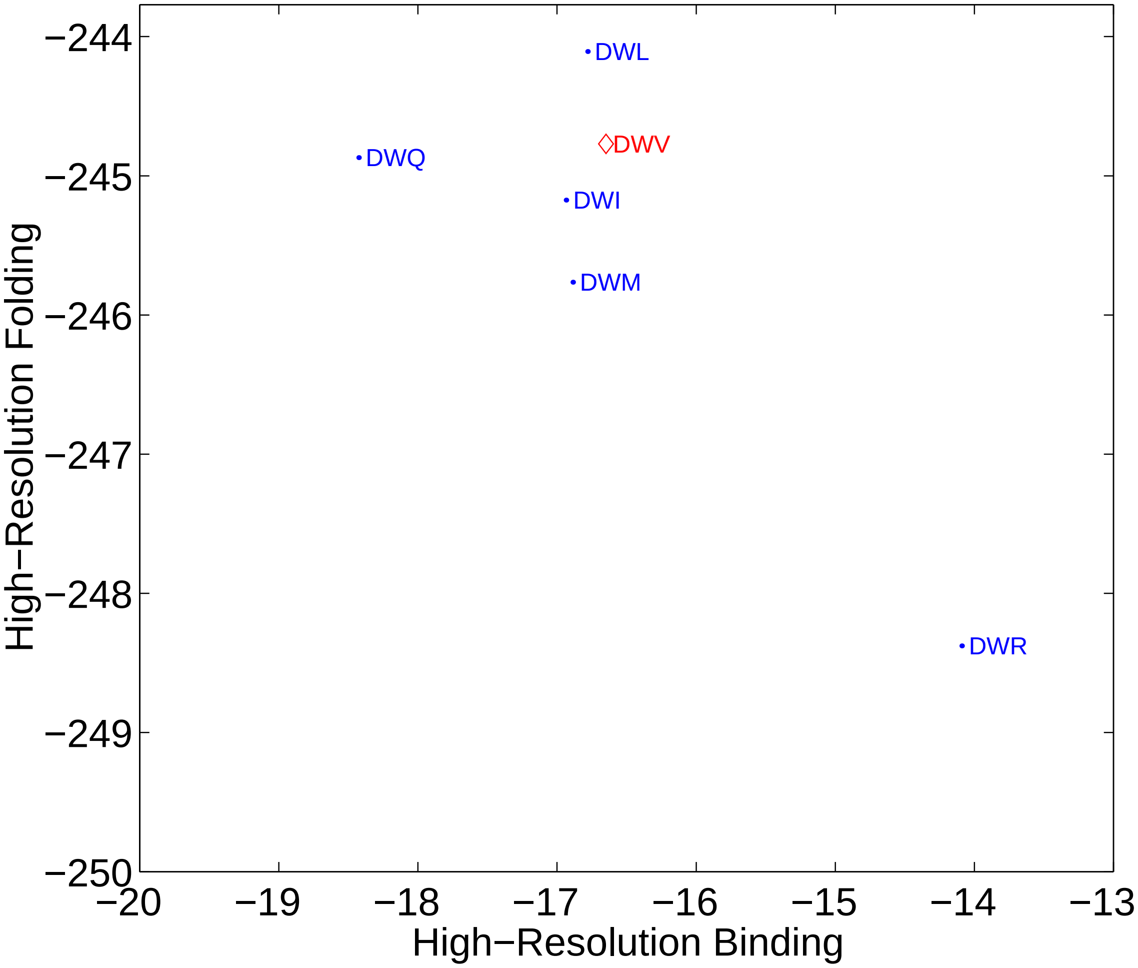

Figure 6-22: High-resolution folding $\Delta G^{\text{fold}}_{\text{high}}$ vs. binding $\Delta G^{\text{bind}}_{\text{high}}$ energy from DEE/A* on the "center 3" residues (Asp35, Trp38, Val73). For each sequence, we only kept the best-binding structure with folding no more than one kcal mol^{-1} worse than the wild-type sequence structure shown as a red diamond ($\Delta G^{\text{fold}}_{\text{high}} \leq \Delta G^{\text{fold}}_{\text{high}}(\text{wildtype}) + 1$ kcal mol^{-1}). The sequences shown are the only ones that meet this criterion. Each point is labelled with the sequence of positions 35, 38, and 73. The energy on each axis has an arbitrary constant term. Both dimensions are in kcal mol^{-1}.

If we rank the sequences by binding energy using the best-binding structure for each sequence, then, as we go from low- to medium- to high-resolution energy functions, the ranking of the wild-type sequence's binding energy improves from #67 to #24 to #14. And the ranking of the wild-type sequence's folding energy improves from #70 to #11 to #14.

Promising Non-Wild-type Sequences

There are only 4 non-wild-type sequences with high-resolution binding as good as wild type, and high-resolution folding no more than 1 kcal mol^{-1} worse than wild type. (See Table 6.7.) They are all single-position mutants, with barstar's short Val73 side chain mutated to Gln, Ile, Met, or Leu. Only the Val73Gln mutant, pictured in 6-23, has significantly better binding (by -1.77 kcal mol^{-1}). The power of DEE is that it considers all possible combinations of rotamers, but in this case, the few promising mutations are single-residue mutations.

The charge optimization study of Lee and Tidor [89] sought to optimize the atomic charges on barstar's side chains, *with its shape held constant.* They found that binding was best when Val73 had a 0 charge, as opposed to a -1 or $+1$ charge, but the optimized charges on Val73 were $+0.85$ on CB, -0.46 on CG1, and -0.39 on CG2. These optimized charges form a strong dipole to interact well with barnase Arg39. In hindsight, this result strongly suggests that a favorable electrostatic interaction is possible at this position. But with the protein shape held constant, Lee and Tidor could only conclude that Val73 was close to optimal, because it had the optimal total charge of 0, and because the optimized charges only gave a small benefit of < 1 kcal mol^{-1} over the all-zero charges of the wild-type Val73. Our present study, which gives the side chains the freedom to move and grow larger, predicts that the polar Gln73, oriented to interact well with barnase Arg39, will improve binding and not hurt folding.

The -1.77 kcal mol^{-1} binding improvement of the Val73Gln mutation is mostly from the van der Waals term (-1.71 van der Waals, $+0.02$ hydrophobic, -0.08 FDPB

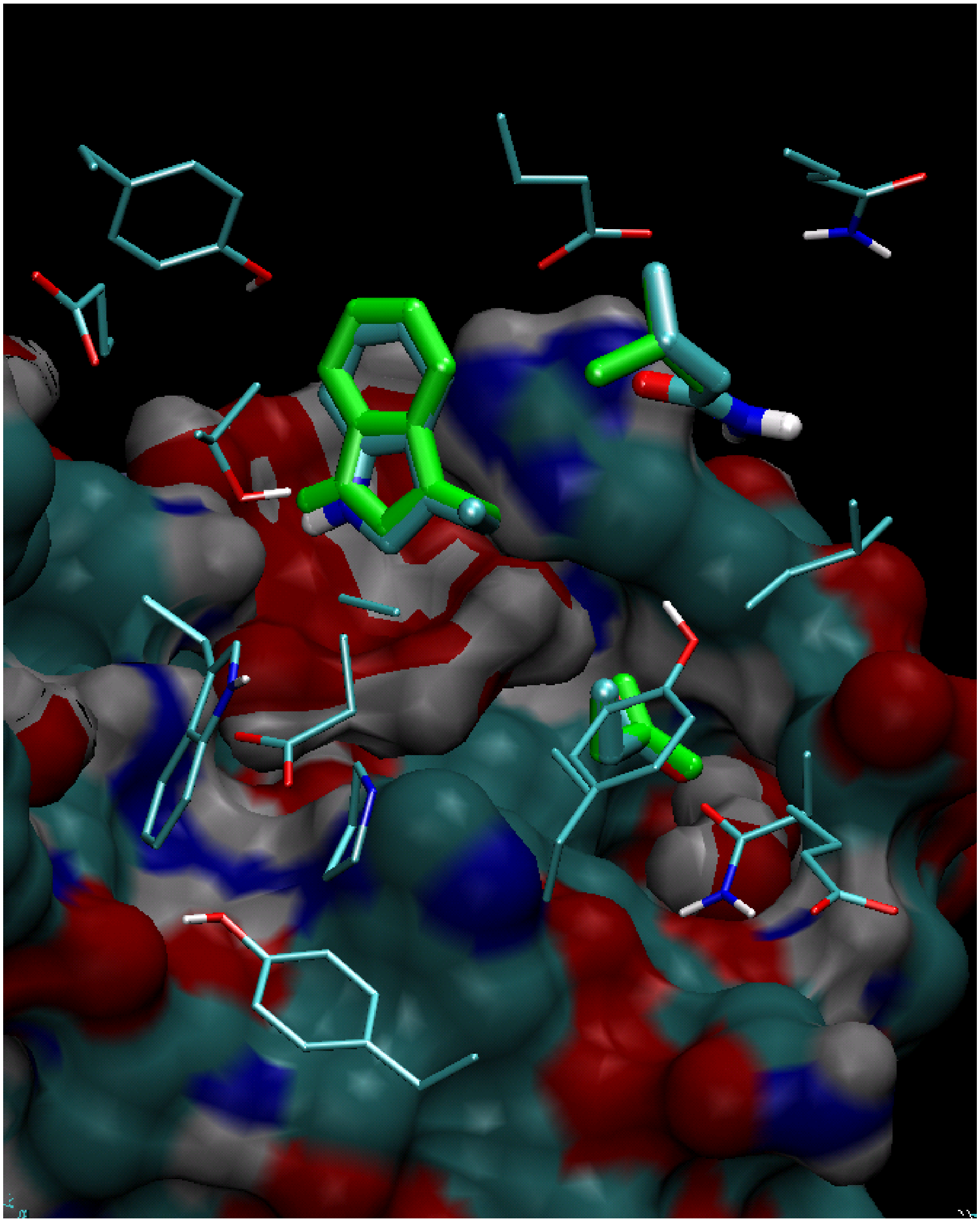

Figure 6-23: DEE/A* Redesign of “center 3” Residues: Best Structure with Sequence DWQ. Barnase and the interfacial waters are shown as surfaces colored by element. We are looking through barnase, which is invisible except for side chains that bury solvent-accessible surface area upon binding. The “center 3” barstar residues (Asp35, Trp38, Val73) are shown in fat licorice. The crystal structure positions are shown in green. The DEE/A*-designed structure with the sequence DWQ is colored by element. This sequence differs from the wild type by the mutation Val73Gln. The Gln73 is positioned to interact with barnase Arg59 (visible here as the 3 blue patches arranged in a triangle at the top of the barnase surface).

electrostatics); but this does not mean that the electrostatics are not important. Consider the mutation Val73Met: the medium-sized, hydrophobic Met73 is roughly like a hydrophobic isostere of Gln73. The Val73Met mutation improves binding by only -0.24 kcal mol^{-1} (-0.84 van der Waals, $+0.01$ hydrophobic, 0.60 FDPB electrostatics). That $+0.60$ kcal mol^{-1} electrostatic cost of desolvating this region is offset in the case of Gln73 by a favorable electrostatic interaction. The moral is that the charge distribution is important, even if the net electrostatic binding energy difference is zero.

Minimization

As a test of whether the rotamer library is fine enough, we minimized all 4629 structures found by DEE/A* in the design of the "center 3" barstar residues, and then re-calculated their high-resolution energies. The minimization was done on the bound state, with only the mobile side chains allowed to move, using the low-resolution energy function G_{low} but with full van der Waals radii.

The folding energies of many conformations improve after minimization, as shown in Figure 6-24, due mainly to improved van der Waals in the bound state. Conformations with folding energies on the upper, unfavorable, end of the range of values before minimization drop down into the lower, favorable, end of the range. For most conformations with folding energies in the lower end of the range, the value does not change as much after minimization. This demonstrates that some conformations had van der Waals clashes which minimization relieved, but most conformations did not have major van der Waals clashes.

After all structures were minimized, we calculated high-resolution energies for all structures, and screened the sequences using the high-resolution energies of the minimized structures. Each sequence was represented by its best binding structure which folds no more than one kcal mol^{-1} worse than wild type. The most promising structures are shown in Table 6.8, which should be compared to the same list from before minimization, Table 6.7. The only major change is that the single-position mutant Val73His now

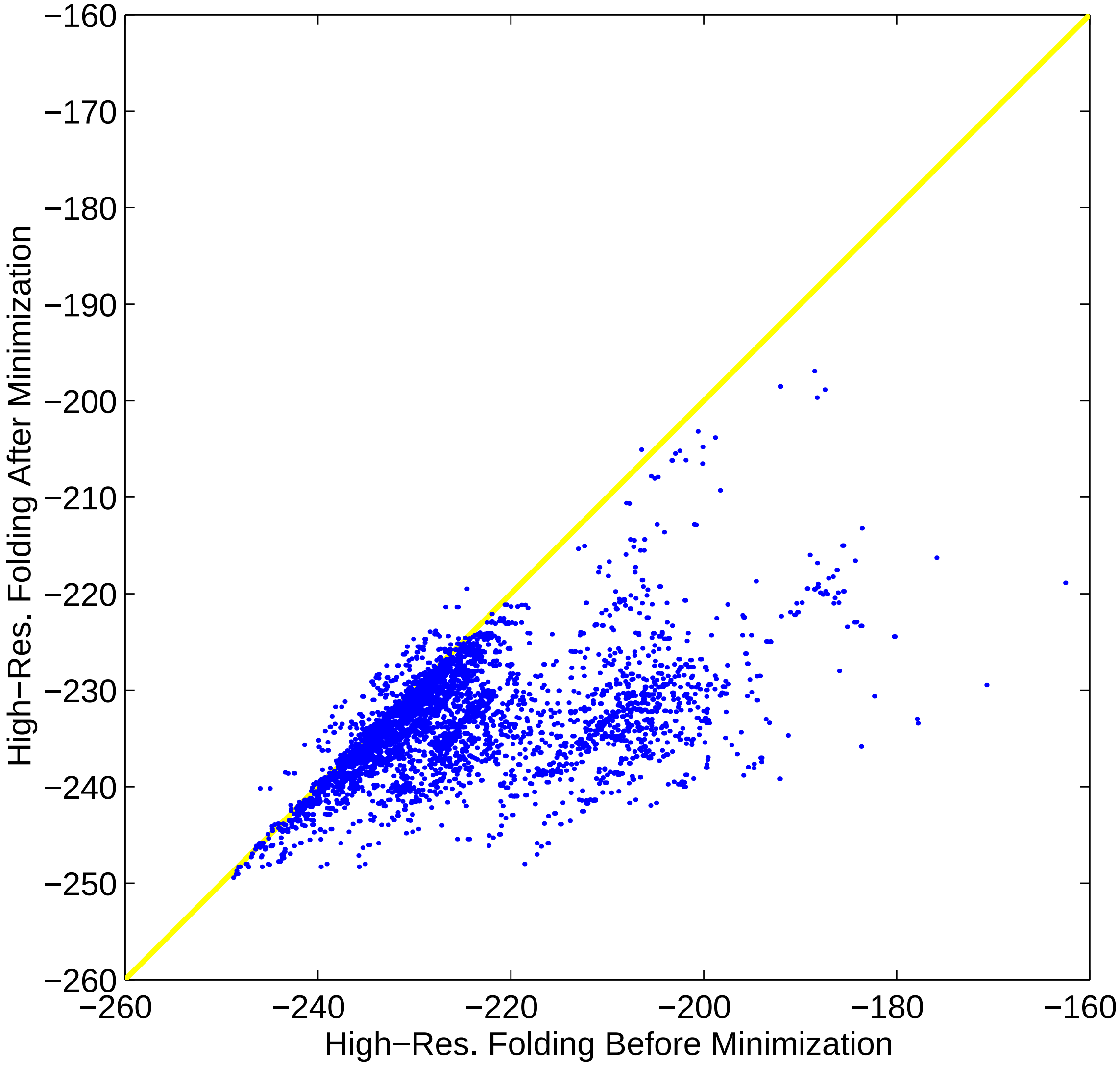

Figure 6-24: High-resolution folding free energy $\Delta G^{\text{fold}}_{\text{high}}$ before and after minimization, for 4629 mutant structures from DEE/A* on the "center 3" residues (Asp35, Trp38, Val73). Both axes are in kcal mol^{-1} and have the same arbitrary constant term.

appears as a promising sequence, with predicted binding affinity 1. kcal mol^{-1} tighter than the wild type. All of the structures of this Val73His mutant have a van der Waals clash between the His73 and Leu34 side chains of barstar. The histidine conformations in the rotamer library we used are not spaced finely enough to fit a histidine in this position, whereas minimization finds such a fit. After minimization, the electrostatic term of the binding free energy for the Val73His mutation is about the same as for the hydrophobic mutation Val73Met (+0.52 kcal mol^{-1}). The Val73His mutation improves binding by -1.05 kcal mol^{-1} mainly by making a better steric fit (-1.57 van der Waals, +0.01 hydrophobic, +0.52 FDPB electrostatics).

Table 6.8: Barstar "center 3" sequences found with good high-resolution binding and folding after minimization. Minimization followed by high-resolution energy calculation was done on all 4629 conformations of the 481 sequences within 30 kcal mol^{-1} of the minimum low-resolution binding energy before minimization $\Delta G^{\text{bind}}_{\text{low}}$. They are sorted by high-resolution binding energy after minimization, relative to the wild type $\Delta\Delta G^{\text{bind,mini}}_{\text{high}} = \Delta G^{\text{bind,mini}}_{\text{high}} - \Delta G^{\text{bind,mini}}_{\text{high}}$(wildtype). The high-resolution folding energy is also shown relative to the wild type. The wild-type sequence is shown in boldface; it appears in the list, but is also shown at the beginning of the table for reference. Note that "H" and "h" are HIS and HSD, the PARAM19 forms of histidine protonated on the $N_{\delta 1}$ and the $N_{\epsilon 2}$ atoms, respectively. For clarity, every position which matches the wild type is shown as a blank space.

$\Delta\Delta G^{\text{bind,mini}}_{\text{high}}$	$\Delta\Delta G^{\text{fold,mini}}_{\text{high}}$	Sequence		
(kcal mol^{-1})		35	38	73
Wild type:				
0.	**0.**	**D**	**W**	**V**
Good folders in order of binding, after minimization:				
-2.43	-0.09			Q
-1.05	0.20			H
-0.36	0.19			M
-0.16	-1.97			I
-0.03	0.62			h
0.	**0.**	**D**	**W**	**V**
2.65	-1.31			R

Computational Expense

For each variant structure of barnase/barstar, calculating the low-resolution binding and folding energies takes a small fraction of a second on a 1 GHz Intel Pentium III. Calculating the medium-resolution energies takes 10 seconds for each structure; most of this time is for ACE calculations of solvation energies in the bound and unbound states. Calculating the high-resolution energies takes 29 minutes for each structure; most of this time is for FDPB calculation: 10 translations each, with 65 grids/side, of $(4 + N)$ configurations (bound with each binding partner charged, unbound with each binding partner charged, and a separated desolvated state of each of the N mobile side chains).

Note that recompiling DELPHI with the Intel Fortran compiler, which optimizes the code for use on Intel x86 processors, reduces its run time by a factor of about 0.5. Using an improved DELPHI wrapper script (unpublished code, improved by David F. Green) reduces its run time by another factor of about 0.5. Before it was improved, a significant portion of the total run time was taken up by the wrapper script, which sets up the charges on the grid, calculates energy terms as sums over the point charges, etc. (everything except for the actual finite-difference solution, which DELPHI itself does). These two improvements combined reduce the DELPHI run time to 7 minutes per rotamer state, still 42 times longer than ACE.

Results for Barnase/Barstar “center 3” Redesign with Higher Hydrophobic γ Coefficient

The preceding results for the redesign of the “center 3” barstar residues (Asp35, Trp38, Val73) used a hydrophobic parameter of $\gamma = 5$ cal mol^{-1} Å^{-2}. Because this parameter is not generally agreed upon, and as a test of our method’s robustness, we repeated the ligand design procedure with the values 22.8 and 75 cal mol^{-1} Å^{-2}. The final results, which should be compared to Table 6.7, are shown in Tables 6.9 and 6.10. The top 5 sequences are in the exact same order for $\gamma =$ 5 and 22.8 cal mol^{-1} Å^{-2}. For 75 cal mol^{-1} Å^{-2}, the top 5 sequences are the same, but are slightly reordered, with their $\Delta\Delta G^{\text{bind}}_{\text{high}}$ changed by less than 0.3 kcal mol^{-1}. (The high hydrophobic term has more of an effect on their folding free energies than their binding free energies, favoring the larger side chains by up to 3 kcal mol^{-1} or more.) The insensitivity of our ligand design method to the choice of γ does not just show that the choice is unimportant; it also demonstrates the robustness of our method under perturbations of the energy function.

Table 6.9: Barstar "center 3" sequences found with good high-resolution binding and folding, using hydrophobic $\gamma = 22.8$ cal mol^{-1} Å^{-2}. Compare to Table 6.7 for $\gamma = 5$ cal mol^{-1} Å^{-2}.

$\Delta\Delta G^{\text{bind}}_{\text{high}}$	$\Delta\Delta G^{\text{fold}}_{\text{high}}$	Sequence		
(kcal mol^{-1})		35	38	73
Wild type:				
0.	**0.**	**D**	**W**	**V**
Good folders in order of binding:				
-1.72	-0.81			Q
-0.28	-0.88			I
-0.22	-1.95			M
-0.07	-0.05			L
0.	**0.**	**D**	**W**	**V**
2.56	-5.44			R
4.46	-0.25		F	R
4.76	-0.28		Y	R

6.4.2 Hierarchical Screening Protocol

In this section, we describe a protocol that uses the low-, medium-, and high-resolution energies as successive screens to reduce the number of high-resolution energy calculations required. The protocol was developed in an attempt to find all of the rotamer states with good high-resolution binding and folding without actually calculating all of the high-resolution energies. It is founded on the observed correlation, and lack of false negatives, of the low-, medium-, and high-resolution energies for the "center 3" redesign.

The correlation of G_{low} and G_{med} with G_{high} allows our algorithm to determine which rotamer states have low-resolution energies promising enough to warrant a medium-

Table 6.10: Barstar "center 3" sequences found with good high-resolution binding and folding, using hydrophobic $\gamma = 75$ cal mol^{-1} Å^{-2}. Compare to Table 6.7 for $\gamma = 5$ cal mol^{-1} Å^{-2}.

$\Delta\Delta G^{\text{bind}}_{\text{high}}$	$\Delta\Delta G^{\text{fold}}_{\text{high}}$	Sequence		
(kcal mol^{-1})		35	38	73
Wild type:				
0.	**0.**	**D**	**W**	**V**
Good folders in order of binding:				
-1.56	-2.88			Q
-0.51	-0.76			M
-0.26	-2.28			I
0.	**0.**	**D**	**W**	**V**
0.11	-2.13			L
2.59	-10.80			R
2.71	-1.71			K
4.10	-3.43		F	R

resolution energy calculation, and then which states have medium-resolution energies promising enough to warrant a high-resolution energy calculation.

Empirical Relationship of Medium- to High-Resolution Energies

Consider the data in Figure 6-13: the slanted line on the plot is an empirical lower bounding line; there are no points below the line, and therefore no false negatives, or structures with very poor $\Delta G^{\text{bind}}_{\text{med}}$ but very good $\Delta G^{\text{bind}}_{\text{high}}$. So with this data, to be sure to get all structures whose $\Delta G^{\text{bind}}_{\text{high}}$ is within, for example, 7 kcal mol^{-1} of the minimum (i.e., below the horizontal line segment), only structures with $\Delta G^{\text{bind}}_{\text{med}}$ below a cutoff (i.e., to the left of the vertical line) need to have their $\Delta G^{\text{bind}}_{\text{high}}$ calculated.

Based on the data in Figure 6-13, we chose the empirical bounding line for the high- vs. medium-resolution binding energies to have slope $m^{\text{bind}} = 0.4$ and to pass 2 kcal mol^{-1} below the lowest point. Similarly, based on the data in Figure 6-17, we chose the empirical bounding line for the high- vs. medium-resolution folding energies to have slope $m^{\text{fold}} = 0.8$ and to pass 5 kcal mol^{-1} below the lowest point.

Of course, to determine these empirical bounding lines, we calculated the high-resolution energies for all the same conformations as the medium-resolution energies. Now, for other DEE/A* redesign experiments, particularly those with more structures to consider, one could simply use this protocol to save on high-resolution energy calculation time.

Description of Hierarchical Screening Protocol

DEE/A* makes a list of rotamer states sorted by low-resolution binding energy (it is computationally inexpensive to make this list longer than needed, or to extend it if needed). Since the low-resolution energy function can favor over-charging the binding interface, we employ a simple method which quickly finds out if some values of the total charge get good low-resolution energies but poor high-resolution energies: the list of rotamer states found by DEE/A* is sorted into separate lists for each value of the total charge, and these list are then interleaved. The effect is to consider, for each value of the total charge, the one structure with the best low-resolution binding energy, then the second best structure for each charge, and so on.

For each rotamer state on this sorted and interleaved list, medium-resolution energies are calculated. Then, for the first structure considered, high-resolution energies are calculated immediately. For subsequent structures, however, high-resolution energies are only calculated if both the binding and folding medium-resolution energies pass cutoffs. These cutoffs are represented as vertical lines in Figures 6-13 and 6-17. Throughout the screening procedure, these cutoffs are updated whenever a new minimum high-resolution binding or folding energy is found.

In order to save on computation time for medium-resolution as well as high-resolution energies, several criteria are used to determine when a given value q of the total charge is not expected to have any more medium-resolution energies promising enough to merit a high-resolution energy calculation. After the criteria are met, all structures with that charge q are discarded. The criteria are (i.) that at least 20 structures with charge q in a row have had unpromising medium-resolution energies, (ii.) that at least 30% of all the structures with charge q considered so far have had unpromising medium-resolution energies, and (iii.) that we have proceeded up the list far enough that the low-resolution binding energy is at least 2 kcal mol^{-1} beyond that of the last structure which had promising medium-resolution energies. Pseudo-code for the detailed protocol is given in Figure 6-25.

Results of Screening Protocol on Barnase/Barstar "Center 3" Redesign

We chose to do the "center 3" system's redesign first, because it has few enough structures that low-, medium-, and high-resolution energies can be calculated for all structures. This allowed us to develop a screening protocol which can be applied to design experiments with many more structures.

We applied the screening protocol described above to the redesign of the "center 3" residues of barstar, setting for the protocol the goal of finding all structures within 7 kcal mol^{-1} of the minimum for both the binding and folding high-resolution energy. The choice of 7 kcal mol^{-1} was arbitrary, except that we wanted the the structures sought by the protocol to include the wild type. The performance of the protocol is depicted in Figure 6-26. In green are the 36 structures which we aimed to find with the protocol; i.e., those within 7 kcal mol^{-1} of the minimum for both the binding and folding high-resolution energy. The protocol only calculates high-resolution energies for the 1099 structures shown in black or green, and does not calculate them for the rest of the 4629 structures, shown in blue. So, the protocol significantly reduced the number of high-resolution energy calculation required in this case, from at least 4629 down to 1099, while still succeeding in finding the 36 structures of interest.

Validation of Screening Protocol on gp41 Redesign

Having determined the empirical bounding lines assumed by the screening protocol using data from the redesign of the "center 3" residues of barstar, we next chose an unrelated protein binding system in order to validate the screening protocol. This second system

Figure 6-25: Pseudo-Code for Hierarchical Screening Protocol

The goal is to find all rotamer states k with $\Delta G^{\text{bind}}_{\text{high}}(k)$ within c^{bind} (e.g., 7 kcal mol^{-1}) of the minimum, $\min(\Delta G^{\text{bind}}_{\text{high}})$; and $\Delta G^{\text{fold}}_{\text{high}}(k)$ within c^{fold} (e.g., 7 kcal mol^{-1}) of the minimum, $\min(\Delta G^{\text{fold}}_{\text{high}})$.

We found empirically that all points on a plot of $\Delta G^{\text{bind}}_{\text{high}}$ vs. $\Delta G^{\text{bind}}_{\text{med}}$ lie above a bounding line of slope $m^{\text{bind}} = 0.4$ passing $b^{\text{bind}} = 2$ kcal mol^{-1} below the point with minimum $\Delta G^{\text{bind}}_{\text{high}}$. So we do not need to consider rotamer states k with $\Delta G^{\text{bind}}_{\text{med}}(k) > \Delta G^{\text{bind}}_{\text{med, cutoff}}$, where

$\Delta G^{\text{bind}}_{\text{med, cutoff}} = (\Delta G^{\text{bind}}_{\text{med}}$ of state with minimum $\Delta G^{\text{bind}}_{\text{high}}) + \frac{c^{\text{bind}}+b^{\text{bind}}}{m^{\text{bind}}}$

We also found empirically that all points on a plot of $\Delta G^{\text{fold}}_{\text{high}}$ vs. $\Delta G^{\text{fold}}_{\text{med}}$ lie above a bounding line of slope $m^{\text{fold}} = 0.8$ passing $b^{\text{fold}} = 5$ kcal mol^{-1} below the point with minimum $\Delta G^{\text{fold}}_{\text{high}}$. So we do not need to consider rotamer states k with $\Delta G^{\text{fold}}_{\text{med}}(k) > \Delta G^{\text{fold}}_{\text{med, cutoff}}$, where

$\Delta G^{\text{fold}}_{\text{med, cutoff}} = (\Delta G^{\text{fold}}_{\text{med}}$ of state with minimum $\Delta G^{\text{fold}}_{\text{high}}) + \frac{c^{\text{fold}}+b^{\text{fold}}}{m^{\text{fold}}}$

```
1. Beginning with the list of structures from DEE/A*, sorted by low-resolution binding energy, break the list up by total charge, and interleave these lists.
2. Going through the list, for each rotamer state $k$:
    (a) Calculate its medium-resolution energies $\Delta G^{\text{bind}}_{\text{med}}(k)$ and $\Delta G^{\text{fold}}_{\text{med}}(k)$. Increment $n_{\text{calc, med}}(q(k))$, the total number of medium-resolution energies calculated so far for integer charge $q(k)$ of the mobile residues for state $k$
    (b) If its medium-resolution energies look promising; i.e. $(\Delta G^{\text{bind}}_{\text{med}}(k) \leq \Delta G^{\text{bind}}_{\text{med, cutoff}})$ and $(\Delta G^{\text{fold}}_{\text{med}}(k) \leq \Delta G^{\text{fold}}_{\text{med, cutoff}})$, then
        i. Calculate high-resolution energies $\Delta G^{\text{bind}}_{\text{high}}(k)$ and $\Delta G^{\text{fold}}_{\text{high}}(k)$.
        ii. Reset $n_{\text{misses}}(q(k)) = 0$, the number of medium-resolution energies in a row which have not warranted a high-resolution energy calculation.
        iii. Set $\Delta G^{\text{bind}}_{\text{low, last high}}(q(k))$, which is the $\Delta G^{\text{bind}}_{\text{low}}$ of the last rotamer state of charge $q(k)$ for which we have calculated high-resolution energies, to $\Delta G^{\text{bind}}_{\text{low}}(k)$.
        iv. Update $\Delta G^{\text{bind}}_{\text{med, cutoff}}$ with equation above if $\Delta G^{\text{bind}}_{\text{high}}(k)$ is the new minimum $\Delta G^{\text{bind}}_{\text{high}}$.
        v. Update $\Delta G^{\text{fold}}_{\text{med, cutoff}}$ with equation above if $\Delta G^{\text{fold}}_{\text{high}}(k)$ is the new minimum $\Delta G^{\text{fold}}_{\text{high}}$.
    (c) Otherwise, increment $n_{\text{misses}}(q(k))$.
    (d) Decide that it's time to stop looking at rotamer states of charge $q(k)$ if
        i. $(n_{\text{misses}}(q(k)) \geq n_{\min})$ and
        ii. $(n_{\text{misses}}(q(k)) > f_{\max} \cdot n_{\text{calc, med}}(q(k)))$ and
        iii. $\left[(\Delta G^{\text{bind}}_{\text{low}}(k) - \Delta G^{\text{bind}}_{\text{low, last high}}(q(k))) > \Delta\Delta G^{\text{bind}}_{\text{low, quit}}\right]$
```

where $\Delta\Delta G^{\text{bind}}_{\text{low, quit}} = 2$ kcal mol^{-1} is a minimum distance in $\Delta G^{\text{bind}}_{\text{low}}$ to proceed up the list, even if none of their medium-resolution energies prove promising enough to calculate high-resolution energies. $f_{\max} = 0.3$ is the minimum fraction of the rotamer states of a charge q considered so far that must have unpromising medium-resolution energies, before giving on charge q entirely. $n_{\min} = 20$ is the minimum number of unpromising medium-resolution energies of structures with a charge q in a row that we require before giving up on charge q entirely.

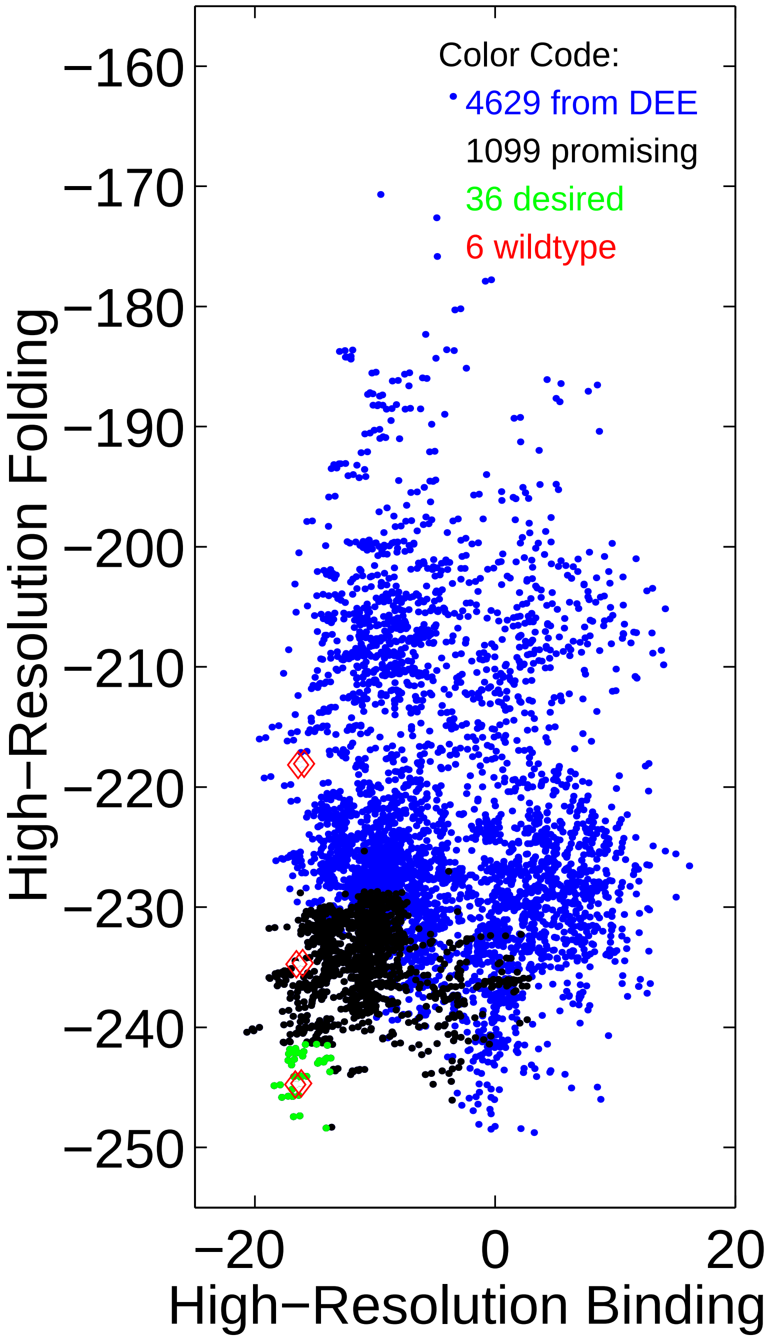

Figure 6-26: High-resolution $\Delta G^{\text{bind}}_{\text{high}}$ vs. medium-resolution binding $\Delta G^{\text{bind}}_{\text{med}}$ energy for 4629 mutant structures from DEE/A* on the "center 3" residues (Asp35, Trp38, Val73). Structures with the wild-type sequence are shown as red diamonds. The colors illustrate the protocol in Section 6.4.2: In green, the 36 structures which the protocol aimed to capture: those within 7 kcal mol^{-1} of the minimum of both the binding and folding high-resolution energy. The protocol decides to calculate high-resolution energies for the 1099 structures shown in black or green, and decides not to calculate them for the rest of the 4629 structures, shown in blue. The energy on each axis has an arbitrary constant term.

is a DEE/A* redesign of three residues of the D chain of gp41 (Trp2, Trp5, Asp6) to optimize its binding to the ABC 3-chain core. It was not computationally feasible for the screening protocol to find *all* conformations within 7 kcal mol^{-1} of the minimum high-resolution binding and folding energies, because the gp41 structures are more closely spaced in energy than those of the "center 3" system. The gp41 conformations are more closely spaced in energy because the binding interface is more open to solvent, and more hydrophobic. Within 30 kcal mol^{-1} of the minimum low-resolution binding energy, there are 76307 conformations , as compared to 4629 for the redesign of the "center 3" barstar residues. To limit computational expense, we only used the 9996 gp41 structures of the 1000 sequences within 19.0 kcal mol^{-1} of the minimum low-resolution binding energy. To validate the screening protocol, we ran it on the gp41 system, but we also calculated the high-resolution energies of all 9996 structures to check if the protocol missed any structures that it was supposed to find.

Data for high- vs. medium-resolution binding energies from this gp41 redesign are shown in Figure 6-27. We see that a line constructed as before, with a slope of $m^{\text{bind}} = 0.4$ passing through a point 2 kcal mol^{-1} below the lowest point, is a lower bounding line for this data as well. So the gp41 data obey the empirical bounding line determined from the "center 3" data.

Similarly, the folding energy data from this gp41 redesign is shown in Figure 6-28. A few of these gp41 structures fall slightly below the barnase/barstar "center 3" empirical bounding line.

Figure 6-29 shows that the protocol calculated high-resolution energies for 4699 out of 9996 structures, and successfully found all 185 of the desired structures (i.e., those within 7 kcal mol^{-1} of the minimum of both the binding and folding high-resolution energy). This validates the method.

By only taking the first 9996 structures with the best low-resolution binding energy, we stopped the protocol early; as the low-resolution binding energy goes up, fewer structures pass the screening protocol. Therefore, the time savings will generally be greater than

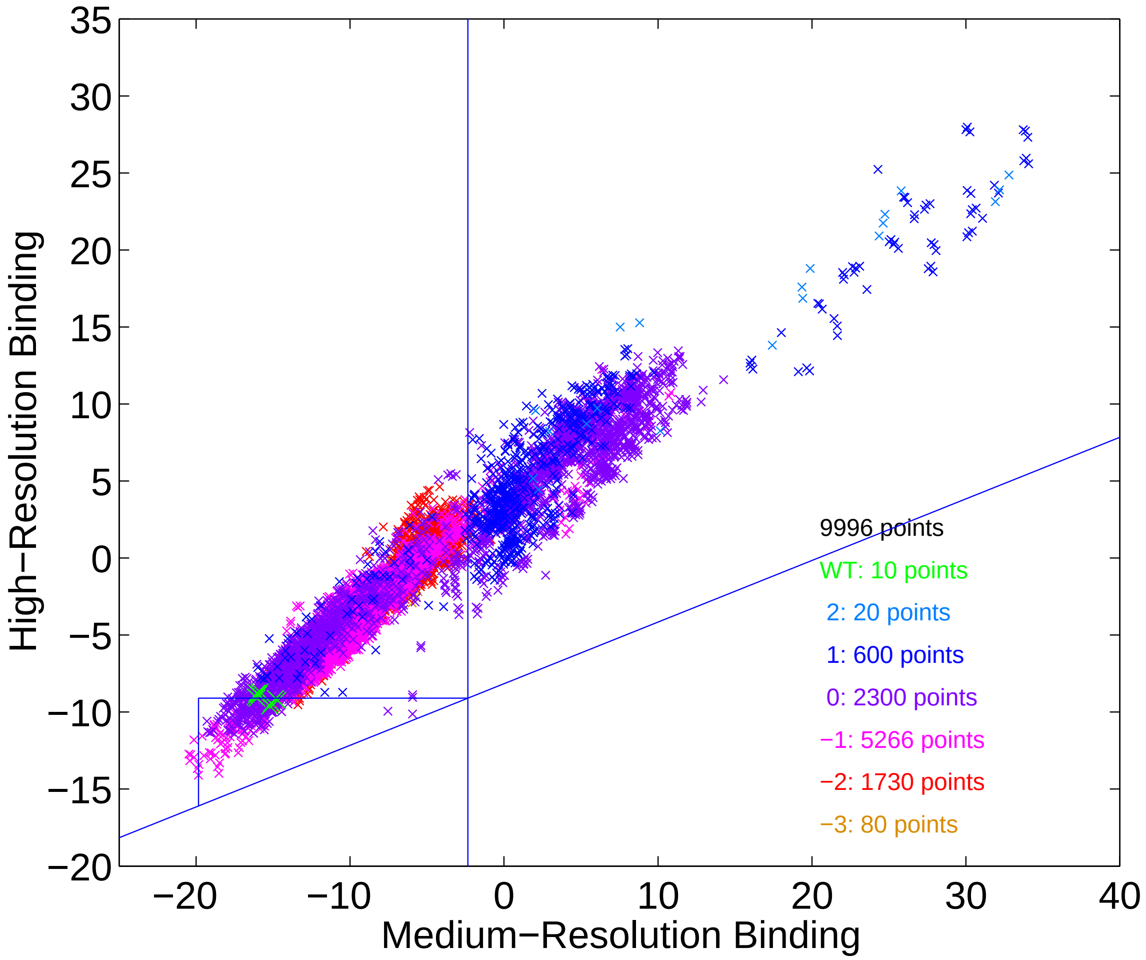

Figure 6-27: High-resolution $\Delta G^{\text{bind}}_{\text{high}}$ vs. medium-resolution binding $\Delta G^{\text{bind}}_{\text{med}}$ energy for 9996 mutant structures from DEE/A* on three residues of gp41 (Trp2, Trp5, Asp6) to optimize its binding to the ABC 3-chain core. Structures are colored by the charge of these three residues. Structures with the wild-type sequence are shown as green "X"s. The lines illustrate the verification of the protocol developed from in Section 6.4.2: The slanted line is the empirical lower bounding line determined from the barnase/barstar "center 3" DEE/A* data in Figure 6-13: the line with slope 0.4 passing 2 kcal mol^{-1} below the lowest point. There are no points below the line, and therefore no "false negatives". The figure has data for the 9996 structures with $\Delta G^{\text{bind}}_{\text{low}}$ within 19.0 kcal mol^{-1} of the minimum. The energy on each axis has an arbitrary constant term.

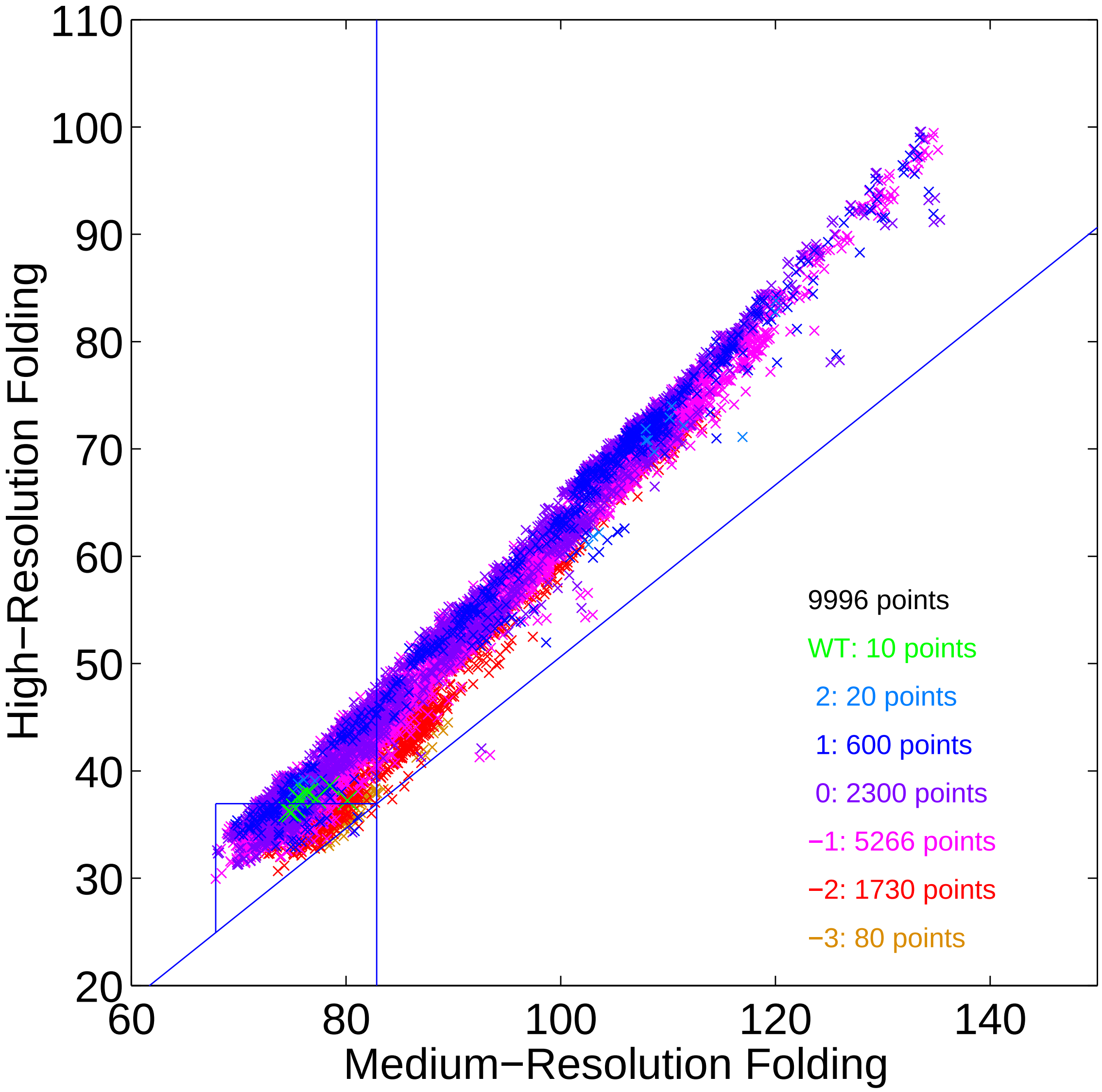

Figure 6-28: High-resolution $\Delta G^{\text{fold}}_{\text{high}}$ vs. medium-resolution folding $\Delta G^{\text{fold}}_{\text{med}}$ energy for 9996 mutant structures from DEE/A* on three residues of gp41 (Trp 2, Trp 5, Asp 6) to optimize its binding to the ABC 3-chain core. Structures are colored by the charge of these three residues. Structures with the wild-type sequence are shown as green "X"s. The lines illustrate the verification of the protocol developed from in Section 6.4.2: The slanted line is the empirical lower bounding line determined from the barnase/barstar "center 3" DEE/A* data in Figure 6-17: the line with slope 0.8 passing 5 kcal mol^{-1} below the lowest point. The energy on each axis has an arbitrary constant term.

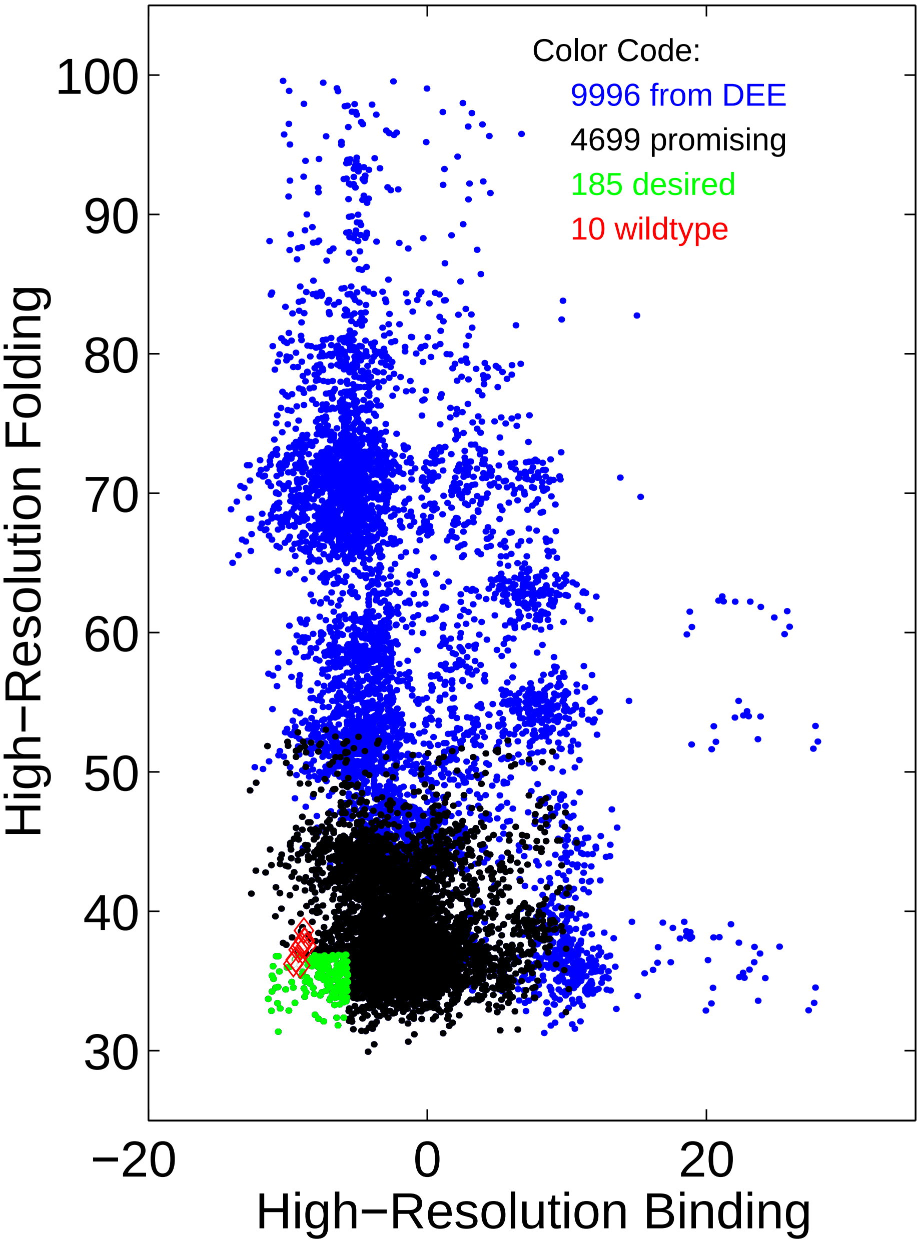

Figure 6-29: High-resolution $\Delta G^{\text{bind}}_{\text{high}}$ vs. medium-resolution binding $\Delta G^{\text{bind}}_{\text{med}}$ energy for 9996 mutant structures from DEE/A* on three residues of the D chain of gp41 (Trp 2, Trp 5, Asp 6) to optimize its binding to the ABC 3-chain core. Structures with the wild-type sequence are shown as red diamonds. The colors illustrate the protocol in Section 6.4.2: The 185 structures which we aimed for the protocol to capture are shown in green: those within $c^{\text{bind}} = c^{\text{fold}} = 7$ kcal mol^{-1} of the minimum of both the binding and folding high-resolution energy. The protocol decided to calculate high-resolution energies for the 4699 structures shown in black or green, and decided not to calculate them for the rest of the 9996 structures, shown in blue. The energy on each axis has an arbitrary constant term.

4699/9996.

6.4.3 Results for Barnase/Barstar "Lee 7" Redesign

We used our design methods to redesign the "Lee 7" residues of barstar (Asn33, Asp35, Trp38, Asp39, Thr42, Val73, Glu76) for optimal binding to barnase. There were 743780 structures representing 75148 sequences within 30 kcal mol^{-1} of the minimum low-resolution binding energy. The diversity of amino acids that these 75148 sequences had in the 7 positions is shown in Table 6.11. Figure 6-30 shows the low-resolution folding vs. binding energy for all 743780 structures.

Table 6.11: Amino Acid Frequency in the 75148 sequences within 30 kcal mol^{-1} of the minimum binding energy found by DEE/A* for the "Lee 7" barstar residues (Asn33, Asp35, Trp38, Asp39, Thr42, Val73, Glu76). The wild-type amino acids are shown in boldface.

Amino	barstar residue						
Acid	#33	#35	#38	#39	#42	#73	#76
ALA	781	16	704	10	1171	1854	650
ARG	6203		787		4863	2790	481
ASN	**3061**	484	2136	357	2755	4158	884
ASP	5956	**74388**	15423	**64265**	24948	9285	8242
CYS	986	42	826	29	1624	2050	744
GLN	1806		2816	44	2532	5471	3992
GLU	11483		12722	10213	10474	25279	**43418**
GLY	527	5	473	4	887	1773	613
HIS	4461		4820		2898	3464	1589
HSD	4411		4376		3409	1558	2140
ILE	2540		747	53	1250	2726	1110
LEU	1880		1927	2	1789	2887	1088
LYS	13936		784		2078	2062	437
MET	1696		2033	40	1803	3175	1435
PHE	3386		5098		1660		1518
SER	1202	55	1372	29	2615	2173	882
THR	1356	135	1439	71	**2515**	2259	952
TRP	4482		**9484**		2553		2117
TYR	3411		6879		1726		2032
VAL	1584	23	302	31	1598	**2184**	824

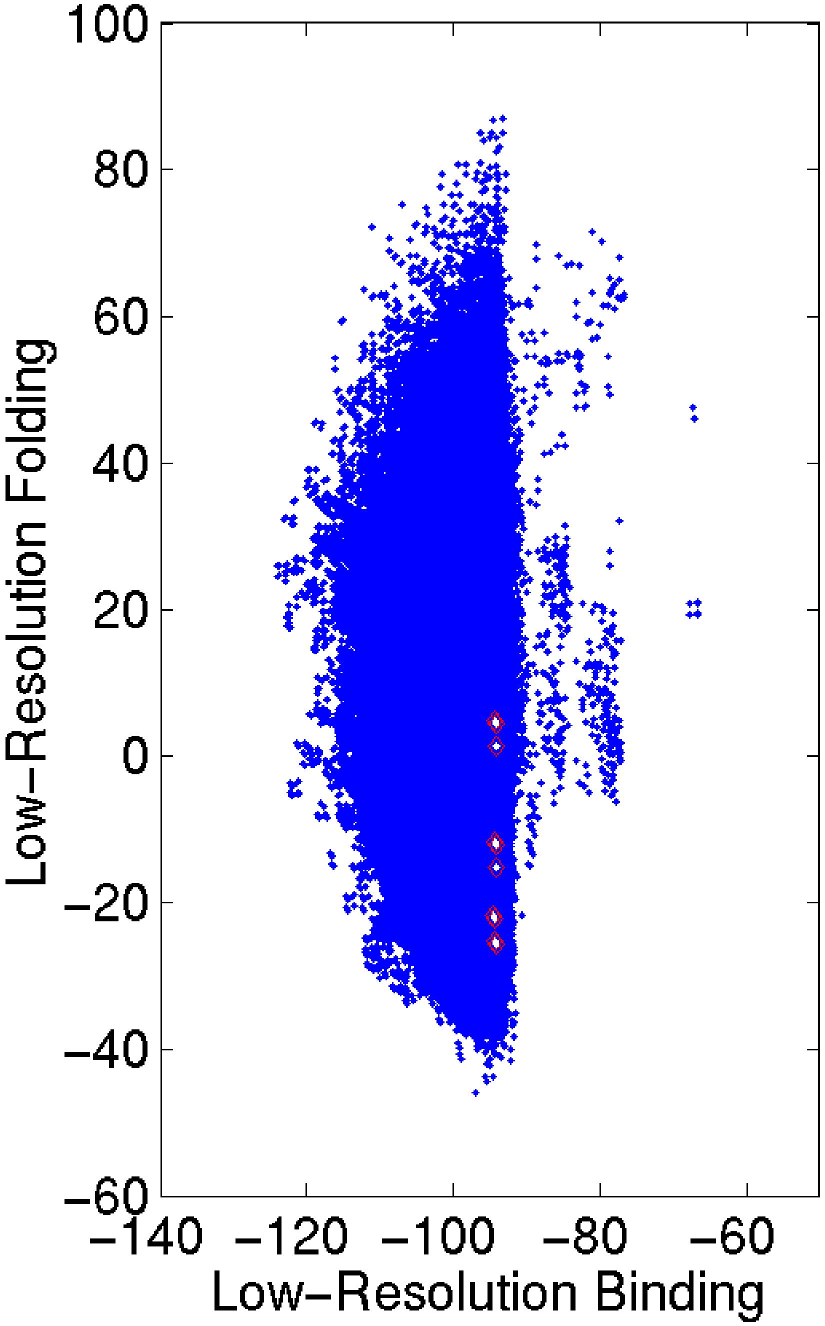

Figure 6-30: Low-resolution folding $\Delta G_{\text{low}}^{\text{fold}}$ vs. binding $\Delta G_{\text{low}}^{\text{bind}}$ energy for 743,780 mutant structures from DEE/A* on the "Lee 7" residues (Asn33, Asp35, Trp38, Asp39, Thr42, Val73, Glu76). Structures with the wild-type sequence are shown as red diamonds. The energy on each axis has an arbitrary constant term. Both dimensions are in kcal mol^{-1}.

With this many structures, it is not feasible to follow the protocol of Section 6.4.2; trying to find all structures with good high-resolution energies would be too computationally expensive. Happily, in ligand design, there is rarely any need to find *all* good ligands, if the good ligands that one *can* afford to find are interesting. We therefore followed a simpler procedure to find structures with low high-resolution binding and folding free energies. For all structures found by DEE/A* using low-resolution energies, medium-resolution energies were calculated. Figure 6-31 shows folding vs. binding energy for the medium-resolution energy function, for all 743,780 structures.

Wild-Type Sequence Structures

Of the 10 structures representing the wild-type sequence (shown in Figure 6-32), the structure most similar to the crystal structure is one of the 3 that are approximately tied for best medium-resolution binding energy, as shown in Figure 6-31. Among those 3 structures, it is the one with the best medium-resolution folding energy by almost 10 kcal mol^{-1}. (The same is true of their high-resolution binding and folding energies.) This is the structure that we chose to represent the wild-type sequence. The only significant difference it has from the crystal structure is that its Glu76 side chain orients to make a bidentate interaction with barnase Arg59. All 10 of the structures representing the wild-type sequence have this bidentate conformation, as do almost all of the other designed structures with Glu at position 76. All 3 crystallographic subunits of PDB code 1BRS have the same conformation of Glu76, but two other crystal structures of the barnase/barstar complex, PDB codes 1B27 and 1B3S, each have 1 of 3 crystallographic subunits with Glu76 in the bidentate conformation relative to barnase Arg59.

If we rank the sequences by the best binding energy among their structures, then as we go from low- to medium-resolution binding energy functions, the ranking of the wild-type sequence's binding energy improves from #64447 to #5329. This demonstrates that the medium-resolution energy function, which uses the ACE electrostatics treatment, does much better at picking out the wild type as a tight-binding sequence. The ranking

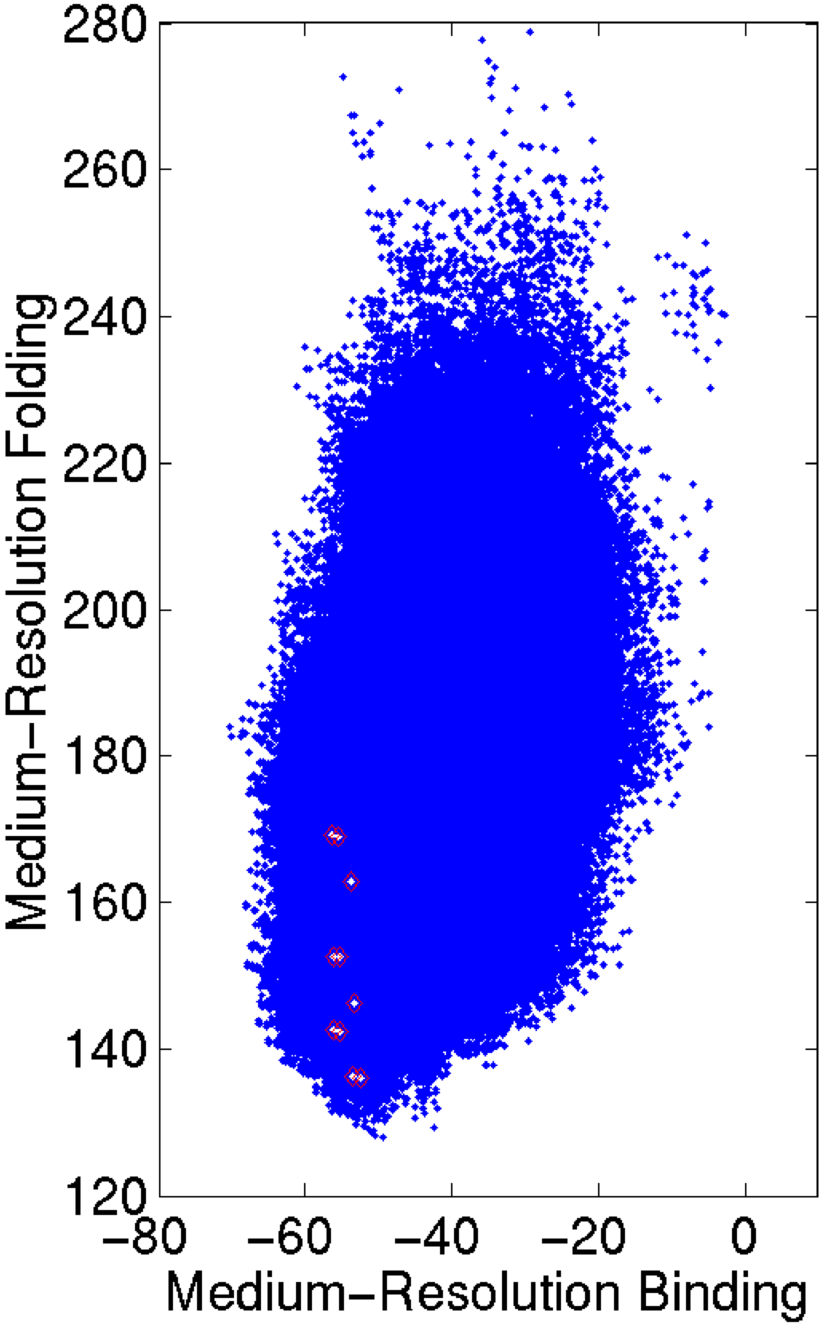

Figure 6-31: Medium-resolution folding $\Delta G^{\text{fold}}_{\text{med}}$ vs. binding $\Delta G^{\text{bind}}_{\text{med}}$ energy for 743,780 mutant structures from DEE/A* on the "Lee 7" residues (Asn33, Asp35, Trp38, Asp39, Thr42, Val73, Glu76). Some points with very poor $\Delta G^{\text{fold}}_{\text{med}}$ fall above the rectangular area shown. Structures with the wild-type sequence are shown as red diamonds. The energy on each axis has an arbitrary constant term. Both dimensions are in kcal mol^{-1}.

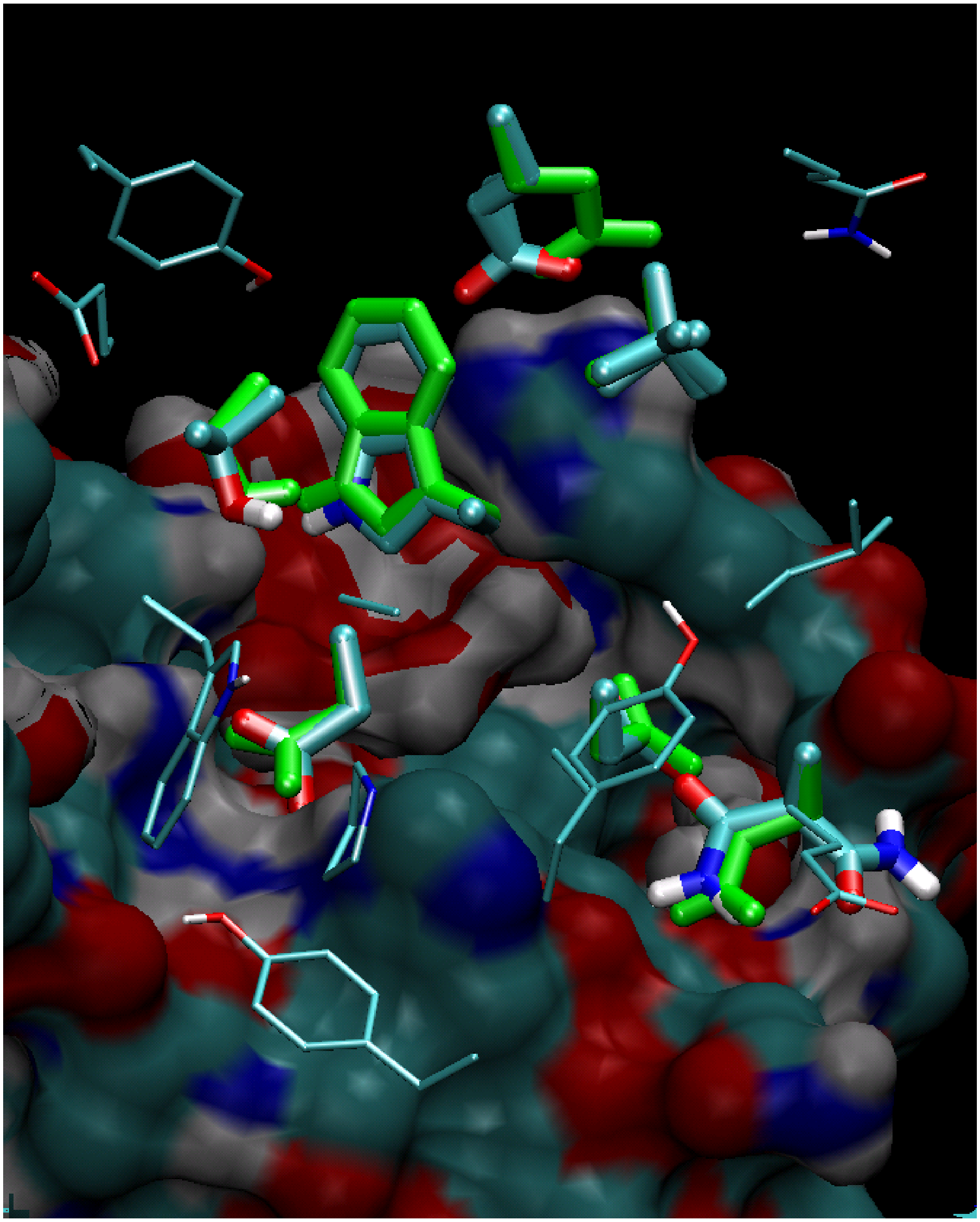

Figure 6-32: DEE/A* Redesign of “Lee 7” Residues: Structures with Wild-type Sequence NDWDTVE. Barnase and the interfacial waters are shown as surfaces colored by element. We are looking through barnase, which is invisible except for side chains that bury solvent-accessible surface area upon binding. The “Lee 7” barstar residues (Asn33, Asp35, Trp38, Asp39, Thr42, Val73, Glu76) are shown in fat licorice. The crystal structure positions are shown in green. The 10 DEE/A*-designed structures with the wild-type sequence are colored by element. They are very much like the crystal structure, except that all of them have Glu76 in a bidentate orientation to barnase Arg59, some of them have Asn33 flipped out of its pocket, and some of them have Val73 rotated.

of the wild-type sequence's folding energy improves from #2034 to #217. So, in this case, it appears that the medium-resolution folding energy is more than 24 times better (5329/217) at identifying the wild type than the medium-resolution binding energy. These numbers show how important the folding free energy can be as a discriminator of native-like tight-binding structures.

The disparity in how the wild-type sequence is ranked by the low- and high-resolution energy functions also demonstrates the value of our search procedure, because the structures with the wild-type sequence are far down on the list sorted by low-resolution binding energy, but our procedure has kept the list short enough that they can be found in a feasible amount of time. By using DEE/A* first to rank sequences, we get enough diversity of sequence so that sequences which the high-resolution energy function prefers can make it onto the long list of sequences. Our choice to keep only 10 structures from the second and third levels of DEE/A* has proven to be a good balance between sampling a diversity of structures and keeping the total number of structures low enough that the medium- and high-resolution energy functions can sort through them in a feasible amount of time.

Prediction of Stable Tight-Binding Structures

Having identified the wild-type-sequence structure found by DEE/A* which most closely resembles the crystal structure, we use its binding and folding energies as benchmarks for the other structures. For each structure with medium-resolution binding energy no worse than wild type and medium-resolution folding no more than one kcal mol^{-1} worse than wild type, high-resolution energies were calculated for all structures (there are up to 10) with the same sequence.

The final result of this design procedure is the small collection of promising sequences shown in Table 6.12, a ranking by high-resolution binding energy of each sequence's best high-resolution binder with high-resolution folding energy not more than one kcal mol^{-1} worse than wild type. The wild-type sequence does very well, ranking #89 out of

$1.28{\times}10^9$ possible sequences (20^7). Each sequence is represented by only one sequence at this final stage of the screening, so this simplified version of our protein design method has ranked a wild-type-like structure in its top 100 out of $3{\times}10^{24}$ total conformations (3098^7).

Promising Non-Wild-type Sequences

The non-wild-type sequences in Table 6.12 that appear promising largely owe their advantage to a few single-position mutations: Val73Gln, Val73Glu, and Asn33Leu. Many of the sequences in Table 6.12 are actually combinations of one of these favorable mutations and one or more other slightly unfavorable mutations.

Val73Gln Mutation

The structure of this single-position mutant is shown in Figure 6-33. The polar Gln73 makes a nice hydrogen bond with barnase Arg59, improving binding by -1.76 kcal mol^{-1} while not affecting the folding free energy ($+0.05$ kcal mol^{-1}), very much as we found when we redesigned the "center 3" residues (Asp35, Trp38, Val73) in Section 6.4.1.

Val73Glu and Asn33Leu Mutations

Although these mutations appear promising in Table 6.12, minimization reveals that their folding free energies before minimization have an unfair and arbitrary advantage. We discuss this in the next section.

Table 6.12: Barstar “Lee 7” sequences found with good high-resolution binding and folding. (For each sequence, the structure with lowest high-resolution binding free energy, which also has high-resolution folding free energy no more than one kcal mol^{-1} worse than wild type, is used.) They are sorted by high-resolution binding energy. All binding energies are shown relative to the wild-type binding energy $\Delta G^{\text{bind}}_{\text{high}}$(wildtype). Likewise, the folding energies are shown relative to the wild-type folding energy. The wild-type sequence is shown in boldface; we curtail the list at the wild type, but it is also shown at the beginning of the table for reference. For clarity, every position which matches the wild type is shown as a blank space.

$\Delta\Delta G^{\text{bind}}_{\text{high}}$	$\Delta\Delta G^{\text{fold}}_{\text{high}}$	Sequence						
(kcal mol^{-1})		33	35	38	39	42	73	76
Wild type:								
0.	**0.**	**N**	**D**	**W**	**D**	**T**	**V**	**E**
Good folders in order of binding:								
-3.74	-1.04	L				S	E	
-3.62	-3.07	L					E	
-3.11	0.83						E	
-2.74	-0.30	F				S	E	
-2.65	-0.03	L				C	E	
-2.63	-2.33	F					E	
-2.52	-2.23	K				S	E	
-2.41	-0.30	Y					E	
-2.40	-4.26	K					E	
-2.20	-1.11	L				F	E	
-1.88	-3.27	R				S	E	
-1.76	-5.30	R					E	
-1.76	0.05						Q	
-1.71	-4.77	L				I	E	
-1.57	0.25	L				A	E	
-1.53	-0.40	F				M	E	
-1.37	-1.23	E					E	
-1.37	-3.27	V				S	E	
-1.25	-5.30	V					E	
-1.23	-3.34	F				L	E	
-1.21	-0.85	C				S	E	
-1.21	0.12	S				S	E	
-1.17	-5.22	L				S	Q	
-1.12	-0.40	F				Q	E	

continued ↪

$\Delta\Delta G^{\text{bind}}_{\text{high}}$	$\Delta\Delta G^{\text{fold}}_{\text{high}}$	Sequence						
(kcal mol^{-1})		33	35	38	39	42	73	76
Continued:								
-1.09	-2.88	C					E	
-1.09	-1.91	S					E	
-1.01	-1.31	Y				L	E	
-1.01	0.05	K				M	E	
-1.00	-5.28	K				L	E	
-0.89	-2.34	K				Q	E	
-0.80	-2.27	R				C	E	
-0.79	-1.48					S	Q	
-0.77	-0.69	Y				M	E	
-0.76	-1.38	A					E	
-0.73	-2.26	Y				Q	E	
-0.72	-1.49	M				S	E	
-0.71	-4.03	F				I	E	
-0.69	0.54	W					Q	
-0.66	-3.38	R				M	E	
-0.61	-3.51	M					E	
-0.55	-2.94	Q				S	E	
-0.50	-2.00	Y				I	E	
-0.49	-5.96	K				I	E	
-0.44	-4.96	Q					E	
-0.40	-0.56					S	I	
-0.39	-1.66	L				S	N	
-0.36	-6.32	R				L	E	
-0.36	-1.07					L	Q	
-0.36	-0.55						M	
-0.35	-3.34	R				F	E	
-0.34	-0.25	T				S	E	
-0.30	0.75						L	
-0.30	-0.40						I	
-0.29	-2.26	V				C	E	
-0.27	0.70	E				M	E	
-0.26	-3.38	R				Q	E	
-0.22	-2.28	T					E	
-0.17	0.52	F				N	Q	
-0.15	-3.37	V				M	E	
-0.15	-4.52	F				S	Q	
-0.11	-0.45					M	Q	

continued ↪

$\Delta\Delta G^{\mathrm{bind}}_{\mathrm{high}}$	$\Delta\Delta G^{\mathrm{fold}}_{\mathrm{high}}$	Sequence						
(kcal mol^{-1})		33	35	38	39	42	73	76
Continued:								
-0.04	-6.55	F					Q	
-0.01	-0.29	R		Y			E	
0.	**0.**	**N**	**D**	**W**	**D**	**T**	**V**	**E**
…								

Minimization

For every structure with high-resolution folding energy no more that one kcal mol^{-1} worse than wild type, we performed minimization as described previously on all structures of the same sequence. This was a total of 3390 structures of 339 sequences. The energy function used was the low-resolution bound state energy, but using full van der Waals radii. Only atoms of the mobile side chains were allowed to move.

High-resolution energies were then calculated for all minimized structures, and the sequences were screened using the high-resolution energies of the minimized structures. Each sequence was represented by its best binding structure which folds no more than one kcal mol^{-1} worse than wild type. The most promising structures are shown in Table 6.13, which should be compared to the same list from before minimization, Table 6.12.

Since we only minimized a small subset of the structures, including the promising sequences before minimization, it is not surprising that the list of promising sequences gets shorter after minimization: sequences initially on the list can drop off after minimization, but we did not give all the other sequences a chance to be promoted onto the list. If this were done, we would at least expect the minimization procedure to promote the Val73His mutant found by the “center 3” redesign experiment onto the list of promising structures.

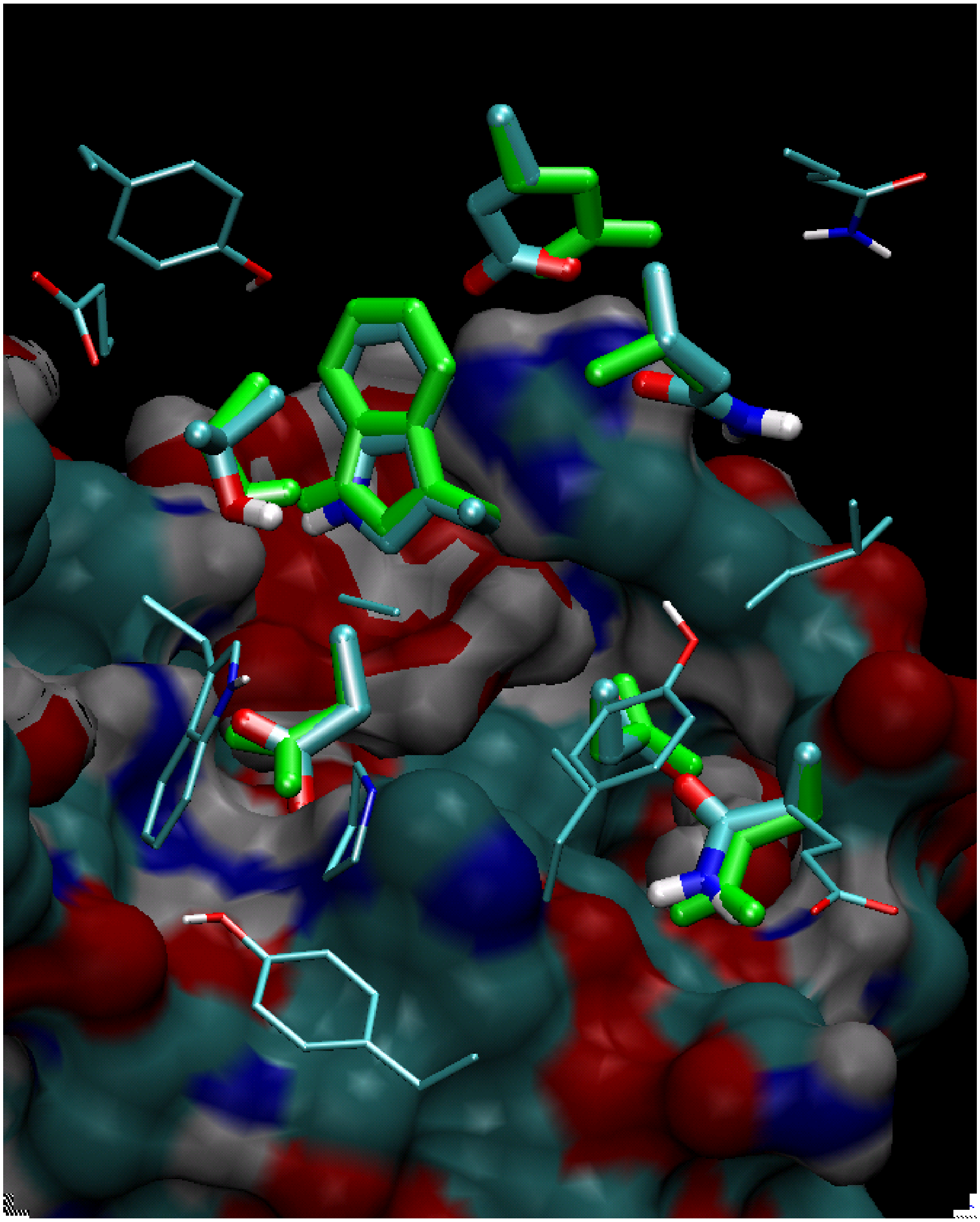

Figure 6-33: DEE/A* Redesign of "Lee 7" Residues: Best Structure with Sequence NDWDTQE. This sequence differs from the wild type by the mutation Val73Gln. Barnase and the interfacial waters are shown as surfaces colored by element. We are looking through barnase, which is invisible except for side chains that bury solvent-accessible surface area upon binding. The "Lee 7" barstar residues (Asn33, Asp35, Trp38, Asp39, Thr42, Val73, Glu76) are shown in fat licorice. The crystal structure positions are shown in green. The DEE/A*-designed structure with sequence NDWDTQE is colored by element. Both Gln73 and Glu76 interact with barnase Arg59 (visible here as the 3 blue patches arranged in a triangle at the top of the barnase surface).

Table 6.13: Barstar "Lee 7" sequences found with good high-resolution binding and folding after minimization. Minimization followed by high-resolution energy calculation was done on 3390 conformations of the 339 sequences listed in Table 6.12. They are sorted by high-resolution binding energy after minimization, relative to the wild type $\Delta\Delta G^{\text{bind,mini}}_{\text{high}} = \Delta G^{\text{bind,mini}}_{\text{high}} - \Delta G^{\text{bind,mini}}_{\text{high}}(\text{wildtype})$. The high-resolution folding energy is also shown relative to the wild type. The wild-type sequence is shown in boldface; it appears in the list, but is also shown at the beginning of the table for reference. Note that "H" and "h" are HIS and HSD, the PARAM19 forms of histidine protonated on the $N_{\delta 1}$ and the $N_{\epsilon 2}$ atoms, respectively. For clarity, every position which matches the wild type is shown as a blank space.

$\Delta\Delta G^{\text{bind,mini}}_{\text{high}}$	$\Delta\Delta G^{\text{fold,mini}}_{\text{high}}$	Sequence						
(kcal mol^{-1})		33	35	38	39	42	73	76
Wild type:								
0.	**0.**	**N**	**D**	**W**	**D**	**T**	**V**	**E**
Good folders in order of binding, after minimization:								
-2.44	0.02						Q	
-1.67	-0.90					L	Q	
-1.24	-2.35					Q	Q	
-1.02	0.11	W				Q	Q	
-1.01	-0.04					M	Q	
-0.97	-1.46					I	Q	
-0.36	0.78	H				Q	Q	
-0.34	0.40						M	
-0.22	-1.94						I	
-0.18	-0.30					S	I	
0.	**0.**	**N**	**D**	**W**	**D**	**T**	**V**	**E**
...								

After minimization, the sequences with the mutations Asn33Leu and Val73Glu are no longer on the list of promising sequences.

Why is the Asn33Leu mutation no longer on the list of promising structures after minimization? DEE/A* chose rotamers for Asn33 that have a slight van der Waals clash (about a 10 kcal mol^{-1} penalty) with their own molecule, barstar. Recall that we chose to have DEE/A* rank on the binding energy, and disqualify any rotamer or rotamer pair whose folding energy term is above a cutoff, 25 kcal mol^{-1}. So, in this case, DEE/A* chose Asn33 rotamers unfavorable for folding, but not so unfavorable as to be disqualified by the cutoff. This unfairly penalized the folding energies of all sequences with the wild-type residue Asn33, which made many sequences in Table 6.12 with the Asn33Leu mutation incorrectly appear to be as stable as the wild type. The minimized structures reveal that the wild-type Asn33 residue is much more beneficial for folding stability than any other residue at that position.

Why is the Val73Glu mutation no longer on the list of promising structures after minimization? The reason is essentially the same as in the previous explanation, but slightly more complicated. Asn33 occupies one of two rotamers in most of the structures found by DEE/A*. These two Asn33 rotamers are nearly tied in their contribution to the binding energy, so one of them is chosen by a small margin for some sequences, and the other is chosen by a small margin for other sequences. It happens that the sequences with a Val73Glu mutation choose the Asn33 rotamer that is better for folding than the other rotamer, by -3.5 kcal mol^{-1}. So sequences with the Val73Glu mutation incorrectly seemed to have better folding energies than sequences with Val73.

Both of these cases reveal that it could be beneficial to use the folding energy more directly in the DEE/A* ranking of structures.

6.5 Future Directions

In general, molecules can change conformation upon binding. The present method could be extended to allow this by this simple procedure: Use DEE/A* to redesign the bound complex and each unbound molecule separately for maximum stability. (This is straightforward using the present method; one can do two redesign runs to optimize $G^{\text{bound}} - G^{\text{unfolded}}$ and $G^{\text{unbound}} - G^{\text{unfolded}}$.) These could be used to calculate, for each sequence, a non-rigid binding free energy: the free energy difference of the most stable bound configuration versus the most stable unbound configuration. The final analysis of the sequences could use both this non-rigid binding free energy and the folding free energy.

As we concluded in the previous section, it could be beneficial to use the folding energy more directly in the DEE/A* ranking of structures. For example, ranking on $\Delta G^{\text{fold}}_{\text{low}} + \Delta G^{\text{bind}}_{\text{low}}$ rather than $\Delta G^{\text{bind}}_{\text{low}}$ would be one way to design structures that have good binding *or* folding. Then the structures could be screened to select those with good binding *and* folding.

Including a pairwise approximation of solvation in the low-resolution electrostatic energy function could increase the accuracy of the first stage of DEE/A*. The Generalized Born approximation of the atomic interactions could be used, but to make it pairwise, a method could be developed to assign each atom one approximate value for its solvation radius, regardless of the rotamers placed at the mobile residues.

A rotamer library should be developed which uses finer sampling of the dihedral angles of histidine, tryptophan, and phenylalanine, for example, which have long bulky shapes, but only two dihedral angles. Using our screening protocol on a combined set of minimized and unminimized structures could help to overcome the coarseness of the rotamer library while still preserving the wider range of structures before minimization.

When redesigning residues, it would be advantageous, in addition to giving full mutational freedom to these residues, to also give conformational freedom to some neighboring residues.

When cutoffs are applied to the self and pair terms of the low-resolution energies, it would be wiser to set the cutoff a given distance above the minimum value for that term across all rotamers. Currently, we set these cutoffs at a given absolute value.

Although we did not treat salt in the present study, it is trivial to set the ionic strength (to a physiological 0.145 M, for example) in both the medium- and high-resolution electrostatic energy functions.

6.6 Conclusion

Our protein design method is novel in that (1) it optimizes the binding free energy while maintaining a stable folding free energy, (2) it uses three stages of DEE/A* searching in order to get a diversity of sequences as well as multiple structures for each sequence, (3) it includes more accurate solvation and electrostatic free energy terms, and (4) it uses a hierarchy of three energy functions to successively screen candidate structures.

We redesigned three residues of gp41 (Trp2, Trp5, Asp6), three residues of barstar (Asp35, Trp38, Val73), and a larger set of seven residues of barstar (Asn33, Asp35, Trp38, Asp39, Thr42, Val73, Glu76). We found that the low-resolution binding free energy function ranks the barstar wild-type sequence, a known tight binder to barnase, much less favorable than the high-resolution function does. The low-resolution binding free energy naively favors sequences which are as negatively charged as possible. However, the higher-resolution energy functions do not make this mistake, because they include the desolvation and indirect interaction terms of the electrostatic binding free energy. Our search procedure finds the wild type, even though the low-resolution free energy function scores it poorly, by ensuring that a diversity of sequences is passed on to the higher-resolution free energy functions.

The wild-type sequence, experimentally known to be a very tight binder, ranks #5 out of 8000 for the barstar three-residue redesign and #89 out of 1,280,000,000 for the seven-residue redesign. The folding free energy was also found to be very favorable for the wild type compared to other sequences; for example, the wild-type sequence ranks #217 in medium-resolution folding free energy, but only #5329 in the corresponding binding free energy. So, in this case, it appears that the medium-resolution folding energy is more than 24 times better at identifying the wild type than the medium-resolution binding energy. These numbers show how important the folding free energy can be as a discriminator of native-like tight-binding structures. The conformations found by our search procedure for the barstar wild-type sequence are very similar to the crystal structure, which is, of course, experimentally known to be the stable conformation.

We do not incorporate minimization of structures into our procedure, but the minimization done in this study highlighted two improvements needed by our method. The DEE/A* structures for the single-position mutant Val73His are incorrectly given poor folding free energies, but this is corrected by minimization. The coarseness of the rotamer library prevents a histidine at this position from fitting properly to avoid steric clashes. His, Trp, and Phe may be particularly vulnerable to such problems because they have long bulky shapes but only two flexible dihedral angles. Minimization also revealed that ranking only on the binding free energy during DEE/A* can cause variations in structures' relative folding free energies which have no physical basis. Therefore it should be beneficial to use the folding energy more directly in the DEE/A* ranking of structures; for example, by ranking on $\Delta G_{\text{low}}^{\text{fold}} + \Delta G_{\text{low}}^{\text{bind}}$ rather than $\Delta G_{\text{low}}^{\text{bind}}$.

The disadvantages of minimization, as we have implemented it, are that it uses a low-resolution energy function, that it minimizes the bound state energy rather than the binding or folding energies, and that it causes many conformations of a residue to all minimize to the same conformation, reducing the diversity of the structures available to pass on to higher-resolution energy functions.

The analytical continuum electrostatics (ACE) method, which we used in our medium-resolution energy function, has proven invaluable as a rapid screening function which includes electrostatic solvation effects.

The single-position barstar mutant Val73Gln is predicted to bind about 2 kcal mol^{-1}

more tightly than the wild type, and to have about the same folding stability as the wild type. The optimum atomic charges on Val73 suggest that this side chain should be polar to optimize its binding free energy. Our design method improves on that charge optimization method by using different structures, and therefore different dielectric boundaries, appropriate to each sequence. A second single-position barstar mutant, Val73His, is predicted to bind about 1 kcal mol^{-1} more tightly than the wild type, and to have about the same folding stability as the wildype. Both of these mutants, Val73Gln and Val73His, are promising candidates for synthesis.

Appendix A

Protein Design: Treatment of Crystallographic Water Molecules

A.1 Methods

Many X-ray crystal structures of binding complexes, including barnase/barstar, have water molecules at well-defined positions in the binding interface. When redesigning for tight binding, one strategy would be to remove some water molecules from the bound conformation, in order to allow side chains to occupy the space instead, or to allow larger side chains to fit. The complex of barnase and barstar has a particularly high number of interfacial water molecules; it is commonly believed that this is the result of a trade-off that evolution has made to gain binding speed at the cost of binding affinity. It would also be advantageous for any ligand design effort with interfacial water molecules to allow them to rotate or move in order to interact as well as possible with each candidate structure.

We developed a method to allow the interfacial water molecules more freedom by treating each of them just like a mobile residue, with a discrete set of rotamers. The water rotamer set we used has 12 conformations, plus the choice of removing the water molecule, which is treated as a 13th conformation.

A.1.1 Allowed Conformations for Mobile Water Molecules

The 12 conformations of each water molecule are determined based on the position of its oxygen atom in the wild type crystallographic structure, and the default positions of its two hydrogen atoms as determined by the HBUILD facility [90] of the molecular modeling package CHARMM [11]. Consider the oxygen atom as the center of a cube, with the 2 hydrogen atoms approximately at 2 corners of the cube separated by a face diagonal. Allowing the 2 hydrogen atoms to be on any pair of cube corners separated by a face diagonal results in 12 conformations. We forced the 8 possible hydrogen positions to be exactly on the corners of a cube, which forces the hydrogens to move from their CHARMM HBUILD positions by about 0.1 Å, stretching the water bond angle from 104.52° to the exact tetrahedral angle, 109.471°. (Allowing the rotamers to have the correct 104.52° angle made very little difference in the relative energies of the rotamers or in the outcome of a DEE/A* run.)

A.1.2 Defining Sequence, Fleximers, and Rotamers for Mobile Water Molecules

We chose to treat each of these 12 conformations, plus the 13th state for absence of the water molecule, as fleximers consisting of 1 rotamer each. Splitting the states at the fleximer level rather than the "sequence" level or the rotamer level strikes a good balance by allowing the waters to try different conformations in each of the 10 fleximer states we keep for each sequence, while not making the total number of states any longer than it is without freedom for the waters, i.e. 10 states per amino acid sequence.

Treating each of the 13 water states as different "sequences" is very computationally expensive because rather than a ranked list of protein sequences, DEE/A* gives a much longer list, because every possible conformation of the waters is considered a different "sequence". Treating only the presence or absence of each water as 2 different "sequences" still lengthens the list of sequences made by DEE/A*, by up to $2^{N_{\text{water}}}$ for N_{water} mobile waters, and promises little benefit over the simpler treatment which we use.

A.2 Results and Discussion

A.2.1 Choosing Protein Residues to Redesign along with Waters

To find barstar residues to redesign along with some explicit water molecules, we did DEE/A* redesigns for each individual barstar residues that has any solvent-accessible surface area (SASA) burial upon binding. These were compared to DEE/A* redesigns of each of these 18 barstar residues with all crystallographic water molecules removed from the structure. For each residue, the difference between these results (the details are not shown here) indicates whether the presence of the waters has any effect on the possible conformations of that residue. We found very few obviously good conformations made possible by giving the waters the freedom to rotate or vanish. Most of the residues (Gly27, Tyr29, Tyr30, Glu32, Leu34, Ala36, Ala40, Glu46, Gln72, Val73) were not significantly affected by the presence of the waters. Five of the residues (Asn33, Thr42, Trp44, Tyr47, Glu76) were somewhat affected, but no promising new conformations were made possible by the removal of all of the waters. Only three of the residues (Asp35, Trp38, Asp39) had amino acid types with promising conformations made possible by the removal of the waters. With the waters removed, Asp35 is able to mutate to Glu, Trp38 is able to mutate to Arg or Lys, and Asp39 is able to mutate to Leu. Based on this analysis, and the proximity of the residues to each other, we chose to redesign these two sets of residues, pictured in Figure A-1:

1. Asn33, Asp35, and 3 neighboring waters.

2. Trp38, Asp39, Thr42, and 8 neighboring waters.

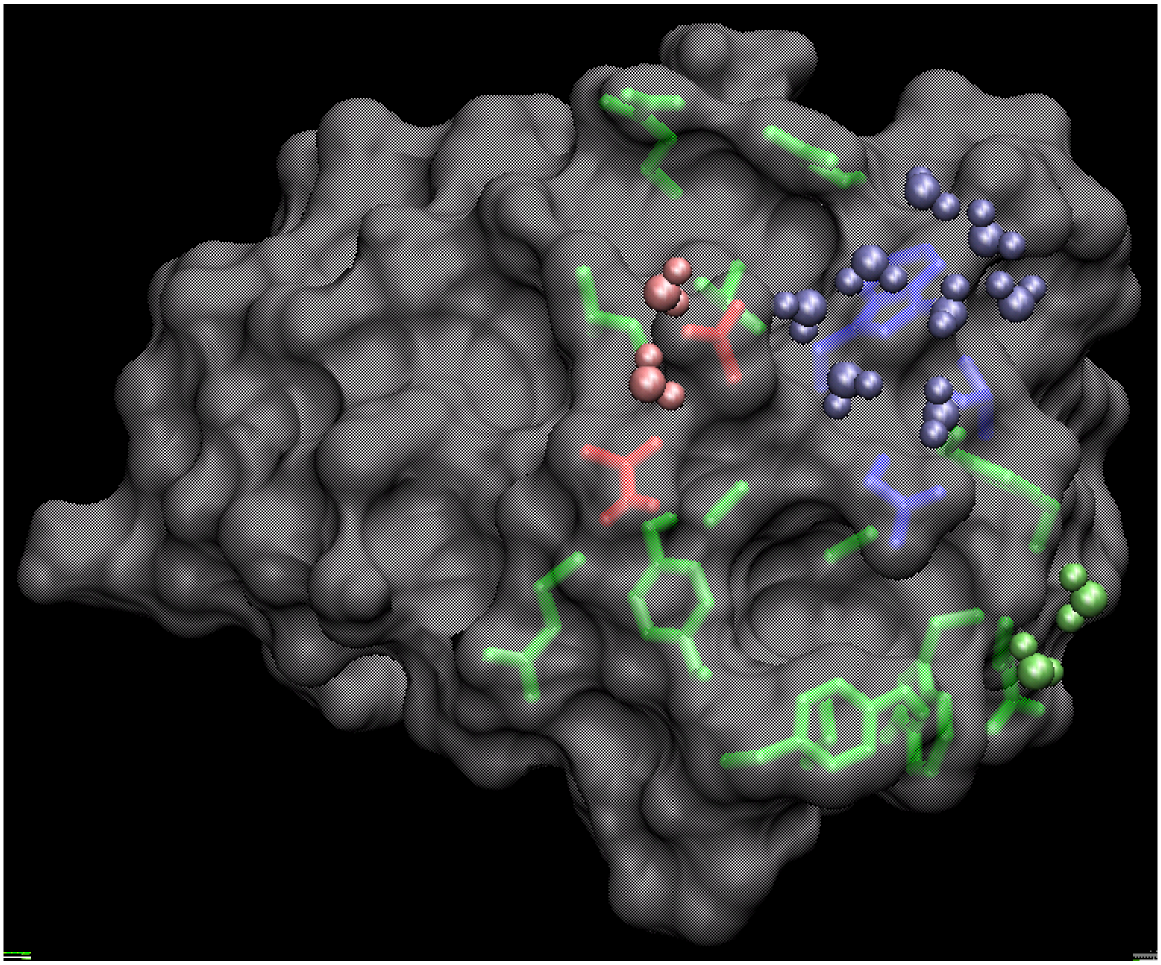

Figure A-1: Two Sets of Barstar Residues to Mutate Along with Waters. All barstar side chains that bury solvent-accessible surface area (SASA) upon binding to barnase, shown as licorice, can be seen through the translucent gray surface of barstar. Interfacial water molecules are shown as small atomic balls. Asn33, Asp35, and 3 waters numbered 22, 29, and 128 are shown in red. Trp38, Asp39, Thr42, and 8 waters numbered 14, 29, 33, 48, 56, 60, 116, and 361 are shown in blue. The other waters, and the other barstar side chains that bury solvent-accessible surface area upon binding to barnase, are shown in green. Barnase is not shown; it would be in front of barstar in this view.

A.2.2 Results for Redesign of Barstar Asn33, Asp35, and 3 Waters

Our full design procedure was applied to barstar Asn33, Asp35, and the 3 water molecules numbered 22, 29, and 128. Actually, since there were only 197 sequences within 30 kcal mol^{-1} of the minimum low-resolution binding energy, we calculated high-resolution energies for all of them rather than using the screening protocol in Section 6.4.2 of Chapter 6. The final results of our design procedure are shown in Table A.1. The chosen wild-type sequence structure is almost identical to the crystal structure, including the conformations of the 3 mobile waters. For each other sequence, we kept the best-binding structure which folds no more than 1 kcal mol^{-1} worse than the chosen wild-type sequence conformation. None of them bind as tightly as the wild type, and all the ones with promising binding and folding free energies keep all 3 waters. Some conformations (with Glu, Gln, Arg, Lys, Met, or Ile at position 35) keep only 1 of the 3 waters, but do not score well.

The bottom line is that the wildtype sequence is predicted to be the best binder for these 2 barstar residues, even when the neighboring water molecules are given the freedom to rotate or vanish.

A.2.3 Results for Redesign of Barstar Trp38, Asp39, Thr42, and 8 Waters

For the barstar residues Trp38, Asp39, Thr42, and the 8 water molecules numbered 14, 29, 33, 48, 56, 60, 116, and 361, we followed our full design procedure and also calculated high-resolution energies for all 352 sequences within 30 kcal mol^{-1} of the minimum low-resolution binding energy, rather than using the protocol in Section 6.4.2 of Chapter 6. All 10 structures with the wildtype protein sequence have protein side chain conformations almost exactly like the crystal structure. Their waters are all present, and are in a variety of conformations. The final results of our design procedure are shown in Table A.2. The

Table A.1: Sequences of Barstar residues 33 and 35 found with good high-resolution binding and folding energy, when redesigned along with 3 neighboring waters. Also shown are the number of the 3 waters present, and the number of them in the crystal structure position. They are sorted by high-resolution binding energy relative to the wild type $\Delta\Delta G^{\text{bind}}_{\text{high}} = \Delta G^{\text{bind}}_{\text{high}} - \Delta G^{\text{bind}}_{\text{high}}(\text{wildtype})$. The high-resolution folding energy is also shown relative to the wild type. The wild-type sequence is shown in boldface. For clarity, every position which matches the wild type is shown as a blank space.

$\Delta\Delta G^{\text{bind}}_{\text{high}}$	$\Delta\Delta G^{\text{fold}}_{\text{high}}$	Sequence		waters	waters
(kcal mol^{-1})		33	35	present	xtal
Wild type:					
0.	**0.**	**N**	**D**	**3**	**3**
Good folders in order of binding:					
0.	**0.**	**N**	**D**	**3**	**3**
0.67	-3.71	F		3	2
0.91	-1.68	Y		3	2
0.92	-5.16	K		3	2
1.11	0.50	W		3	3
1.54	-6.25	R		3	2
1.91	-6.55	V		3	2
2.03	-4.30	E		3	2
2.05	0.09	H		3	3
2.07	-1.75	D		3	2
2.11	-0.36	L		3	2
2.22	0.72	h		3	3
2.52	-2.76	A		3	2
2.89	-3.57	T		3	2
2.99	-4.61	S		3	2
3.03	-5.56	C		3	2
3.04	-3.33	M		3	1
3.17	-4.81	Q		3	1
3.80	-1.52	G		3	2

chosen wildtype-sequence structure is almost tied for the best binding free energy among these 10 structures, and also almost tied for the best folding free energy. For each other sequence, we kept the best-binding structure which folds no more than one kcal mol^{-1} worse than the chosen wildtype-sequence conformation. None of them bind as tightly as the wildtype, and all the ones with promising binding and folding free energies keep all 8 waters. All conformations with sequences within 30 kcal mol^{-1} of the minimum low-resolution binding energy have no more than 2 absent waters. Only one sequence has 2 absent waters, and it does not score well or have a good steric fit.

Table A.2: Sequences of Barstar residues 38, 39, and 42 found with good high-resolution binding and folding energy, when redesigned along with 8 neighboring waters. Also shown are the number of the 8 waters present, and the number of them in the crystal structure position. They are sorted by high-resolution binding energy relative to the wildtype $\Delta\Delta G^{\text{bind}}_{\text{high}} = \Delta G^{\text{bind}}_{\text{high}} - \Delta G^{\text{bind}}_{\text{high}}(\text{wildtype})$. The high-resolution folding energy is also shown relative to the wildtype. The wildtype sequence is shown in boldface. For clarity, every position which matches the wildtype is shown as a blank space.

$\Delta\Delta G^{\text{bind}}_{\text{high}}$	$\Delta\Delta G^{\text{fold}}_{\text{high}}$	Sequence			waters	waters
(kcal mol^{-1})		38	39	42	present	xtal
Wildtype:						
0.	**0.**	**W**	**D**	**T**	**8**	**2**
Good folders in order of binding:						
0.	**0.**	**W**	**D**	**T**	**8**	**2**
2.75	0.48			I	8	2
3.67	0.87			F	8	1
4.53	-0.14			A	8	1
5.86	0.62	F		I	8	2

Again, the bottom line is that the wildtype sequence is predicted to be the best binder for these 3 barstar residues, even when the neighboring water molecules are given the freedom to rotate or vanish.

A.3 Conclusion

We performed a search for ways to improve the binding free energy of barnase and barstar by redesigning barstar residues and removing or rotating interfacial water molecules, and we found no redesigned sequences that bind better than the wildtype *as a result of* giving the waters this freedom. The search was thorough in its consideration of which water molecules and barstar residues to redesign. Our two most limiting assumptions were that we always used the crystal structure backbone conformation and binding geometry, and that we only used 12 possible conformations for each water molecule. Nevertheless, it is surprising that we found no evidence for the commonly-held theory that the interfacial water molecules are there as a result of a “trade-off” of binding affinity for binding speed.

Bibliography

[1] 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.

[2] Michael K. Gilson, Kim A. Sharp, and Barry H. Honig. Calculating the electrostatic potential of molecules in solution: Method and error assessment. *J. Comput. Chem.*, 9:327–335, 1988.

[3] M. K. Gilson and B. H. Honig. Calculation of electrostatic potentials in an enzyme active site. *Nature (London)*, 330:84–86, 1987.

[4] 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.

[5] K. A. Sharp and B. Honig. Electrostatic interactions in macromolecules: Theory and applications. *Annu. Rev. Biophys. Biophys. Chem.*, 19:301–332, 1990.

[6] M. Schaefer and M. Karplus. A comprehensive analytical treatment of continuum electrostatics. *J. Phys. Chem.*, 100:1578–1599, 1996.

[7] M. Schaefer, C. Bartels, and M. Karplus. Solution conformations and thermodynamics of structured peptides: molecular dynamics simulation with an implicit solvation model. *J. Mol. Biol.*, 284:835–848, 1998.

[8] E. Kangas. *Optimizing Molecular Electrostatic Interactions: Binding Affinity and Specificity*. PhD thesis, Massachusetts Institute of Technology, 2000.

[9] T. L. Hill. *An Introduction to Statistical Thermodynamics*. Dover, New York, 1986.

[10] 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.

[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] B. Jayaram, K. J. McConnell, S. B. Dixit, and D. L. Beveridge. Free energy analysis of protein-DNA binding: The ecori endonuclease-DNA complex. *J. Comput. Phys.*, 151:333–357, 1999.

[13] R. M. Levy and E. Gallicchio. Computer simulations with explicit solvent: Recent progress in the thermodynamic decomposition of free energies, and in modeling electrostatic effects. *Annu. Rev. Phys. Chem.*, 49:531–567, 1998.

[14] L. Lins and R. Brasseur. The hydrophobic effect in protein folding. *FASEB Journal*, 9:535–540, April 1995.

[15] K. A. Sharp. Calculation of hyhel10-lysozyme binding free energy changes: Effect of ten point mutations. *Proteins*, 33:39–48, 1998.

[16] W. C. Wimley, T. P. Creamer, and S. H. White. Solvation energies of amino acid residues and backbone in a family of host guest pentapeptides. *Biochemistry*, 35(16):5109–5124, 1996.

[17] H. S. Chan and K. A. Dill. Solvation, how to obtain microscopic energies from partitioning and solvation experiments. *Annu. Rev. Biophys. Biomol. Struct.*, 26:425–459, 1997.

[18] B. Lee and F. M. Richards. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.*, 55:379–400, 1971.

[19] J. D. Jackson. *Classical Electrodynamics.* John Wiley & Sons, 1999.

[20] J. O'M. Bockris and A. K. N. Reddy. *Modern Electrochemistry.* Plenum, New York, 1973.

[21] Michael K. Gilson and Barry H. Honig. Calculation of electrostatic potentials in an enzyme active site. *Nature*, 330:84–86, 1987.

[22] D. A. McQuarrie. *Statistical Mechanics.* Harper & Row, New York, 1976.

[23] K. A. Sharp and B. Honig. Calculating total electrostatic energies with the nonlinear Poisson–Boltzmann equation. *J. Phys. Chem.*, 94:7684–7692, 1990.

[24] E. Alexov and M. R. Gunner. Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties. *Biophys. J.*, 74:2075–2093, 1997.

[25] J. J. Havranek and P. B. Harbury. Tanford–Kirkwood electrostatics for protein modeling. *Proc. Natl. Acad. Sci. U.S.A.*, 96:11145–11150, 1999.

[26] M. Schaefer, M. Sommer, and M. Karplus. pH-dependence of protein stability: Absolute electrostatic free energy differences between conformations. *J. Phys. Chem. B*, 101:1663–1683, 1997.

[27] D. Bashford and M. Karplus. pK_a's of ionizable groups in proteins: Atomic detail from a continuum electrostatic model. *Biochemistry*, 29:10219–10225, 1990.

[28] M. McNutt, L. S. Mullins, F. M. Raushel, and C. N. Pace. Contribution of histidine residues to the conformational stability of ribonuclease T1 and mutant Glu58Ala. *Biochemistry*, 29:7572–7576, 1990.

[29] R. Loewenthal, J. Sancho, and A. R. Fersht. Histidine-aromatic interactions in barnase. *J. Mol. Biol.*, 224:759–770, 1992.

[30] T. E. Creighton. *Proteins: Structures and molecular properties, 2nd ed.* W.H. Freeman & Co., New York, 1993.

[31] M. Oliveberg, V. L. Arcus, and A. R. Fersht. pKa values of carboxyl groups in the native and denatured states of barnase: The pKa values of the denatured state are on average 0.4 units lower than those of model compounds. *Biochemistry*, 34:9424–9433, 1995.

[32] P. Beroza and D. A. Case. Including side chain flexibility in continuum electrostatic calculations of protein titration. *J. Phys. Chem.*, 100:20156–20163, 1996.

[33] A. M. Buckle, G. Schreiber, and A. R. Fersht. Protein–protein recognition: Crystal structural analysis of a barnase–barstar complex at 2.0-Å resolution. *Biochemistry*, 33:8878–8889, 1994.

[34] G. Schreiber, A. M. Buckle, and A. R. Fersht. Stability and function: Two constraints in the evolution of barstar and other proteins. *Structure*, 2:945–951, 1994.

[35] P. J. Kraulis. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.*, 24:946–950, 1991.

[36] W. Humphrey, A. Dalke, and K. Schulten. Vmd - visual molecular dynamics. *J. Mol. Graphics*, 14(1):33–38, 1996.

[37] M. Tanokura. H—NMR study on the tautomerism of the imidazole ring of histidine residues. *Biochimica et Biophysica Acta*, 742:576–585, 1983.

[38] K. B. Wiberg and K. E. Laidig. Rotational barriers adjacent to carbonyl groups 3. amide resonance and the C-O barrier in acids and esters. *J. Am. Chem. Soc.*, 109:5935–5943, 1987.

[39] C. Tanford and R. Roxby. Interpretation of protein titration curves: Application to lysozyme. *Biochemistry*, 11:2192–2198, 1972.

[40] A.-S. Yang, M. R. Gunner, R. Sampogna, K. Sharp, and B. Honig. On the calculation of pK_a's in proteins. *Proteins: Struct., Funct., Genet.*, 15:252–265, 1993.

[41] D. Bashford, D. A. Case, C. Dalvit, L. Tennant, and P. E. Wright. Electrostatic calculations of side-chain pK(a) values in myoglobin and comparison with NMR data for histidines. *Biochemistry*, 32(31):8045–8056, 1993.

[42] H. W. T. van Vlijmen, S. Curry, and M. Schaefer. Titration calculations of foot-and-mouth disease virus capsids and their stabilities as a function of pH. *J. Mol. Biol.*, 275(2):295–308, 1998.

[43] A. T. Brünger and M. Karplus. Polar hydrogen positions in proteins: Empirical energy placement and neutron diffraction comparison. *Proteins: Struct., Funct., Genet.*, 4:148–156, 1988.

[44] R. M. H. Gordon-Beresford, D. Van Belle, J. Giraldo, and S. J. Wodak. Effect of nucleotide substrate binding on the pKa of catalytic residues in barnase. *Proteins*, 25:180–194, 1996.

[45] D. Šali, M. Bycroft, and A. R. Fersht. Stabilization of protein structure by interaction of α-helix dipole with a charged side chain. *Nature (London)*, 335:740–743, 1988.

[46] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, 112:6127–6129, 1990.

[47] J. Srinivasan, M. W. Trevathan, P. Beroza, and D. A. Case. Application of a pairwise generalized Born model to proteins and nucleic acids: Inclusion of salt effects. *Theor. Chem. Acc.*, 101:426–434, 1999.

[48] M. Born. Volumen und Hydrationwärme der Ionen. *Z. Phys.*, 1:45–48, 1920.

[49] L. Onsager. Electric moments of molecules in liquids. *J. Am. Chem. Soc.*, 58:1486–1493, 1936.

[50] L.-P. Lee and B. Tidor. Optimization of electrostatic binding free energy. *J. Chem. Phys.*, 106:8681–8690, 1997.

[51] G. D. Hawkins, C. J. Cramer, and D. G. Truhlar. Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.*, 246:122–129, 1995.

[52] G. D. Hawkins, C. J. Cramer, and D. G. Truhlar. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.*, 100:19824–19839, 1996.

[53] D. Qiu, P. S. Shenkin, F. P. Hollinger, and W. C. Still. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A*, 101:3005–3014, 1997.

[54] A. Ghosh, C. S. Rapp, and R. A. Friesner. A generalized born model based on a surface integral formulation. *J. Phys. Chem. B*, 102:10983–10990, 1998.

[55] M. Schaefer, C. Bartels, F. Leclerc, and M. Karplus. Effective atom volumes for implicit solvent models: Comparison between Voronoi volumes and minimum fluctuation volumes. *J. Comput. Chem.*, 22(15):1857–1879, 2001.

[56] A. Mondragon, S. Subbiah, S. C. Almo, M. Drottar, and S. C. Harrison. Structure of the amino-terminal domain of phage 434 repressor at 2.0 A resolution. *J. Mol. Biol.*, 205:189, 1989.

[57] I. Morize, E. Surcouf, M. C. Vaney, Y. Epelboin, M. Buehner, F. Fridlansky, E. Milgrom, and J. P. Mornon. Refinement of the C222(1) crystal form of oxidized uteroglobin at 1.34 A resolution. *J. Mol. Biol.*, 194:725, 1987.

[58] B. E. Raumann, M. A. Rould, C. O. Pabo, and R. T. Sauer. DNA recognition by β-sheets in the Arc repressor–operator crystal structure. *Nature (London)*, 367:754–757, 1994.

[59] L. J. Stern, J. H. Brown, T. S. Jardetzky, J. C. Gorga, R. G. Urban, J. L. Strominger, and D. C. Wiley. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature*, 368:215, 1994.

[60] S. E. Phillips and B. P. Schoenborn. Neutron diffraction reveals oxygen-histidine hydrogen bond in oxymyoglobin. *Nature*, 292:81, 1981.

[61] C. A. Bewley, K. R. Gustafson, M. R. Boyd, D. G. Covell, A. Bax, G. M. Clore, and A. M. Gronenborn. Solution structure of cyanovirin-N, a potent HIV-inactivating protein. *Nature Struct. Biol.*, 5(7):571–578, 1998.

[62] F. Yang, C. A. Bewley, J. M. Louis, K. R. Gustafson, M. R. Boyd, A. M. Gronenborn, G. M. Clore, and A. Wlodamer. Crystal structure of cyanovirin-N, a potent HIV-inactivating protein, shows unexpected domain swapping. *J. Mol. Biol.*, 288:403–412, 1999.

[63] C. A. Bewley and G. M. Clore. Determination of the relative orientation of the two halves of the domain-swapped dimer of cyanovirin-N in solution using dipolar couplings and rigid body minimization. *J. Am. Chem. Soc.*, 122:6009–6016, 2000.

[64] 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.

[65] M. Scarsi and A. Caflisch. Comment on the validation of continuum electrostatics models. *J. Comput. Chem.*, 20:1533–1536, 1999.

[66] Y. M. Chook, H. Ke, and W. N. Lipscomb. Crystal structures of the monofunctional chorismate mutase from *Bacillus subtilis* and its complex with a transition state analog. *Proc. Natl. Acad. Sci. U.S.A.*, 90:8600–8603, 1993.

[67] M. Elrod-Erickson, M. A. Rould, L. Nekludova, and C. O. Pabo. Zif268 protein–DNA complex refined at 1.6 Å: a model system for understanding zinc finger–DNA interactions. *Structure*, 4:1171–1180, 1996.

[68] M. Elrod-Erickson, T. E. Benson, and C. O. Pabo. High-resolution structures of variant Zif268–DNA complexes: Implications for understanding zinc finger–DNA recognition. *Structure*, 6:451–464, 1998.

[69] J. A. Caravella. *Electrostatics and Packing in Biomolecules: Accounting for Conformational Change in Protein Folding and Binding.* PhD thesis, Massachusetts Institute of Technology, 2002.

[70] J. A. Nelder and R. Mead. A simplex method for function minimization. *Comput. J.*, 7:308–313, 1965.

[71] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes, 2nd Edition.* Cambridge University Press, Cambridge, 1992.

[72] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[73] F. M. Richards. The interpretation of protein structures: total volume, group volume distributions, and packing density. *J. Mol. Biol.*, 82:1–14, 1974.

[74] 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.

[75] M. Shimaoka, J. M. Shifman, H. Jing, L. Takagi, S. L. Mayo, and T. A. Springer. Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nature Struct. Biol.*, 7:674–678, 2000.

[76] 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.

[77] R. F. Goldstein. Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys. J.*, 66:1335–1340, 1994.

[78] N. A. Pierce, J. A. Spriet, J. Desmet, and S. L. Mayo. Confomational splitting: A more powerful criterion for dead-end elimination. *J. Comp. Chem.*, 21:999–1009, 2000.

[79] I. Lasters and J. Desmet. The fuzzy-end elimination theorem: Correctly implementing the side chain placement algorithm based on the dead-end elimination theorem. *Protein Eng.*, 6:717–722, 1993.

[80] P. H. Winston. *Artificial Intelligence.* Addison-Wesley, Reading, Massachussetts, 1992.

[81] 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., Func., Genet.*, 33:227–239, 1998.

[82] D. B. Gordon and S. L. Mayo. Branch-and-terminate: A combinatorial optimization algorithm for protein design. *Structure*, 7:1089–1098, 1999.

[83] R. L. Dunbrack, Jr. and M. Karplus. Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *J. Mol. Biol.*, 230:543–574, 1993.

[84] J. Mendes, A. M. Baptista, M. Arménia Carrondo, and C. M. Soares. Improved modeling of side-chains in proteins with rotamer-based methods: A flexible rotamer model. *Proteins: Struct., Funct., Genet.*, 37:530–543, 1999.

[85] B. I. Dahiyat and S. L. Mayo. De novo protein design: Fully automated sequence selection. *Science (Washington, D.C.)*, 278:82–87, 1997.

[86] B. Kuhlman, J. W. O'Neill, D. E. Kim, K. Y. J. Zhang, and D. Baker. Conversion of monomeric protein L to an obligate dimer by computational protein design. *Proc. Natl. Acad. Sci. U.S.A.*, 98(19):10687–10691, 2001.

[87] G. Schreiber and A. R. Fersht. Interaction of barnase with its polypeptide inhibitor barstar studied by protein engineering. *Biochemistry*, 32:5145–5150, 1993.

[88] R. W. Hartley. Directed mutagenesis and barnase–barstar recognition. *Biochemistry*, 32:5978–5984, 1993.

[89] L.-P. Lee and B. Tidor. Barstar is electrostatically optimized for tight binding to barnase. *Nature Struct. Biol.*, 8:73–76, 2001.

[90] A. T. Brünger and M. Karplus. Polar hydrogen positions in proteins: Empirical energy placement and neutron diffraction comparison. *Proteins*, 4:148–156, 1988.

[91] W. R. Tulip, V. R. Harley, R. G. Webster, and J. Novotny. N9 neuraminidase complexes with antibodies NC41 and NC10: empirical free energy calculations capture specificity trends observed with mutant binding data. *Biochemistry*, 33:7986–7997, 1994.

[92] B. I. Dahiyat and S. L. Mayo. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. U.S.A.*, 94:10172–10177, 1997.

[93] 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.

[94] A. G. Street and S. L. Mayo. Pairwise calculation of protein solvent-accessible surface areas. *Folding and Design*, 4:253–258, 1998.

[95] M. H. Abraham. Free energies, enthalpies, and entropies of solution of gaseous nonpolar nonelectrolytes in water and nonaqueous solvents. the hydrophobic effect. *J. Am. Chem. Soc.*, 104:2085–2094, 1982.