METHODS IN COMPUTATIONAL PROTEIN DESIGN

Thesis by

Victor Wai Tak Kam

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2008

(Defended May 30th, 2008)

Acknowledgements

I dedicate this thesis to my parents. Without all their support and encouragement, I would not even be here writing this.

Getting a Ph.D. is hard. But it was enjoyable when you are given the freedom to explore problems from directions of your choice. I thank my advisor, Professor Bill Goddard, for the support and faith he entrusted in me. I have much to learn from him—the creativity he displays towards solving problems, the breadth and depth of his knowledge, and his boundless energy and enthusiasm. I would also like to thank Dr. Vaidehi Nagarajan, who put in an enormous amount of time and effort and has always been a great friend.

I had a great group of friends at Caltech to help me get through this journey. I thank Julius Su, Candy Tong, Sam Cheung, Jiyoung Heo, Youyong Li and David Wei. I also thank John Keith, an exceptional cellist, for all the time we spent playing chamber music.

Finally, I thank my loving wife, Wing Li Leung. I do not know how I would get through these years without her steadfast love and support. I have not spent as much time with her as she truly deserved over these years, and I hope that with the completion of my doctorate degree I can begin to start making it up to her.

Abstract

In silico design of protein has generated enormous interest with the rapid advances in computational power. Biological systems are known for their complexity, and we have made a series of computational developments that allow us to perform computational protein design. In this work we present a methodology for the design and prediction of protein active sites.

We begin by presenting SCREAM, a program developed to accurately position sidechains in proteins. We show how using an improved scoring function and placement algorithm allow us to achieve better accuracy in the placement and prediction of sidechains in proteins compared to other methods.

We then describe the development of an accurate treatment for describing hydrogen bonding. This is done by refining the hydrogen bond term in the force field DREIDING. We also need to properly describe electrostatics effects in proteins, and to this end, we introduce neutralized residues for proteins. We found that this improves the variance in our predictions dramatically.

Finally, having established the components described above, we describe a protein design methodology encompassing the above methods and tools. We show predictions we made and those having been verified by experiments.

Table of Contents

LIST OF ILLUSTRATIONS

LIST OF TABLES

1 Computational Protein Design: Overview

Protein design is becoming a practical option for solving problems in protein engineering. The growth in this field has been greatly facilitated by the rapid increase in computational prowess and algorithmic advances made in the past two decades, and investigations today address a wide variety of problems. Progress is not only limited to academic researchers, but also to the industry, as computational methods are gradually being accepted as part of the drug discovery process. The permeation of computational methodologies is certain to continue as pharmaceutical companies face growing pressure to reduce development costs of new drugs, estimated at over $800 million[1] as of 2006.

Computational protein design is the process of introducing mutations in the original sequence of a target protein in order to introduce desired properties through computational means. Examples include improving stability between protein-protein interactions[2] and mutation of an active site of a protein to incorporate a non-cognate ligand[3]. There are two major difficulties with regards to introducing mutations. The first is how to predict computationally the positioning of other sidechains or atoms after the mutations, a problem known as the conformation search problem. The second is how to accurately assess the favorability of the mutations introduced, known as the energy function problem.

1.1 Protein Sidechain Conformation Search

The protein design problem can be thought of as an inverse version of the protein folding problem. In protein folding, the only input is the amino acid sequence, the output being the structure of the protein in all its three-dimensional glory. As of today, there are no known general methods that can reliably produce such an output. In protein design, typically we wish to introduce a special functionality for the protein, and we achieve this by altering the sequence of the protein. The location of the protein backbone often is included as part of

the input, either from the crystal structure of the protein, from homology modeling, or from other *ab initio* methods.

The assumption of a known protein backbone simplifies the search problem considerably, but the search space is still huge. There are 20 possible amino acids for each residue, and each amino acid sidechain have multiple conformations. The sheer number of combinations is huge. Consider the following. If each amino acid sidechain can take on 5 possible conformations, each amino acid can have 5 possible mutations, and there are 100 possible sites on the protein, then there are a total of $5^{100} \cdot 5^{100} \approx 10^{280}$ possibilities! This number is greater than the number of atoms in the universe and certainly not tractable by any computational means. Fortunately, many of these possibilities can be eliminated quickly and it is possible to arrive at reliable conformations by using various methods, procedures and algorithms[4-9] that have been developed in the past two decades.

1.1.1 Rotamers

Instead of working in the continuous conformational search space when placing amino acid sidechains on the protein backbone, Ponders and Richards[10] introduced using discrete sidechains configurations that occur frequently in protein structures. These low-energy conformations are called "rotamers" (see Figure 1-1 for two rotamers of the glutamine sidechain), short for rotational isomers.

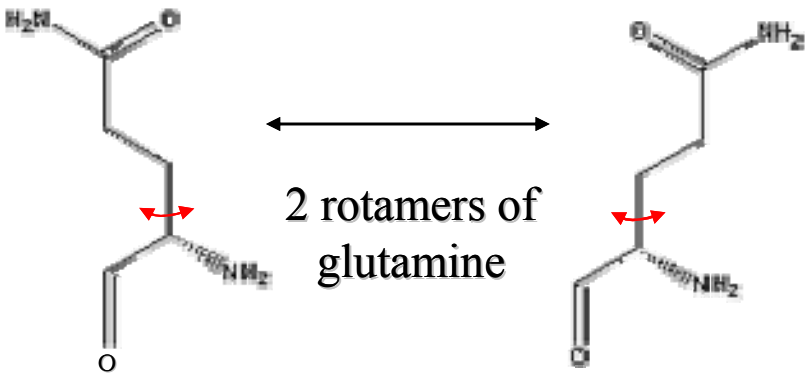


Figure 1-1 Illustration of 2 rotamers of the amino acid glutamine.

Each amino acid sidechain needs to be represented by a number of these rotamers. For amino acids such as arginine and lysine that have more torsional freedoms, more rotamers are needed to cover the conformation space. Conversely, simpler amino acids such as serine only need a small number of rotamers. A collection of rotamers that represent the amino acids is called a rotamer library[5,11]. The structures of the rotamers can be taken from crystal structures in proteins or be constructed according to chemical principles, such as energy minima of torsional angles. Clearly, the more rotamers we use, the more complete the rotational search space we are sampling and more accurate the results, but at an increased performance cost. Therefore, there is a trade-off between the accuracy (more rotamers) and speed (fewer rotamers) in our search problem. In Chapter 2, we explore methods to reduce the cost of the trade-off.

1.1.2 Placement of Sidechains

After we place the rotamers onto the protein backbone, we need to decide which configurations are better than others. This is done by using a scoring function[8], which can be energy-based, statistics-based, or some combination of the two. After deciding on a scoring function, there is also the global optimization problem of finding the combination of rotamers that yield the best score given the scoring function. Because of the combinatorial nature of the problem, a brute-force search will not be successful in finding a good solution to this problem. Many approaches have been developed[12].

In Chapter 2, we go into more details on energy-based scoring functions and a mean-field approach to the placement of sidechains.

1.2 Energy Expressions

Ab initio quantum chemistry methods are very accurate in calculating molecular structures and properties. Based on fundamental axioms in quantum mechanics, it is also the most expensive in terms of time needed to perform calculations. For many applications,

including simulation of large biomolecules, the level of information obtained from quantum mechanics is unnecessary.

Molecular mechanical force fields[13,14] have been developed to perform accurate calculations in a more reasonable amount of time. These force fields use simpler energy expressions to describe interactions between molecules, and are parameterized against experimental or quantum mechanical calculations. Popular force fields in the study of biological systems include CHARMM[13], AMBER[15] and DREIDING[14]. The accuracy of these force fields allows the study of such complex problems in molecular biology such as protein folding[16,17], molecular docking[18], and *ab initio* structure prediction[19,20]. Indeed, many commercial software, such as Accelrys' Quanta and Cerius2, include these biological force fields in the distribution.

We briefly review components of a biological force field. The total overall energy of a system is divided into valence terms and non-valence terms:

$$E = E_{val} + E_{non}$$

In the following section, we describe the functional forms that are used in DREIDING. Similar functions are used for other force fields.

1.2.1 Covalent Terms

DREIDING divides the valence terms into four components:

$$E_{val} = E_b + E_a + E_t + E_i$$

where E_b is the bond stretch energy between two atoms, E_a the bond angle energy between three atoms, E_t the torsional angle energy between four atoms, and E_i the inversion term between four atoms. For E_b, E_a and E_t, a harmonic expression is often employed to capture the fact that there structures tend to reside in a local neighborhood around equilibrium values. Periodicity is included for the angle term and the torsion term. The equilibrium

values, for these terms are parameterized. $E_{i,}$ the inversion term, is a special term in DREIDING that ensures the molecules stay planar (as in benzene) or maintain the correct chirality during simulations. A harmonic expression can also be used here.

Valence bonds are never broken in DREIDING and most other biological force fields.

1.2.2 Non-covalent Terms

Many interesting phenomena in bio-molecules involve non-covalent interaction between molecules, such as a ligand binding to a receptor. Thus, the accurate description between non-bonded atoms is crucial to the accurate prediction of inter-molecular properties such as binding energies. In DREIDING, non-valence terms include:

$$E_{non} = E_{VDW} + E_Q + E_{HB}$$

where E_{VDW} is the two-body van der Waals (VDW) energy (also known as dispersion), E_Q the two-body Coulombic or electrostatic interaction, and E_{HB} a special three-body hydrogen bond term.

Historically, the Lennard-Jones 12-6 potential is often used as the expression for the VDW term:

$$E_{VDW}(R) = D_{VDW}[(R_{VDW} / R)^{12} - 2(R_{VDW} / R)^6]$$

where D_{VDW} is the equilibrium VDW well-depth of a pair of atoms, R_{VDW} the equilibrium radial distance, and R the current distance between the two atoms. While this potential is considered to be too repulsive at short distance, it became popular back in the days when computers were not as powerful because it is simpler to compute than other alternative energy expressions. The Lennard-Jones 12-6 potential is still the dominant form of VDW today although other VDW expressions, such as Morse, Exponential-6 and softer Lennard-Jones potentials such as 9-6 are also employed. All VDW expressions satisfy a $1/r^6$

asymptotic behavior at large distances, reflecting the fact that its origin lies in the instantaneous but fluctuating dipole moments of neutral atoms.

Typically, the equilibrium radial distance between two atoms is the sum of the atomic radii of the two interacting atoms. There are instances when we do not want this, and we will show an example in Chapter 3.

The classic Coulomb interaction form is used for the electrostatic term:

$$E_Q = (322.0637)Q_i Q_j / \varepsilon R_{ij}$$

where the energy is expressed in kcal/mol and R_{ij} is expressed in Å. Q_i, Q_j are point charges assigned to center of atoms. The dielectric constant is usually taken to be ε=1.0, but values greater than 1.0 have been used to account for the effect of polarization. We go into more detail on this issue in Chapter 3.

DREIDING uses a 3-body hydrogen bond term that is not common in many other force fields:

$$E_{hb} = D_{hb}[5(R_{hb} / R_{DA})^{12} - 6(R_{hb} / R_{DA})^{10}]\cos^4(\theta_{DHA})$$

where D_{hb} stands for the well-depth of the hydrogen bond potential, R_{hb} the equilibrium distance and θ_{DHA} the angle between the hydrogen bond donor atom, hydrogen and the acceptor atom. This expression was inherited from the early days of molecular mechanical force fields 30 years ago when there was no reliable method of assigning atomic charges. The situation is different today, with charges from quantum mechanics readily available. Despite readily available quantum mechanics charges, it is difficult to obtain accurate interaction energies between a pair of hydrogen-bonding molecules because of constraints set forth by VDW and charges. Considering that polar interactions are such an important

component of bio-molecular recognition, we explore different hydrogen bond functions in Chapter 3 in an effort to improve the accuracy.

1.3 Summary

Protein design has two main components: a search component and a scoring function component. The search component arises from the combinatorial nature of mutation selection and sidechain placement. A force field is often used as the scoring function. In subsequent chapters, we will provide more in-depth explorations of these issues. We will also present a few examples of protein design using all these new developments in the final chapter.

2 SCREAM

2.1 Introduction

In developing general predictive approaches for structures of membrane proteins[21-23], we found that current available sidechain placement methods, e.g. SCWRL, did not provide sufficiently accurate results to determine the helix-helix relative orientations within the membrane. Consequently, we developed the SCREAM (*S*ide*C*hain *R*otamer *E*nergy *A*nalysis *M*ethodology) approach reported here, which we have found to lead to dramatically improved protein structures. Here, we present validation of SCREAM against standard libraries of crystal structures.

As briefly mentioned in Chapter 1, sidechain placement methods play a major role in recent applications in the field of computational molecular biology; from protein design[24-26], flexible ligand docking[27], loop-building[28], to prediction of protein structures[29]. Much attention has been paid to this important problem, which is difficult because it is in a category of problems known as NP-hard[30], for which no efficient algorithm is known to exist. Since the groundbreaking work by Ponder and Richards[10], many approaches have been developed, including mean-field approximation[6,31], Monte Carlo algorithms[7,32], and Dead-End Elimination (DEE)[4,33-35]. In practice, however, studies have also concluded that the combinatorial issue may not be as severe as originally thought[36,37]. Compared to the placement methods and rotamer libraries, scoring functions have not been studied as extensively[8,38,39]. Here, we focus on the scoring function.

Our scoring function is based on the all-atom forcefield DREIDING[14] which includes an explicit hydrogen bond term. The use of a rotamer library is widely used in sidechain prediction methods, and many authors have introduced quality rotamer libraries[11,37,40] since the Ponder library. To account for the discreteness of rotamer libraries, several approaches have been introduced, such as reducing van der Waals radii[41,42], capping of repulsion

energy[43], rotamer minimization[7,44] and the use of subrotamer ensembles for each dominant rotamer[45]. We introduce a flat-bottom region for the van der Waals (VDW) 12-6 potential and the DREIDING hydrogen bond term (12-10 with a cosine angle term). The width of the flat-bottom depends on the specific atom of each sidechain, as well as the coarseness of the underlying rotamer library used.

We show in this study that accuracy can be improved substantially by introducing the flat-bottom potential, and in a systematic way. In addition to showing that placement accuracy is dependent upon the number of rotamers used in a library, we find that it is possible for suitably chosen energy functions to compensate the use of coarser rotamer libraries. We demonstrate a high overall accuracy in sidechain placement, and make comparison to the popular sidechain placement program SCWRL[46].

2.2 Materials and Methods

2.2.1 Preparation of Rotamer Libraries

Rotamer libraries of various diversities are derived from the complete coordinate rotamer library of Xiang[37]. We added hydrogens to the rotamers, and considered both δ and ε versions in the case for histidines. CHARMM charges are used throughout[13]. Since the Xiang library was based on crystal structure data, we minimized each of the conformations so that the internal energies will be consistent with subsequent energy evaluations of the proteins. To do this we placed each sidechain on a template backbone (Ala-X-Ala in the extended conformation) and did 10 steps conjugate gradient minimization using the DREIDING forcefield.

We generated rotamer libraries of varying coarseness by a clustering procedure, using the heavy atom RMSD between minimized rotamers as the metric. Starting with the closest rotamers, we eliminated those within the specific threshold RMSD value choosing always the rotamer with the lowest minimized DREIDING energy. This threshold RMSD value is defined as the *diversity* of the resulting library. To ensure that rotamers can make proper hydrogen bonds, each sidechain conformation for serine, threonine, and tyrosine was repeated with each possible polar hydrogen position. Thus, for serine and threonine, the three sp^3 position hydrogens were added to the hydroxyl oxygen, while for tyrosine, we add the out-of-place OH bonds 90 degrees from the phenyl ring in addition to two sp^2 positions in the plane. The final number of rotamers for libraries of different diversities is shown in Table 2-1.

Diversity	Starting	0.2Å	0.6Å	1.0Å	1.4Å	1.8Å	2.2Å	3.0Å	5.0Å	All-Torsion
Rotamer Count	35828	14755	3195	1014	378	214	136	84	44	382

Table 2-1 Number of rotamers in libraries of various diversities.

In addition, we constructed the "*All-Torsion*" rotamer library in which one rotamer for each major torsional angle (120 degrees for sp^3 anchor atoms, 180 degrees for sp^2 anchor atoms) was included. The angles were obtained from the backbone independent rotamer library from Dunbrack[5] and built using the same procedure as described above.
All our rotamer libraries are backbone independent.

2.2.2 Preparation of Structures for Validation of SCREAM

We considered three sets of protein for validating and training SCREAM.

Xiang: Xiang[37] considered 33 proteins for testing their method for developing libraries of side chain conformations : 1aac, 1aho, 1b9o, 1c5e, 1c9o, 1cbn, 1cc7, 1cex, 1cku, 1ctj, 1cz9, 1czp, 1d4t, 1eca, 1igd, 1ixh, 1mfm, 1plc, 1qj4, 1ql0, 1qlw, 1qnj, 1qq4, 1qtn, 1qtw, 1qu9,

1rcf, 1vfy, 2pth, 3lzt, 5p21, 5pti and 7rsa. We have tested SCREAM for exactly these cases.

Liang: Liang[38,47] considered 15 proteins for testing their method for scoring functions for choosing side chain conformations. Of these, the 10 that were not in the Xiang set are denoted as the Liang set: 1bpi, 1isu, 1ptx, 1xnb, 256b, 2erl, 2hbg, 2ihl, 5rxn and 9rnt. The proteins that overlap with the Xiang set are not included.

Other: In addition we included 10 proteins with resolution not worse than 1.8Å from the SCWRL dataset: 1a8d, 1bfd, 1bgf, 1c3d, 1ctf, 1ctj, 1moq, 1rzl, 1svy and 1yge. Here we ignored structures with ligands or missing residues or which had a sequence identity of more than 50% with the Xiang or Liang sets. As will be described in later sections, this set is used only for deriving the σ-values and sidechain placement parameters.

For each of these 53 proteins, the raw atom coordinates were downloaded from the PDB database. Hydrogens were added using WHATIF[48] and ligands were typed using PRODRUG[49]. Manual typing of ligands were carried out in cases where they cannot be typed by PRODRUG (~10 cases). Waters, solvents, and metals were kept when present.

These structures were then minimized (100 conjugate gradient steps) using the DREIDING forcefield. In all cases, the minimized structures differed by less than 0.3Å total RMSD compared to the original crystal structures. All metals, prolines, cysteines in disulfide bonds and sidechains in coordination with metals were kept fixed throughout sidechain placement calculations.

2.2.3 Surface Area Calculations

Which residues were considered as buried or exposed was determined from the Solvent Accessible Surface Area (SASA), using a probe of radius 1.4Å. The reference for fully exposed surface area for each sidechain type is a fully extended tri-peptide in the form of Ala-X-Ala. A sidechain with >20% SASA compared with the reference SASA was considered exposed. This percentage is smaller than the typical 50% level in the literature—around 25% for the Xiang set and 39% for the Liang set because we include solvent molecules as part of the structure.

2.2.4 Positioning of Sidechains

Placement of the rotamers on the backbone is decided by the coordinates of the C, C_α, N backbone atoms plus the C_β atom. To specify the position of the C_β atom we use the coordinates with respect to C, C_α, and N based on the statistics gathered from the HBPLUS protein set (see above). This involves three parameters:

1. The angle of the Cα-C_β bond from the bisector of the C-Cα-N angle: 1.81° (from the HBPLUS protein set)

2. The angle of the Cα-C_β bond with the C-C_α-N plane: 51.1° (from the HBPLUS protein set)

3. The Cα-C_β bond length: 1.55Å (average value from the Other protein set).

Thus the C_β atom will generally have a different position from the crystal C_β position. As is common practice in the literature, we did not include this C_β deviation in the RMSD calculations.

2.2.5 Combinatorial Placement Algorithm

The SCREAM combinatorial placement algorithm consists of three stages: self energy calculation for rotamers, clash elimination, and further optimization of sidechains.

2.2.5.1 Stage 1: Rotamer Self Energy Calculation

The all atom forcefield DREIDING[14] was used to calculate the interactions between atoms, with a modification to be described in the scoring function section. The internal energy contributions $E_{internal}$ (bond, angle and torsion terms and non-bonds that involve only the sidechain atoms) were pre-calculated and stored in the rotamer library. For each residue to be replaced, the interaction energy ($E_{sc\text{-}fixed}$) was calculated for each rotamer interacting with just the protein backbone and fixed residues (all fixed atoms). The sum of these two terms is the *empty lattice energy* (E_{EL}) of a rotamer in the absence of all other sidechains to be replaced

$$E_{EL} = E_{internal} + E_{sc-fixed}$$

We use the term *ground state* to refer to the rotamer with the lowest E_{EL} energy. All other rotamer states are termed *excited states.* Excited states with an energy 50 kcal/mol above the ground state were discarded from the rotamer list for the remaining calculations.

2.2.5.2 Stage 2: Clash Elimination

Eisenmenger et al.[36] showed that the sidechain-backbone interaction accounts for the geometries of 74% of all core sidechains and 53% of all sidechains. Thus, the ground state of each sidechain was taken as the starting structure. Of course, this structure might have severe VDW clashes between sidechains since no interaction between sidechains has been included. Elimination of these clashes was done as follows. A list of clashes of all ground state pairs, above a default threshold of 25kcal/mol was sorted by their clashing energies. The pair (A, B) with the worst clash was then subjected to rotamer optimization by considering all pairs of rotamers, and selecting the lowest energy to form a super-rotamer with a new energy:

$$E_{tot}(A,B) = E_{self}(A) + E_{self}(B) + E_{Int.}(A,B) \equiv E_{self}(AB)$$

where E_{Int} indicates the interaction energy between rotamer A and rotamer B, which was the only energy calculation done at this step since the E_{EL} terms were calculated in Stage 1. The ground state for this super rotamer now replaced the rotamer pair in the original structure. Since large sidechains such as ARG and LYS may have as many as 700 rotamers for the 1.0A library, we limited the number of pairs to be calculated explicitly to 1,000, which we selected based upon the sum of the empty lattice energies. Of these interaction pairs we kept the ones with interaction energies below 100 kcal/mol.

After resolving a clash, we considered the lowest rotamer pairs from the above calculation as a super residue. Thus, subsequent clash resolution, say between residue C and residue A, will consider interactions of all sidechains of C with the (A,B) rotamer pairs. Now the spectrum of interaction energies treats (A,B) as a super rotamer so that the (C, (A,B)) energy spectrum is treated the same as for a simple rotamer pair with the spectrum:

$$E_{tot}(A,B,C) = E_{self}(A) + E_{self}(B) + E_{self}(C) + E_{Int}(A,B) + E_{Int}(A,C) + E_{Int}(B,C)$$
$$= E_{self}(AB) + E_{self}(C) + E_{Int}(A,C) + E_{Int}(B,C)$$

$$\equiv E_{self}(AB) + E_{self}(C) + E_{Int}(AB,C)$$

This process continued by generating a new list of clashing residue pairs including the new (A,B,C), resolving the next worst clash as above. The procedure was repeated until no further clashes were identified between two rotamers or super-rotamers.

2.2.5.3 Stage 3: Final Doublet Optimization

It is possible for some clashes to remain after Stage 2, since the number of rotamers pair evaluations is capped (at 1,000) and also the numbers of rotamers in a super-rotamer (20). To solve this problem, the structure from the end of stage 2 was further optimized. Sidechain pairs (termed *doublets*) were now ordered in decreasing energies in the presence of all other sidechains, and one iteration round of local optimization on those residue pairs was performed in that order. Any residue that had already been examined in this stage as

part of a doublet was eliminated from further doublet examination. Always, the doublet with the lowest overall energy was kept.

2.2.5.4 Stage 4: Final Singlet Optimizations

The structure would undergo one final round of optimization, where all residues were examined one at a time, again in order of decreasing energies for the rotamer currently placed in the structure. Again, the rotamer with the best overall energy was retained for the final structure. More iterations rounds on the final result improved the overall RMSD (unpublished results), but we did not pursue this path[50] for the purposes of this presentation.

We illustrate the effects of the doublet and singlet optimization stages by giving a specific example—1aac, using the 1.0Å rotamer library and optimal parameters (to be described in a later section). After the clash elimination stage, the RMSD between the predicted structure and the crystal structure was 0.733Å. The pair clashes remaining in this case included the pairs F57 and L67, V37 and F82, and V43 and W45. Doublets optimization brought the RMSD down to 0.703Å. The final singlet optimization stage brought the RMSD value further down to 0.622Å.

For this case, doublet optimization took 3 seconds, while singlet optimization took 13 seconds. For comparison, clash elimination took 30 seconds to complete, while the rotamer self energy calculation took 8 seconds.

2.2.6 The Flat-Bottom Scoring Function

Since our library is discrete, the best position for a sidechain may lead to some contacts slightly too short. Since the VDW interactions becomes very repulsive very quickly for distances shorter than R_e, a distance too short by even 0.1A may cause a very repulsive VDW energy. This might lead to selecting an incorrect rotamer. In order to avoid this problem, we use a flat-bottom potential in which the attractive region is exactly the same down to R_e but the repulsive region is displaced by some amount Δ so that contacts that are

slightly too short by Δ will not cause a false repulsive energy. The form of this potential is shown in Figure 2-1.

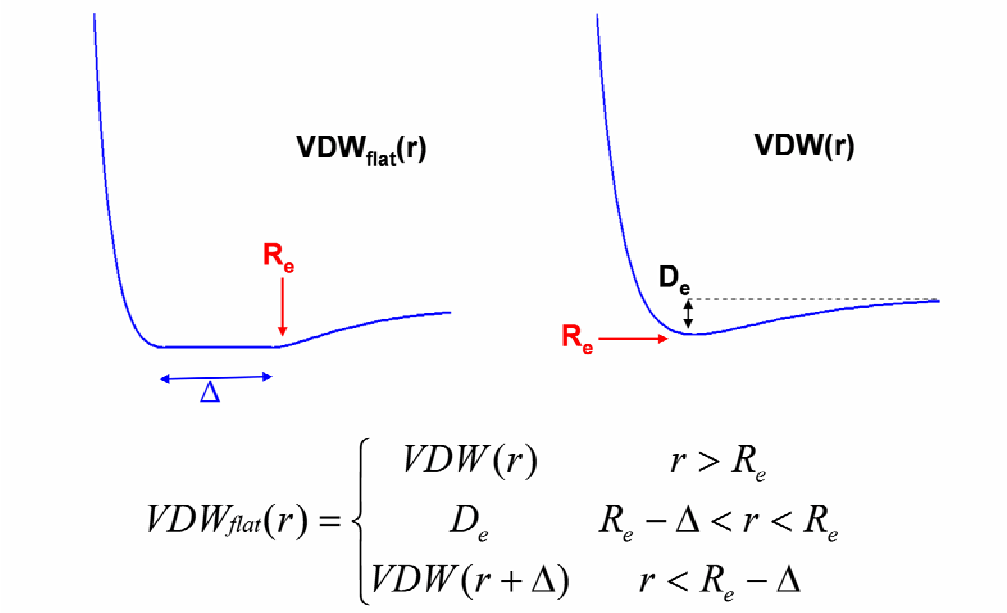


Figure 2-1 The flat-bottom potential. The inner wall is shifted by an amount Δ.

We allow a different Δ for each atom of each residue of each diversity. The way this is done is by writing Δ as:

$\Delta = s \cdot \sigma$

Where *s* is a scaling factor and the σ values are compiled as follows.

2.2.6.1 Compilation of σ values

For each rotamer library we considered the 10 query protein structures in the HBPLUS set (see Materials and Methods). For each sidechain in each query structure, we picked the closest matching rotamer (in RMSD) from the library and record the distance deviation for each atom of the sidechain of that residue. Thus, the atoms at the tip of the longer

sidechains such as arginine and lysine would have greater distance deviations than C_β atoms. The mean distance deviation (δ) for every atom of each amino-acid type over all 10 query proteins is then calculated. As an example, the δ values for arginine and lysine rotamers in the rotamer library of 1.0A diversity (rotamer libraries were described in 2.2.1) are listed in Table 2-2 and Table 2-3.

Dist. Deviation (Å)	Mean (δ)	Corrected Error (σ)
C_β	0.090	0.059
C_γ	0.245	0.153
C_δ	0.439	0.275
N_ε	0.502	0.315
C_ζ	0.588	0.369
$N_{\eta 1}$, $N_{\eta 2}$	0.858, 0.839	0.538, 0.526

Table 2-2 δ and σ values for each atom on the Arginine side-chain, listed in order of distance away from the mainchain. Nη1 and Nη2 are equivalent atoms; the average value is used in actual calculations. These numbers were obtained from the rotamer library of diversity 1.0Å.

Dist. Deviation (Å)	Mean (δ)	Corrected Error (σ)
C_β	0.089	0.056
C_γ	0.259	0.162
C_δ	0.406	0.254
C_ε	0.596	0.373
N_ζ	0.803	0.503

Table 2-3 δ and σ values for each atom on the Lysine side-chain, listed in order of distance away from the mainchain. These numbers were obtained from the rotamer library of diversity 1.0Å

We assume that the error in positioning of any one atom of the sidechain will have a Gaussian distribution of the form:

$$f(r) \propto e^{-\frac{r^2}{2\cdot\sigma^2}}$$

Where r is radial distance and σ represents the standard deviation. Thus,

$$\rho(r) \propto 4\pi r^2 f(r)$$

is the probability of finding an atom at position r from the crystal position (which is weighted by a factor of $4\pi r^2$ from the x, y and z distributions). The uncertainty δ in the Cartesian distance along the line between two atoms is related to σ by the form:

$$\delta = 2 \cdot \sqrt{\frac{2 \cdot \sigma^2}{\pi}}$$

where δ is the value described above. This σ is listed for arginine and lysine in Table 2-2 and Table 2-3.

2.2.6.2 Scaling factor s

The Δ values for each sidechain atom type will depend on their σ values:

$\Delta = s \cdot \sigma$

The deviations for σ above provide a measure of relative uncertainties in the ability of a library to describe the correct position of the sidechain atoms. However, to obtain the absolute value of the flat-bottomness we allow an overall scaling factor for the flat-bottom portion of the potential for all atoms.

The value of *s* was optimized for the Xiang set of 33 proteins for libraries of diversities ranging from 0.2A to 5.0A as discussed in section 3.

2.2.6.3 Flat-bottom potential on Hydrogen bond terms

We use a flat-bottom for the VDW interactions and not for the Coulomb interactions because the VDW inner wall potential becomes repulsive very quickly with distance (e.g. $1/r^{12}$). Such scaling is not important for Coulomb since it scales as 1/r. Most force fields use a modified VDW interaction between hydrogen bonded atoms. Current version of AMBER and CHARMM do this between donor hydrogen and the acceptor heavy atom, treating the interaction as a standard 12-6 Lennard-Jones with modified parameters. The flat-bottom for the other van der Waals interactions should apply equally well for these hydrogen bond terms. However, DREIDING uses an explicit 12-10 hydrogen bond term between the heavy atoms combined with a factor depending upon the linearity of the donor-hydrogen-acceptor triad:

$$E_{hb} = D_{hb}[5(R_{hb} / R_{DA})^{12} - 6(R_{hb} / R_{DA})^{10}]\cos^4(\theta_{DHA})$$

where D_{hb} stands for the well-depth of the hydrogen bond potential, R_{hb} the equilibrium distance and θ_{DHA} the angle between the hydrogen bond donor atom, hydrogen and the acceptor atom. We use a flat bottom potential for this DREIDING hydrogen bond term. However, we now allow both the inner and outer walls to shift by an amount Δ from the equilibrium point. The objective here is to also let the potential to capture the polar contacts that would otherwise be missed, both when a donor-acceptor pair is too close or too far away from each other.

2.2.6.4 Charges

We use the CHARMM[13] charges for the protein and water, since these are standard and well-tested values. For ligands and other solvents, we use QEq[51] charges, which provide values similar to those from quantum mechanics.

The Coulomb interaction between atoms 1 and 2 is written as:

$$E_{Coulomb} = \frac{c_0}{\varepsilon}\frac{q_1 q_2}{r_{12}}$$

where q_1 and q_2 are charges in electron units, r_{12} in Å, ε the dielectric constant and c_0=332.0637 converts to energies in kcal/mol. After optimization on a Xiang set of proteins using the 1.0Å diversity rotamer library and a scaling factor *s*=1.0, we chose the dielectric ε=6.0 (see Figure 2-2). Our calculation of electrostatics used a cubic spline cutoff beginning at 8Å and ending at 10Å.

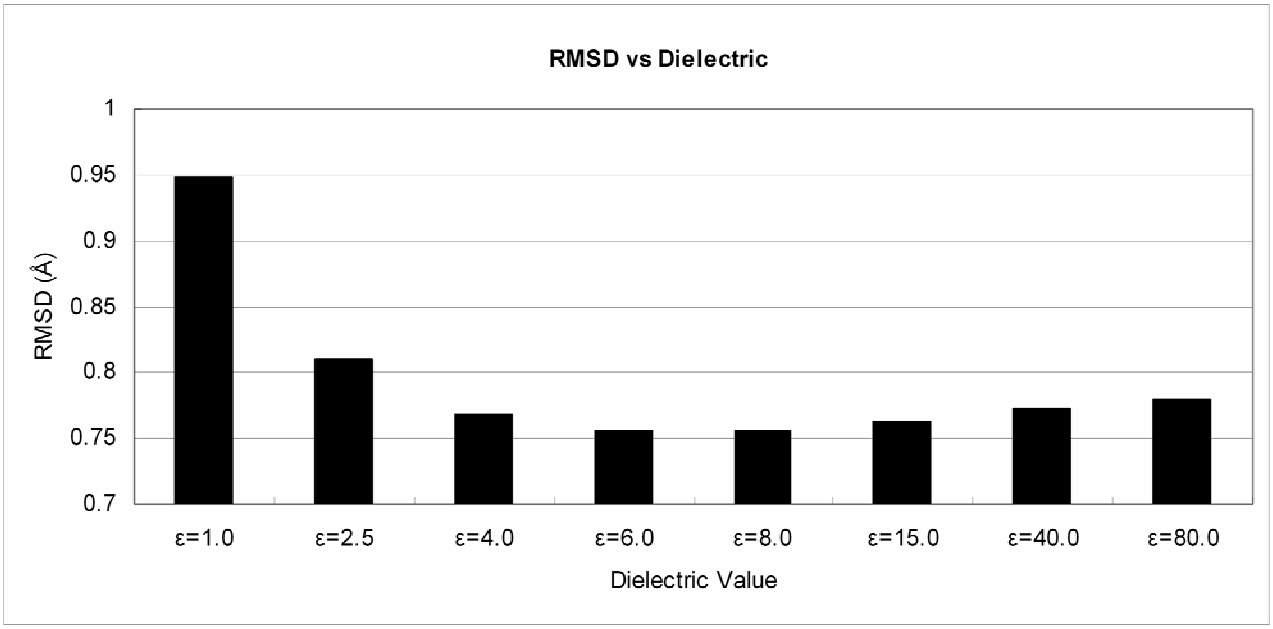


Figure 2-2 Effects on dielectric value on RMSD. The optimum value for the constant dielectric, ε=6.0 shown here, was obtained by fitting results for the Xiang set with a diversity of 1.0Å and a scaling factor s of 1.0.

2.2.6.5 Total rotamer energies

The valence energies (bonds, angles, torsions and inversion) plus the internal HB, Coulomb and VDW energies of the rotamers were calculated beforehand and stored in the rotamer library.

The final form of the scoring function is thus:

$$E_{Total} = \sum_{i} E_{EL} + \sum_{i<j} E_{Pair}$$

where E_{EL} is the sum over internal energies and the backbone interaction energies as described in 2.1 and

$$E_{Pair} = E_{VDW} + E_{HB} + E_{Coulomb}$$

is the total non-bond energy between all pairs of atoms between a pair of residues.
For any particular atoms i and j, the total flat-bottom correction $\Delta_{i,j}$ for the VDW and HB terms is obtained from the individual Δ values of Δ_i and Δ_j using the relation:

$$\Delta_{i,j} = \sqrt{\Delta_i^2 + \Delta_j^2}$$

This value corresponds to the standard deviation from the convolution of two normal distributions with standard deviations Δ_i and Δ_j.

2.3 Results and Discussion

2.3.1 Single Placement of Side-chains

To explore the effect on placement accuracy of using flat-bottom potentials, we increased the scaling factor *s* from 0.0 (no scaling) to 2.0 in 0.1 increments. To isolate the effects of the scaling, we placed sidechains one at a time onto the protein, in the presence of all other sidechains in their crystal positions. The values here represent the best possible results given a scoring function and a rotamer library[8]. The Xiang set of proteins described in Materials and Methods are used here.

Figure 2-3 shows that the best scaling factor is $s \sim 1$ for all rotamer libraries. Note that s=1 for the 1.0A library leads to an accuracy of 0.665A which is much better than the accuracy of 0.71 Å obtained using s=0 (no scaling) for the much bigger 0.6A library.

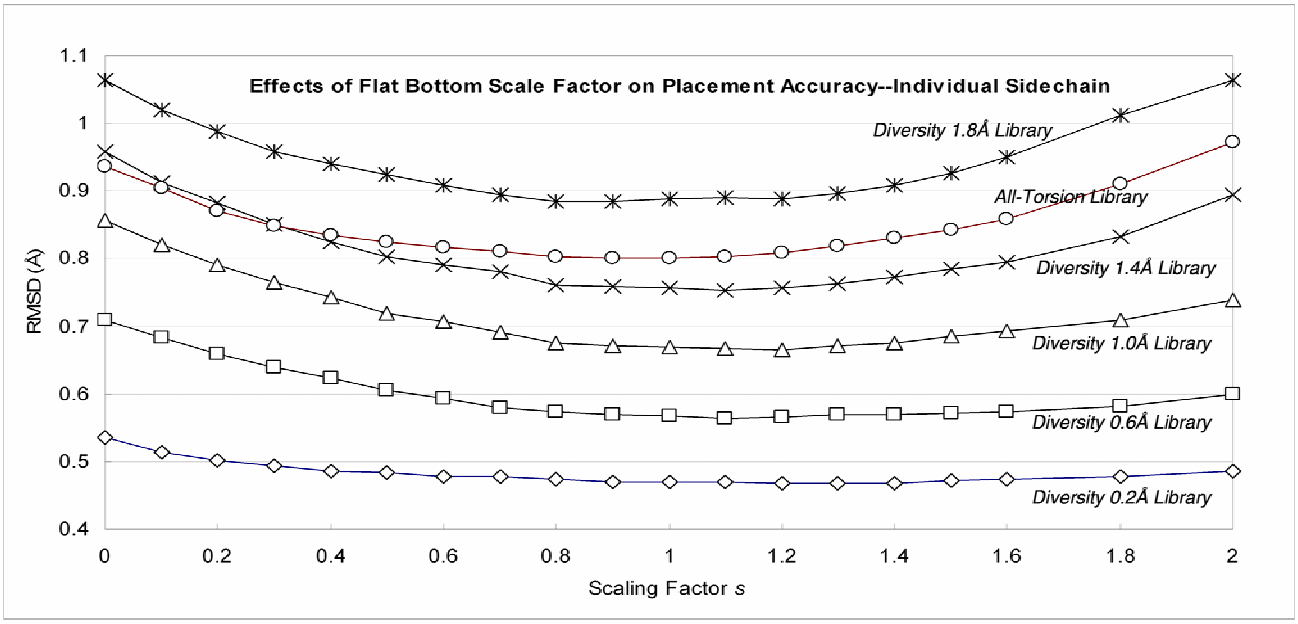


Figure 2-3 Single sidechain placement accuracy for various rotamer libraries at different s values. Shown are the libraries of 0.2Å diversity (14755 rotamers), 0.6Å diversity (3195 rotamers), 1.0Å diversity (1014 rotamers), 1.4Å diversity (378 rotamers), 1.8Å diversity (218) and all-torsion (382 rotamers). The coarser the rotamer library is, the more pronounced the effect of s becomes.

Taking the all-torsion rotamer library as an example, the RMSD improves from 0.94Å for $s = 0$ (no flat bottom) to 0.80Å for $s = 0.9$. This library with 378 rotamers leads to an accuracy of 0.80Å, which compares with the accuracy of 0.75Å obtained using the 1.4Å library, which has 382 rotamers.

We optimized the scaling factors for rotamer libraries of diversities ranging from just 5.0 Å (44 rotamers) to 0.2 Å (13,000 rotamers). Table 2-4 and Table 2-5 lists the optimum scaling factors and accuracies of these rotamer libraries, which lead to accuracies ranging from 0.47 Å (0.2 Å diversity) to 1.86 Å (5.0 Å diversity). We consider that the 1.0 Å library with an accuracy of 0.665 Å using 1014 total rotamers as a good compromise of efficiency and accuracy. These tables also list the results for the unscaled potential.

Library	Number of Rotamers	Unmodified Potential (RMSD, Å)	Best s value	Best RMSD(Å)
0.2Å	14755	0.536	1.3	0.468
0.6Å	3195	0.710	1.1	0.564
1.0Å	1014	0.857	1.2	0.665
1.4Å	378	0.958	1.1	0.753
1.8Å	214	1.064	0.9	0.885
2.2Å	136	1.343	0.8	1.175
3.0Å	84	1.624	0.7	1.487
5.0Å	44	1.890	0.7	1.860
AllTorsion	382	0.937	0.9	0.800

Table 2-4 Optimized s value for rotamer libraries of size ranging from 0.2Å to 5.0Å, plus the all torsion rotamer library. The s values that give the best RMSD value are listed

Library	Number of Rotamers	χ1/χ1+2 accuracy (from unmodified scoring function)	Best scaling factor *s*	χ1/χ1+2 accuracy using best *s value*
0.2Å	14755	95.0% / 91.8%	1.3	96.3% / 93.4%
0.6Å	3195	92.6% / 87.7%	1.1	95.6% / 92.1%
1.0Å	1014	90.0% / 83.4%	1.2	95.3% / 90.4%
1.4Å	378	87.8% / 80.0%	1.2	94.7% / 88.9%
1.8Å	214	84.3% / 75.6%	1.2	91.5% / 83.8%
2.2Å	136	71.9% / 61.0%	0.8	79.1% / 68.0%
3.0Å	84	63.4% / 54.1%	0.7	68.4% / 58.9%
5.0Å	44	53.2% / 44.9%	0.7	54.9% / 45.8%
AllTorsion	382	89.6% / 81.3%	1.1	93.3% / 86.8%

Table 2-5 Effect of s values on χ1/χ1+2 accuracy. Rotamer libraries of diversity ranging from 0.2Å to 5.0Å, plus the all torsion rotamer library are used. The best χ1+2 accuracy is used to determine the most effective scaling factor s. A χ angle is considered correct if within 40◦ of the corresponding χ angle in the crystal sidechain conformation.

2.3.2 Effects of Buried vs. Exposed Residues

The percentage of exposed residues considered in Section 2.3.1 is only 25% because crystallographic waters and solvents were included in the calculation. We consider this as the best test of the scoring function. However, in practical applications, such water and solvent molecules will not be present. This creates additional uncertainties for the surface residues whose positions should be affected by the solvent and water. Without such solvent molecules, the energy functions will tend to distort the sidechains to interact with other residues of the protein. Surface residues have more flexibility and it would be better to have smaller scaling factors for these sidechains. Thus, we optimized separate scaling factors for surface residues versus bulk. To do this, we calculated the SASA for the Xiang

set and assigned all residues > 20% exposed as surface. The resulting optimized scaling factors are in Table 2-6. In Figure 2-4, we see that the accuracy for the 1.4 Å library increases from 0.809 (bulk) and 1.409 (surface) to 0.515 Å (bulk) and 1.107 Å (surface).

The current SCREAM software does not distinguish between surface and bulk residues. In order to predict the surface residues prior to assigning the sidechains, we recommend using the alanized protein and rolling a ball of 2.9 Å instead of the standard 1.4 Å (supplementary material).

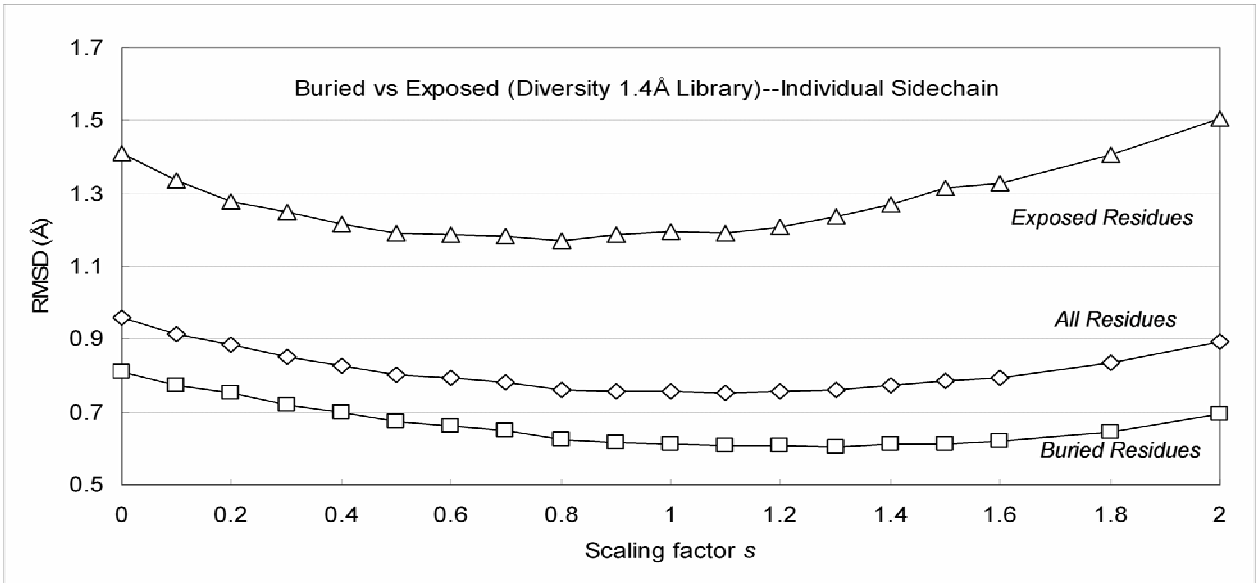


Figure 2-4 The effects of varying the scaling factor s on placement accuracies for the exposed and core residues. Shown are results from the 1.4Å diversity rotamer library results. Exposed residues account for approximately 25% of all residues.

Rotamer Library	Optimal Scaling Factor s for core residues	Optimal Scaling Factor s for surface residues	Core residue RMSD (Å) for optimal s	Surface residue RMSD (Å) for optimal s
0.2Å	1.4	0.6	0.309	0.939
0.6Å	1.2	0.8	0.414	1.010
1.0Å	1.2	0.9	0.515	1.107
1.4Å	1.3	0.8	0.605	1.171
1.8Å	1.2	0.7	0.742	1.227
2.2Å	0.8	0.6	1.105	1.371
3.0Å	0.7	0.6	1.439	1.625
5.0Å	0.7	0.7	1.835	1.935
All-Torsion	0.9	0.8	0.656	1.224

Table 2-6 Accuracy comparison in single sidechain placements for buried and exposed residues for the Xiang test set.

2.3.3 Placement of All Sidechains on Proteins, Comparison with SCWRL

The effectiveness of the flat-bottom potential in the single-placement setting extends to multiple sidechain placements. Based on the same Xiang test set of 33 proteins, we report the placement accuracy shown in Figure 2-5. The optimal *s* values were similar to the values from single placement tests. For example, the 1.0 Å library had an optimum scaling factor s=1.0 leading to an accuracy of 0.747Å (compared to 0.665 Å for single placement). Overall, the accuracy discrepancy in multiple placement and single placement setting comes to a 0.09Å RMSD. Using the $\chi 1/\chi 2$ criterion leads to similar conclusions, as seen in Table 2-8.

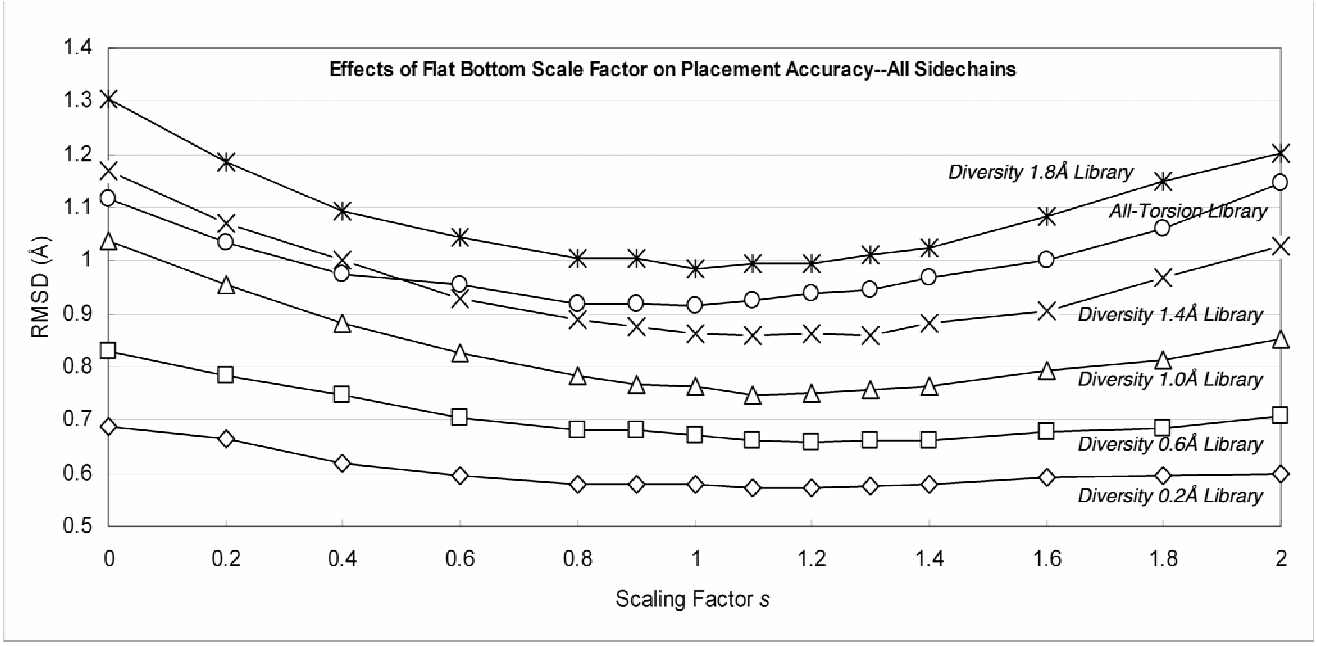


Figure 2-5 Accuracy for simultaneously replacing all sidechains for various rotamer libraries at different s values. Shown are the libraries of 0.6Å diversity (3195 rotamers), 1.0Å diversity (1014 rotamers), 1.4Å diversity (378 rotamers), 1.8Å diversity (218) and all-torsion (382 rotamers).

The overall improvement in RMSD of the optimal *s* values over the exact Lennard-Jones potential, however, is more dramatic than in the single placement tests. For instance, by introducing the optimal *s value* for the float-bottom potential, in the single sidechain placement case, the accuracy improved from 0.834Å to 0.663Å, an improvement of 0.17Å; in the all-sidechain placement case, the improvements went from 1.024Å to 0.755Å, an improvement of 0.27Å.

To compare our results with SCWRL, we applied SCWRL3.0 on the Xiang set of proteins. We found an accuracy of 0.85Å for SCWRL. A direct comparison between SCREAM and SCWRL is difficult since SCWRL uses a backbone dependent rotamer library and a more sophisticated multiple sidechain placement algorithm. However, we note that the 1.8Å SCREAM library, with just 214 rotamers, achieved an accuracy of 0.86Å RMSD which is

comparable to the 0.85 Å for SCWRL, which has a rotamer for each major torsion angle, coming to ~370 rotamers. Of course, SCWRL uses a backbone dependent rotamer library, so the specific torsion angles of those rotamers depend on the backbone φ-ψ angles.

Library	Number of Rotamers	Unmodified Potential (RMSD, Å)	Best Scale Factor *s* value	Best RMSD (Å)
0.2Å	14755	0.689	1.2	0.571
0.6Å	3195	0.830	1.2	0.657
1.0Å	1014	1.036	1.1	0.747
1.4Å	378	1.171	1.1	0.860
1.8Å	214	1.303	1.0	0.985
2.2Å	136	1.545	0.9	1.278
3.0Å	84	1.756	0.8	1.565
5.0Å	44	1.987	0.6	1.909
AllTorsion	382	1.118	1.0	0.916
SCWRL	0.951Å			

Table 2-7 Optimized s value for rotamer libraries of size ranging from 0.2Å to 5.0Å, plus the all torsion rotamer library. The scaling factor s that gives the best RMSD value is included. For comparison, SCWRL gives a RMSD of 0.95Å for the same residues and proteins tested in this set.

Library	Number of Rotamers	χ1/χ1+2 accuracy from unmodified scoring function	Optimal s value	χ1/χ1+2 accuracy using optimal s
0.2Å	14755	91.4% / 86.6%	1.3	94.1% / 89.9%
0.6Å	3195	89.7% / 83.0%	1.1	93.8% / 88.5%
1.0Å	1014	84.5% / 75.6%	1.1	92.9% / 86.7%
1.4Å	378	81.7% / 71.4%	1.3	92.1% / 84.3%
1.8Å	214	77.4% / 67.3%	1.2	88.6% / 80.0%
2.2Å	136	66.8% / 55.0%	1.1	75.7% / 64.6%
3.0Å	84	60.6% / 50.5%	0.8	66.2% / 56.7%
5.0Å	44	52.1% / 43.9%	0.6	54.3% / 45.7%
AllTorsion	382	85.0% / 73.4%	1.0	89.7% / 81.5%
SCWRL	86.4% / 79.7%			

Table 2-8 Effect of s values on χ1/χ1+2 accuracy. Rotamer libraries of diversity ranging from 0.2Å to 5.0Å, plus the all torsion rotamer library are used. The best value for χ1+2 correctness is used to determine the most effective s value. A χ angle is considered correct if within 40° of the corresponding χ angle in the crystal sidechain conformation. The χ1/χ1+2 correctness for SCWRL is 86.4% / 79.7%.

2.3.4 Analysis of Impact of Flat-Bottom on Individual Amino Acids during Combinatorial Placement

We analyzed the impact of flat-bottom on accuracies for each individual amino acid. We have also optimized individual scaling factors for each amino acid based on the Xiang set for each library. This approach is not included in the current SCREAM software but the optimum scaling factors are included in Appendix A.

2.3.5 Effects of Minimization on Structures from Different Scaling Factors

For efficiency in predicting the optimum combination of sidechain conformations, we use the discrete rotamers from the library with no minimization. Because of this, the closest (in

terms of RMSD) rotamer in the library to the correct conformation may have short contacts. That is why we use the flat-bottom potential. Of course, after assigning the sidechains we need to optimize the structures in preparation for docking and other applications. To assess how well this optimization improves the accuracy we have minimized the sidechains for each structure for 100 steps (using DREIDING in vacuum) with the results in Table 2-9.

We see that the initial configurations often have very high energies but after minimization these energies become fairly similar for different scaling factors with the same diversity. As expected, the best energies (in bold face) generally come from a scaling factor of 1.0 or 1.1. We note also that as the diversity of the library decreased, the energy of the final optimized configurations also decreased, indicating increased accuracy.

As expected, the RMSD also decreases as we minimize the structures. These results are shown in Table 2-10. For example, for the 1.0A library, accuracy improved from 0.747A to 0.625A.

s value	0.6Å Library		1.0Å Library		1.4Å Library		All-Torsion Library	
	Starting Energy	Minimized Energy	Starting Energy	Minimized Energy	Starting Energy	Minimized Energy	Starting Energy	Minimized Energy
0	-1234.3	-3163.1	546.8	-2839.2	6957.0	-2544.8	1558154.0	-2317.1
0.2	-2237.0	-3225.5	530.7	-2969.3	2804.0	-2675.2	1260675.0	-2515.2
0.4	-2195.1	-3271.3	417.6	-3053.8	**2610.3**	-2790.4	34774.5	-2767.6
0.6	**-2364.8**	-3312.2	**-624.4**	-3102.8	3454.9	-2871.2	**34628.7**	-2826.2
0.8	-2227.6	-3328.1	-419.9	-3168.6	4970.1	-2929.7	41225.3	-2849.5
0.9	-2130.1	-3325.0	-166.4	-3165.1	10013.7	-2941.8	166369.5	-2836.7
1.0	-2041.5	**-3331.6**	143.2	-3166.3	132017.6	**-2952.7**	173157.0	**-2854.6**
1.1	-1952.9	-3341.3	1431.4	**-3177.5**	136424.5	-2945.5	53846.7	-2845.
1.2	-1764.6	-3338.9	1885.2	-3171.0	146372.5	-2938.1	62057.7	-27949
1.3	-545.0	-3327.5	3278.3	-3161.9	161903.0	-2919.4	101904.8	-278.0

Table 2-9 Average energy values for the 33 proteins over varying s values. All energy values include valence and non-valence terms, and the units are presented in kcal/mol. The energies do not include interaction terms between atoms that are not involved in the sidechain placement calculations. Numbers in bold are the minimum values for each category.

Scaling Factor	0.6Å Library		1.0Å Library		1.4Å Library		All-Torsion Library	
	Starting RMSD	Minimized RMSD	Starting RMSD	Minimized RMSD	Starting RMSD	Minimized RMSD	Starting RMSD	Minimized RMSD
0	0.830	0.737	1.036	0.930	1.171	1.061	1.112	1.003
0.2	0.784	0.694	0.954	0.848	1.071	0.962	1.035	0.916
0.4	0.746	0.658	0.884	0.773	**1.003**	0.887	0.975	0.848
0.6	**0.706**	0.615	**0.827**	0.718	0.930	0.814	**0.954**	0.823
0.8	0.681	0.591	0.784	0.668	0.888	0.767	0.920	0.787
0.9	0.682	0.591	0.766	0.651	0.877	0.752	0.917	0.786
1.0	0.672	0.581	0.764	0.647	0.863	**0.736**	0.916	**0.780**
1.1	0.662	**0.569**	0.747	**0.625**	0.860	0.729	0.923	0.786
1.2	0.657	0.562	0.752	0.629	0.861	0.727	0.937	0.799
1.3	0.662	0.568	0.758	0.632	0.860	0.724	0.946	0.803

Table 2-10 Average RMSD values (in Å) for the Xiang set of 33 proteins, before and after minimization. Entries in bold correspond to those with the lowest DREIDING energies before and after minimization, see Table 2-9 for details.

2.3.6 Program Execution Performance

All tests have been run on Intel Xeon 2.33 GHz CPU single processors. The tradeoff in time vs. rotamer library size is detailed in Table 2-11. Obviously, the size of rotamer libraries affects the time spent on sidechain placement. Compared to SCWRL, the time required by SCREAM is relatively slow. However, SCWRL does not explicitly include hydrogen atoms, and use of united atom should reduce the computational time by SCREAM by a factor of about three[47].

It might appear that the increased accuracy of using SCREAM compared to SCWRL might not justify the increased expense. However, these test cases are all systems for which exact

structures are available. We have found in applications involving predictions of new structures that the SCREAM procedure works better than SCWRL, in particular for predicting GPCRs, as will be presented elsewhere[52].

Library Diversity	Number of Rotamers	Time per protein	X_1 (%)		χ_{1+2} (%)		RMSD (Å)	
			Buried	All	Buried	All	Buried	All
0.2Å	14755	554 s	96.7	93.8	93.7	89.7	0.43	0.58
0.6Å	3195	291 s	96.1	93.5	91.6	88.0	0.53	0.67
1.0Å	1014	146 s	95.5	92.4	89.8	85.9	0.62	0.76
1.4Å	378	110 s	94.4	91.6	87.0	83.8	0.73	0.86
1.8Å	214	1 s	90.9	87.8	83.4	80.0	0.85	0.99
All-Torsion	382	147 s	92.4	89.7	85.2	81.5	0.78	0.92
SCWRL	n/a	3 s	90.3	86.4	84.4	79.7	0.79	0.95

Table 2-11 Performance measure of SCREAM, with rotamer libraries of various diversities. The timing statistics were taken from the runs that gave the best energy values.

2.3.7 Tests on the Liang Set Using the Optimized Scaling Factor

In the previous sections, we optimized the scaling factors for the Xiang set and discussed the accuracy for the Xiang set. As to better indicate how well SCREAM works for new systems we tested the predictions for the Liang set using the scaling factors optimized for the Xiang set.

Rotamer libraries of practical use, including those of diversities 0.6 Å, 1.0Å, 1.4Å, 1.8Å and the all-torsion rotamer library were used for this test. Results are shown in Table 2-12. For example, using the 1.4A library, we found an accuracy of 0.96Å for all residues and 0.74Å for the buried residues, which compares to 0.86Å for all residues and 0.73Å for the

buried residues for the Xiang set. The reason for the decreased accuracy is that 40% of sidechains in the Liang set are solvent exposed compared to 25% for the Xiang set. The prediction of core residues is approximately at the same level of accuracy as reported in previous sections.

Library Diversity	Number of Rotamers	Run time per protein	χ1 (%)		χ1+2 (%)		RMSD (Å)	
			Buried	All	Buried	All	Buried	All
0.6Å / s = 1.2	3195	78.9 s	96.4	90.8	92.6	84.3	0.52	0.80
1.0Å / s = 1.1	1014	41.0 s	93.6	89.1	87.1	80.7	0.69	0.93
1.4Å / s = 1.1	378	29.9 s	94.5	89.4	86.2	79.9	0.74	0.96
1.8Å / s = 1.0	214	27.6 s	90.3	85.2	83.5	77.0	0.84	1.05
All-Torsion / s = 1.0	382	32.5 s	93.4	87.6	87.3	79.4	0.77	0.99
SCWRL	n/a	2 s	90.5	83.7	84.3	75.5	0.82	1.10

Table 2-12 SCREAM predictions on the Liang test set using optimized scaling factor for rotamer libraries of various diversities. The percentage of buried residues in this test set is about 40%, greater than the 25% figure from the previous test set. We include crystal structure solvents in the predictions, and the increase in exposed residues is due to the fewer resolved solvents in those structures.

2.3.8 Parameters for Other Lennard Jones Potentials

While the Lennard-Jones 12-6 potential is the most commonly used, it has been demonstrated that softer potentials improve placement accuracy[53]. Thus, we tested out Lennard-Jones potentials of the 7-6, 8-6, 9-6, 10-6 and 11-6 types on the 1.0Å rotamer library for the Xiang protein set. As expected, the softer potentials performed better, but the results can be improved further by including a flat-bottom region in the potential. Results are shown in Table 2-13. The optimal value of the scaling factor *s* decreases with

softer Lennard-Jones potentials, which was expected and was consistent with the flat-bottom potential approach. It is interesting to note that the 11-6 potential with optimized scaling factor *s* achieved the best overall RMSD value for this test, though the differences across the different Lennard-Jones potentials were small.

LJ Type	Unmodified Potential (RMSD, Å)	Best Scale Factor *s* value	Best Scale Factor RMSD (Å)
7-6	0.831	0.4	0.767
8-6	0.845	0.6	0.752
9-6	0.855	0.7	0.752
10-6	0.911	0.8	0.749
11-6	0.963	1.0	0.741
12-6	1.036	1.1	0.747

Table 2-13 Effect of different Lennard-Jones potentials and their optimal scaling factor s. Tests were done on the Xiang protein set using the 1.0Å rotamer library.

2.3.9 Comparison with VDW Radii Scaling

We also test out using reduced VDW radii values on the 1.0Å rotamer library for the Xiang protein set. The results are shown in Table 2-14. The improvement from using reduced VDW radii is not as pronounced as the improvement from using softer Lennard-Jones potential forms, described in the previous section.

VDW Radii Scaling	RMSD (Å)
75%	0.959
80%	0.884
85%	0.866
90%	0.896
95%	0.956
100%	1.036

Table 2-14 Effects of VDW scaling. Tests were done on the Xiang protein set using the 1.0Å rotamer library.

2.3.10 Extension beyond the Natural Amino Acids

The σ values were calculated for the natural amino acids. To extend the flat-bottom potential approach for ligands and non-natural amino acids, a value for Δ or σ needs to be determined. These values clearly depend on how conformations were generated, but we recommend a simple scheme such as using Δ=0.4Å for all atoms.

2.4 Conclusion

We show that sidechain placement using a flat bottom potential leads to excellent sidechain placement results with a simple combinatorial sidechain placement algorithm. We present a straightforward method for deriving these parameters and applied this to rotamer libraries with a wide range of diversities (0.2Å to 5.0Å). The potential is a simple modification of a Lennard-Jones potential, making it easy to incorporable into existing software. In a later chapter, we present the protein design of Trptophanyl-tRNA synthetase active site to incorporate non-natural amino acids, extensively using SCREAM in that application.

3 DREIDING for Polar Interactions

3.1 Introduction

Molecular mechanics force fields, such as CHARMM[13], AMBER[15], GROMACS[54], OPLS[55] and DREIDING[14] have played an essential role in the successful applications in many computational protein studies in the past few decades[26,56]. Thanks to extensive parameterization, these force fields can reproduce accurate molecular structures. This is especially true for the valence terms such as bonds, angles, and torsions, since high level *ab initio* quantum mechanics calculations are available for parameter fitting.

Intermolecular interactions are more difficult to parameterize. Factors such as charge transfer, polarization and hydrogen bonding make it difficult for standard force fields to accurately describe polar interactions in particular. There is in general a lack of agreement on how to assign partial charges on atoms, and charges alone sometimes do not contribute to the accuracy of structures. Polarizable force fields present a plausible solution to this problem, but their development is still in their infancy[54,57,58]. In addition, hydrogen bonds are highly directional and it is difficult to reproduce quantum mechanical energies given the constraints set forth by electrostatics and van der Waals.

Because of the importance of polar interactions in protein systems, DREIDING, unlike many other force fields in common usage, explicitly includes a hydrogen bond. The original expression of the DREIDING hydrogen bond expression, which was formulated over 30 years ago, has no physical basis and we have updated this hydrogen bond expression based on quantum mechanical calculations on water dimer. Based on this new DREIDING hydrogen bond term, we introduce a consistent approach in describing polar interactions in proteins by performing our

parameterization based on quantum mechanical calculations on model compounds. We achieve excellent results with this approach.

In addition, we recommend neutralizing charged protein residues during calculations to reduce unwanted noise arising from electrostatics. The procedure and justification for this strategy is presented.

3.2 Polar Interactions

3.2.1 Neutralizing Amino Acids

The following amino acids have a net charge at physiological pH: Aspartate (Asp, $pK_a = 4.0$) and Glutamate (Glu, $pK_a = 4.0$) have a net charge of -1, Arginine (Arg, pK_a 12.5) and Lysine (Lys, $pK_a = 10.5$) have a net charge of +1 (Figure 3-1). These residues are typically modeled as charged in force fields such as CHARMM and AMBER. Since proteins might not necessarily have an equal number of positively and negatively charged residues, counter-ions such as Na^+ and Cl^- are added to a simulation system to ensure net neutrality in the system when doing molecular simulations.

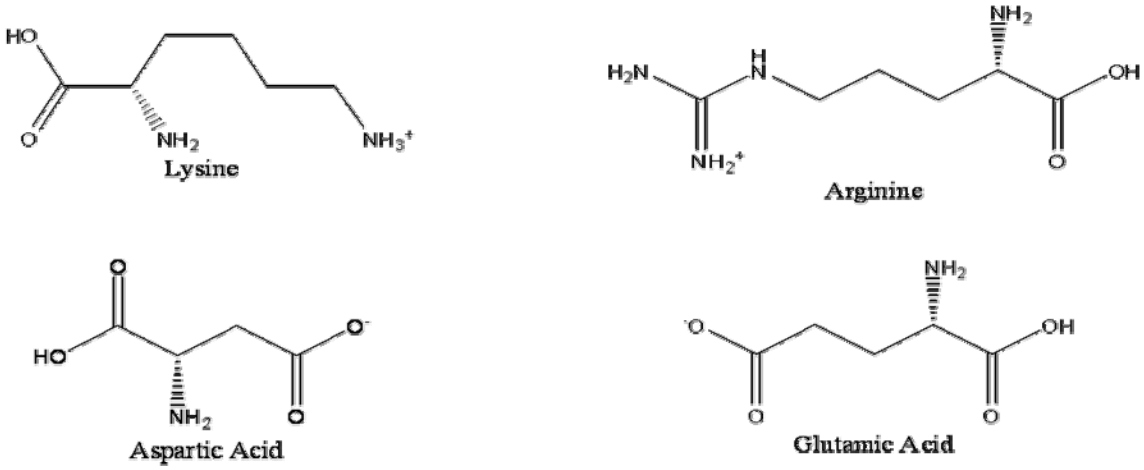


Figure 3-1 The four charged amino acids at physiological pH.

Despite the use of counter-ions, it is recognized that they bring about large fluctuations in electrostatic energies[59]. As mentioned in Chapter 1, electrostatic interaction in force fields is described like the following:

$$E_Q = (322.0637)Q_i Q_j / \varepsilon R_{ij}$$

where the energy is expressed in kcal/mol and R_{ij} is expressed in Å. According to this relation, if we take ε=1.0, two oppositely charged ions even at 30Å apart would have an interaction energy close to 10kcal/mol, which is comparable to the 5-10kcal/mol stability of a typical protein. This level of sensitivity to changes far away from a region of interest such as a ligand binding site leads to unreliable calculations. As a result, researchers have used various schemes to reduce the effect of electrostatics, such as increasing the value of the dielectric constant and using a distance dependent dielectric[60-62]. This is despite the fact that the use of dielectric constants other than 1.0 or a modified Coulombic interaction is justified only on empirical grounds.

Interaction Type	QM Energies (kcal/mol)	Force Field Energies (kcal/mol)
Guanadinium—Carboxylate	120.9	85.5
Amine–Carboxylate	145.1	94.5

Table 3-1 Interaction energies of model charged compounds (see Figure 3-7) representing charged amino acids. Quantum mechanics energies were obtained by constraining the hydrogen-heavy atom distance.

With the considerations mentioned above in mind, we neutralize the net charges on the normally charged amino acid residues. It has been demonstrated empirically and statistically that solvent exposed charged sidechains and exposed salt bridges contribute less than 2 kcal/mol to the self-energy of a protein[63,64], due to the solvent screening effect. For buried salt-bridges, fully charged models leads to an interaction energy of about 90kcal/mol (Table 3-1), which is an unrealistic figure for binding energy calculation purposes. Instead, we perform a proton transfer between the salt bridges. To obtain the correct interaction, parameterization of this salt bridge pair is done by fitting to QM results. Our approach not only reduces the source of error that arises from the use of counter-ions, but also reduces the effect due to random fluctuation of charged residues far away from a region of interest.

3.2.2 Assignment of Protons from Neutralization of Charged Residues

Care is needed when hydrogens are added or removed from charged residues, as the original hydrogen bond network must remain undisrupted. The addition and deletion of protons are needs to satisfy the following rules:

1. Protons are to be transferred from positively charged residues (or functional groups) to negatively charged residues (or functional groups).

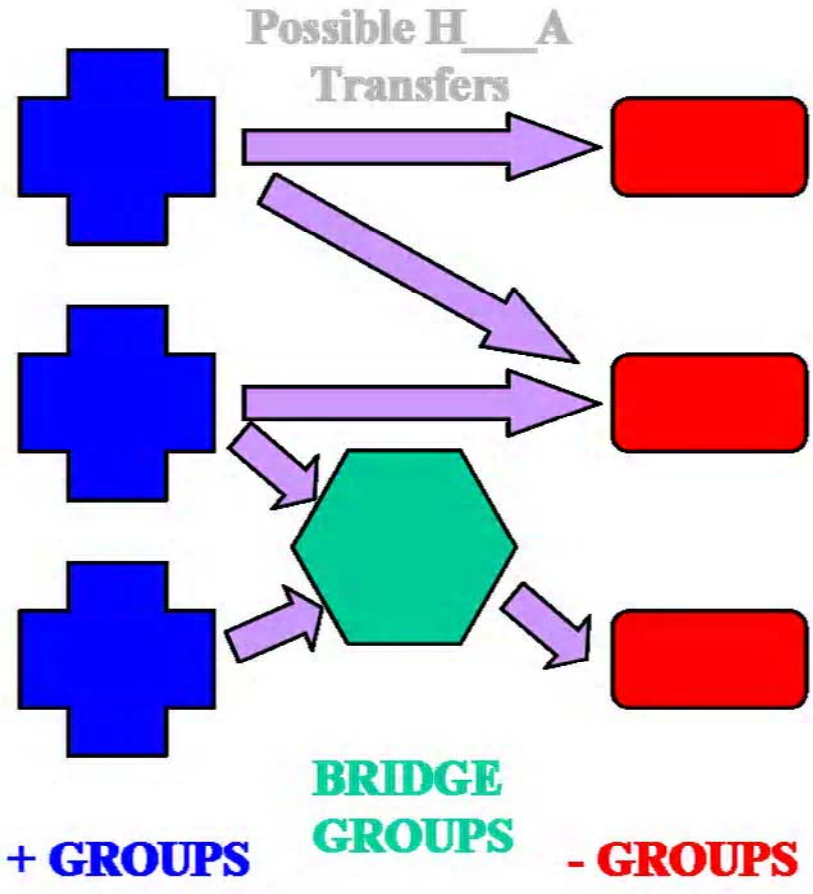
2. If a positively charged residue is not directly involved in a salt bridge with a negatively charged residue but belong to the same hydrogen bond, protons are allows to be transferred via bridging groups, such as water or histidine.

3. If charged residues are isolated even considering its hydrogen bond network, protons are added or deleted so as to maximize the number of hydrogen bonds.

4. If ambiguity arises, such as when a charged residue is involved in two salt bridges, proton assignment is performed by maximizing the total number of hydrogen bonds. See Figure 3-2 for an illustration.

With the rules established above, the problem of proton assignment becomes equivalent to a maximum flow problem, a problem well-studied in computer science. The entire description of the algorithm is in Appendix B.

Figure 3-2 Schematic illustrating the movement and assignment of protons. Arrows denote possible polar hydrogen "+ Groups": functional groups that are positively charged, i.e. those with extra protons. "- Groups": functional groups that are negatively charged, i.e. those that can accept protons. "Bridge Groups": groups such as H_20 or Histidine that can form hydrogen bonds between + Groups and – Groups.



The overall charge of each residue after removing or addition of proton thus becomes zero. As a result of this neutralization procedure, the protein also has a net charge of zero. Calculations of binding energies or effects of mutations are then carried out using a dielectric of 1.0. Clearly, with the elimination of net charges on functional groups, systems like salt bridges are not as stable as prior to the neutralization procedure. Therefore, we have created new atom subtypes for these special neutralized residues (See Table 3-2). We discuss these types in more detail in a subsequent section.

Residue Type	Affected Atoms	Previous Atom Type	Neutralized Atom Type
ASP	$O_{\delta 1}$, $O_{\delta 2}$	O_2	O_2M, O_3M
GLU	$O_{\varepsilon 1}$, $O_{\varepsilon 2}$	O_2	O_2M, O_3M
LYS	N_{ζ}	N_3	N_3P
ARG	$N_{\eta 1}$, $N_{\eta 2}$	N_R	N_RP

Table 3-2 New atom subtypes for atoms on previously charged residues. The suffixes "M" and "P" at the end of a neutralized atom type stand for "Minus" and "Plus", mnemonics for remembering the original net charge of the residue the atom belongs to. For ASP and GLU, the atom type O_3M is used on the oxygen with a proton added onto it.

3.3 DREIDING Hydrogen Bond Term

3.3.1 Introduction

Hydrogen bonds play an important role in biological molecules, as they are largely responsible for the formations of organized three-dimensional structures. In proteins, alpha helices and beta sheets formation are due to backbone hydrogen

bond formation. Many protein-ligand recognitions are due to specific hydrogen bonds.

Because of the importance of hydrogen bonds, an accurate description of hydrogen bond is necessary for biomolecular computations. Some force fields, including CHARMM[13] and AMBER[15], balance VDW and electrostatic interactions to reproduce hydrogen bond interactions. However, with the current force fields using point charges on atomic centers and charges being non-polarizable, it is difficult to produce energies to high accuracy. Thus, in DREIDING[14], a specialized hydrogen bond term is used:

$$E_{hb} = D_{hb}[5(R_{hb} / R_{DA})^{12} - 6(R_{hb} / R_{DA})^{10}]\cos^4(\theta_{DHA})$$

where D_{hb} stands for the well-depth of the hydrogen bond potential, R_{hb} the equilibrium distance and θ_{DHA} the angle between the hydrogen bond donor atom, hydrogen and the acceptor atom.

This hydrogen bond term consists of a 12-10 Lennard-Jones radial term and a $\cos^4$ angular term, with each term independent of the other. We investigate whether this 12-10 Lennard-Jones distance term and the $\cos^4$ angular term accurately describe the behavior for a hydrogen bond interaction. The water dimer is chosen as our model system.

3.3.2 Updating the DREIDING Hydrogen Bond Term

3.3.2.1 Equilibrium Water Dimer Structure

We perform high-level quantum mechanics (X3LYP/aug-cc-pvtz) calculations on water-dimer. The minimized structure is shown in Figure 3-3. A binding energy of 5.0 kcal/mol is obtained for the optimized structure, with the binding energy (E_{bind}) being calculated as:

$$E_{bind} = E_{dimer} - E_{donorWater} - E_{acceptorWater}$$

where E_{dimer} is the energy of the water dimer, $E_{donorWater}$ the energy of the hydrogen bond donating water monomer and $E_{acceptorWater}$ the energy of the hydrogen bond accepting water monomer. The counterpoise correction for basis set superposition error is not included because the basis sets used is of the Dunning[65] type.

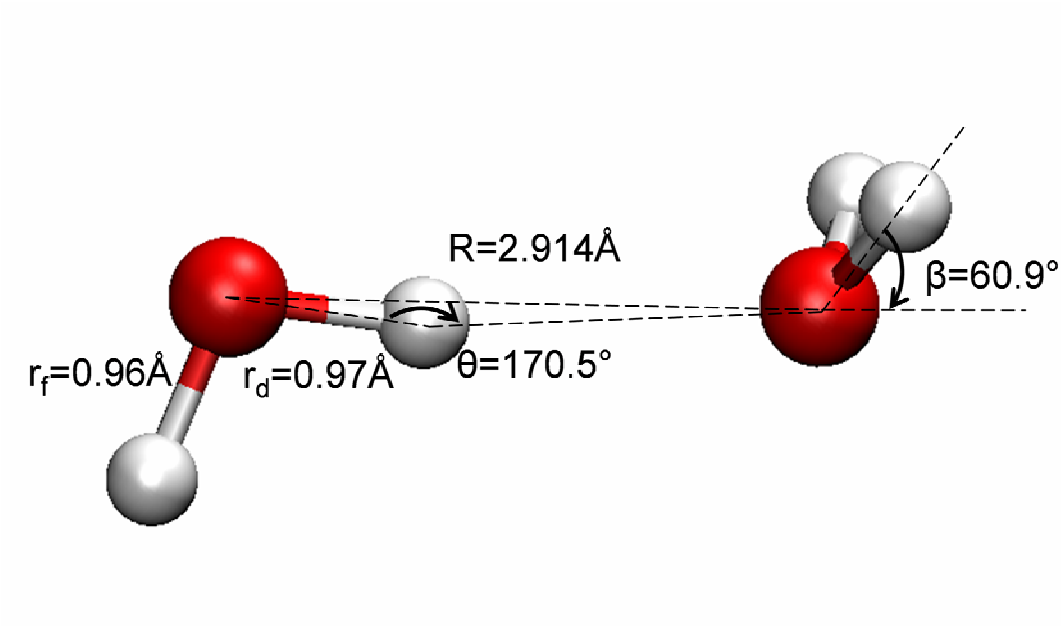


Figure 3-3 Water dimer structure optimized at the X3LYP/aug-cc-pvtz level.

3.3.2.2 Assignment of Charges

A consistent charge assignment scheme is necessary to ensure transferability of parameters. We choose to use Mulliken atomic charges from quantum mechanics. While we used the Dunning type of basis-set for minimization, those basis sets are too diffuse and instead we use the Pople type basis sets for obtaining the atomic Mulliken charges.

Basis Set	Mulliken Charge on Oxygen (e)	Dipole Moment with Mulliken Charges (Debye)
6-311G**	-0.48	1.35
6-31G**	-0.61	1.72
Aug-cc-pvtz-f	-0.67	1.91
6-31G/HF	-0.87	2.43
Experiment	n/a	1.8

Table 3-3 Mulliken charges for water, when using different basis sets. Except for the 6-31G basis set, where the Hartree-Fock method is used, all other calculations are done using the X3LYP method. The dipole moment is calculated by placing the charges on water molecule atoms.

As shown in Table 3-3, the basis set employed in a calculation has a large effect on the charge assignment. The 6-31G/HF calculation is included for comparison purposes, since the charges used in CHARMM is based on these calculations. We have chosen the 6-31G** basis set for all subsequent Mulliken charge assignments in this work. While this basis set underestimates the experimental dipole moment for water (also in general for other polar molecules as well, data not shown), we can easily capture and correct for any missing interaction energies by proper parameterization of the hydrogen bond term in DREIDING.

3.3.2.3 Water Dimer Radial Dependence

We examined what functional form best reproduces the radial dependence of water dimer binding by varying the O-O heavy atom distance at 0.01Å increments and performing quantum mechanics at the X3LYP/cc-pvtz(f++) level. To reduce uncertainty and also because the angular dependence of the DREIDING hydrogen bond has a minimum at 180°, we fix the O-H…O angle to 180.0° (i.e. not at the equilibrium 170.5°) in quantum mechanics calculations. The comparison with

force field one-point energies is done using DREIDING parameters for water (Table 3-4).

Atom Type	D_0 (kcal/mol)	R_0 (Å)	Charges (e^-)
H___A	0.0335	2.9267	+0.3066
O_3W	0.9570	3.4046	-0.6132
O_3W – H___A (off-diagonal term)	0.0001	3.1566	n/a

Table 3-4 DREIDING atom types for oxygen (O_3W) and hydrogen (H___A) in water. Mulliken charges for the water monomer are based on calculations at the X3LYP/aug-cc-pvtz(-f) level. The R_0 off-diagonal O_3 – H___A term is the geometric mean of the two R_0 values for H___A and O_3.

DREIDING previously only includes an O_3 atom type, one that represents sp^3 oxygen. We feel that the water oxygen demands its own atom type, and have used O_3W for the water oxygen, where the "W" stands for water. Except for the properties described in Table 3-4, all other parameters of this O_3W atom type including the valence terms, are kept the same as parameters used for the O_3 atom type.

We tested various common VDW potentials including the 12-6, 11-6, 10-6, 9-6, 8-6, 7-6 Lennard-Jones forms, Morse potential and exponential-6 potential. From these potentials, we chose the Morse potential to describe the radial portion of the hydrogen bond, discarding the 12-10 potential previously. The family of Lennard Jones potentials proved to have too great of an inner repulsion wall to compared to quantum mechanics data. An extremely soft Lennard-Jones 6-4 potential does a good job of fitting, but the $1/r^4$ long term behavior will overwhelm the 12-6 Lennard Jones in VDW calculations. The exponential-6 potential was discarded because of complexities involving an inflection point at small distances. A fit using

the Morse potential against quantum mechanics calculations done on the water dimer is shown in Figure 3-4.

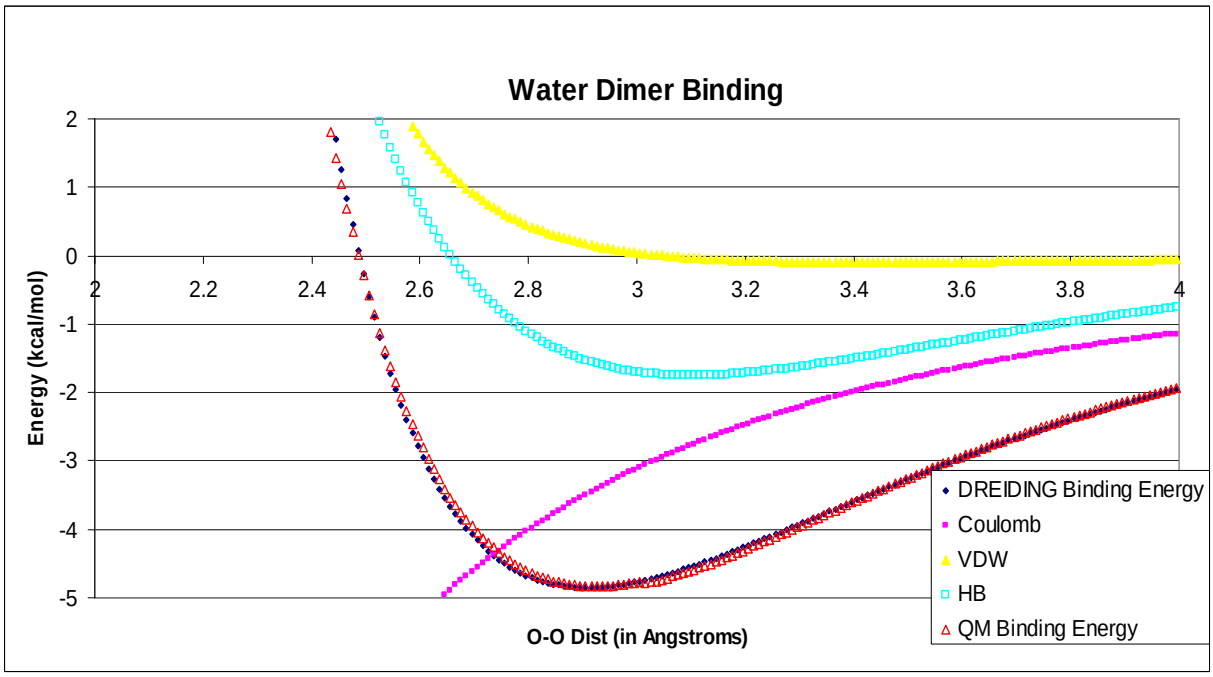


Figure 3-4 Water dimer binding, QM vs. DREIDING using fitted parameters with a Morse hydrogen bond potential with γ=9.70, R_0=3.10 and D_0=1.75.

We chose the Morse potential for two reasons. The γ value in the Morse potential allows one to adjust the steepness of the inner repulsive region of a two-body interaction. In addition, the Morse potential has the desirable asymptotic property that it goes to zero more rapidly than any functional form containing a term with a power of R. In practice, this asymptotic property would be irrelevant because of cutoffs, but we do not wish for cutoffs to determine such a fundamental issue. The Morse potential has the following form:

$$E_{Morse}(R) = D_{HB}[\chi^2 - 2\chi]$$

where $\chi = e^{\left[-\frac{\gamma}{2}\left(\frac{R}{R_{HB}}-1\right)\right]}$.

Here, D_{HB}, R_{HB} and γ are parameters, with D_{HB} being the well-depth of the potential, R_{HB} the equilibrium distance between the hydrogen bond acceptor and donor heavy atoms, and γ a scaling parameter. Using the VDW and electrostatics terms above, γ=9.70 leads to the overall DREIDING water dimer binding energy to be zero at the same radial distance as in quantum mechanics. This γ value will be used for all other pairs of hydrogen bond donor-acceptor pairs.

As noted previously, the H___A – O_3W off-diagonal term (or "cross" term) is specified to have parameters of R_0=3.1566 and D_0=0.0001. The R_0 value is based on the geometric mean of H___A and O_3 R_0 parameters. The small D_0 value is chosen out of necessity. With the default geometric mean combination rule, D_0 between H___A and O_3W would have VDW repulsion energy too great at small distances (around 1.5Å). This adjustment in the off diagonal term is made so that the DREIDING water dimer binding curve can reproduce QM results (Figure 3-4).

3.3.2.4 Angular Dependence

Next, we determine the optimal functional form for the hydrogen bond angular dependence $f(\theta_{AHD})$. We generated 360 water dimer structures by rotating the hydrogen bond donating water around the oxygen at 1 degree increments. The starting position for the hydrogen donating to the acceptor water is such that the O-H…O angle (θ angle, Figure 3-3) is 180°, while fixing all the other atom coordinates of the QM-optimized structure. Single point energy calculation at the X3LYP/aug-cc-pvtz(-f) is done and binding energy of the water dimer is calculated as before. The well depth for the hydrogen bond term used in the fitting is based on the binding energy of the equilibrium water dimer structure in QM. The results are shown in Figure 3-5.

From the plot, it is evident that electrostatic interactions alone cannot reproduce the QM binding curve, even in the important valley where hydrogen bonds are being formed. $\cos^4(\theta_{AHD})$ has too steep of a drop-off when θ deviates from the equilibrium 180° neighborhood. We consider the fact that $\cos^3(\theta_{AHD})$ and $\cos(\theta_{AHD})$ have derivatives that do not vanish at 90° to be a serious shortcoming. As a result, we recommend using $\cos^2(\theta_{AHD})$ functional form in the angular portion of the hydrogen bond expression.

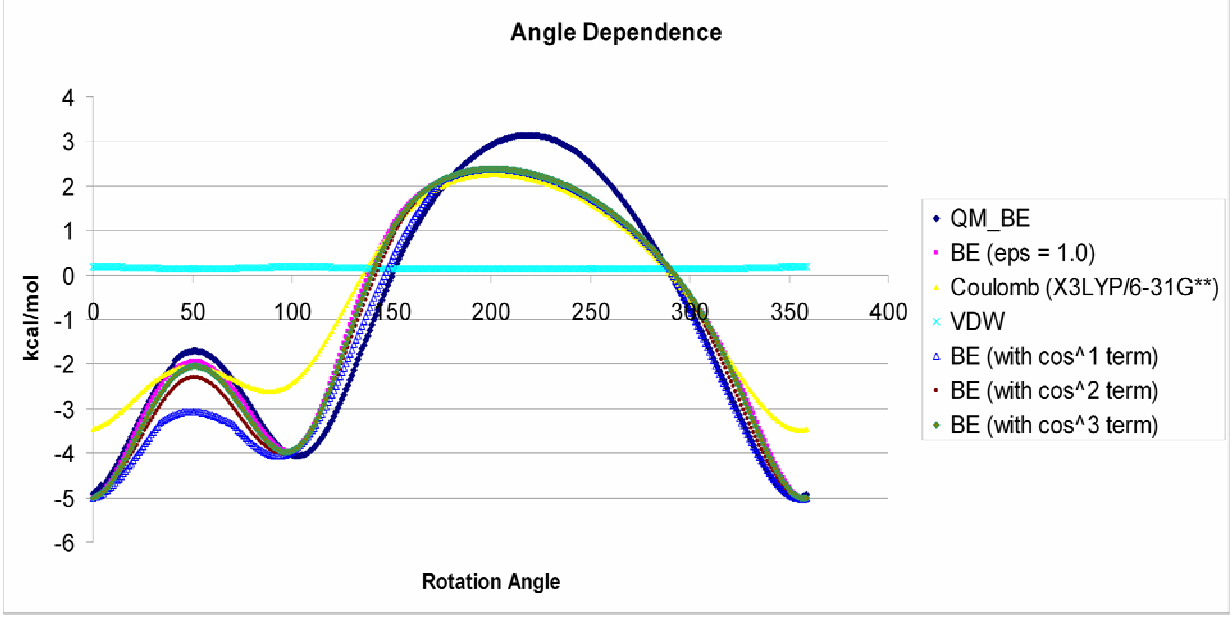


Figure 3-5 Angle dependence of water dimer binding. Note that the angle shown here is the O…O-H angle, not the O-H…O angle in θ_{AHD}.

3.3.2.5 Summary

The new functional form we have chosen for the hydrogen bond is the product of a radial and an angular component. The Morse potential is used for the radial term to allow for a softer interaction potential between the two heavy atoms. $\cos^2(\theta_{AHD})$ is used for the angular term. The off-diagonal interaction between the hydrogen bond acceptor atom and donor atom has been modified to allow the hydrogen and the

acceptor oxygen to approach each other at closer distances. The new DREIDING hydrogen bond is thus:

$$E_{hb} = D_{hb}[\chi^2 - 2\chi]\cos^2(\theta_{DHA})$$

with $\chi = e^{\left[-\frac{\gamma}{2}\left(\frac{R}{R_{HB}}-1\right)\right]}$.

3.3.2.6 Comparison with Other Water Models

We note that in some three-point water models such as F3C and TIP3P, there is no explicit hydrogen terms; instead, the VDW terms and electrostatics are parameterized to reproduce relevant water properties. The equilibrium binding energy values are listed in Table 3-5.

Model	Charge on Hydrogen (e)	Equilibrium O-O Distance (Å)	Binding Energy (kcal/mol)
F3C	0.4100	3.00	7.7
TIP3P	0.4170	2.95	8.3
DREIDING	0.3066	2.91	5.0
QM	n/a	2.91	5.0

Table 3-5 Comparison of binding energies of water dimers in various water models. Final minimized values are reported.

Most water models are developed to reproduce of bulk water properties, which would explain the differences as seen in Table 3-5. This approach may not be appropriate for calculating single-point ligand-protein binding energies, where all the atoms are in fixed positions and have undergone energy minimization. Given

our focus on ligand binding energy calculations, we derive all subsequent parameters based on quantum mechanics energies of various interacting polar dimers.

3.4 Optimization of Hydrogen Bond Parameters

3.4.1 Atom Types for Hydrogen Bond Donors and Acceptors

The assignment of atom types to atoms of the same element but in different chemical environments is crucial to the success of a force field. The more atom types introduced, the more accurate a force field becomes due to the sheer amount of additional parameterization that can be performed. We have intentionally kept the number of atom types to a minimum, in keeping with the philosophy of DREIDING—simplicity and flexibility[14].

In DREIDING, each atom type is uniquely identified as a five-character label. The first two characters denote the element (an underscore "_" is used when the element's chemical symbol is composed of just one character). The third character denotes the hybridization and geometry of the underlying atom: "1" = linear (sp^1), "2" = trigonal (sp^2) and "3" = tetrahedral (sp^3), "R" = resonance. The 4^{th} and 5^{th} characters are used for additional properties. The hydrogen involved in hydrogen bond is given the special type "H___A".

In original DREIDING, there are three types of oxygens (O_2, O_3, O_R) and two types of nitrogens (N_R and N_3) that are present in proteins. For the purpose of hydrogen bond parameterization, we have introduced several additional subtypes (Table 3-6), O_3W for the oxygen in water ("W" for *w*ater, subtype for O_3), O_2M and O_3M for the oxygens in protonated Aspartate and Glutamate ("M" for *m*inus, subtype for O_2 and O_3 respectively), N_A for the nitrogen in Trptophan and Histidine ("A" for *a*romatic, subtype of N_R), N_RP for the nitrogens in

deprotonated Arginine ("P" for *p*ositive, subtype of N_RP) and N_3P (subtype of N_3) for the nitrogen in deprotonated Lysine. In this presentation, we cover only the atom types that occur in proteins, but our parameterization strategy can be extended to the treatment of other systems such as nucleotides and organic molecules.

Each atom subtype has the exact same internal (bond length, angle, torsion) and VDW parameters as the atom type it was derived from. Only hydrogen bond parameters will be modified.

Description	Atom Type	Charge State	Donor/Acceptor?	Model Compound	Protein Examples
Water oxygen	O_3W	0	Both	H_2O	waters
sp^3 oxygen	O_3	0	Both	CH_3OH	Ser, Thr
sp^2 oxygen	O_2	0	Acceptor	CH_3CONH_2 (methyl-amide) (1)	Asn, Gln, backbone
	O_2M	-	Acceptor	CH_3COO^-	Asp, Glu
Resonance oxygen	O_R	0	Both	C_6H_5OH	Tyr
Aromatic nitrogen	N_A	0	Both	$CH_3C_3H_4N_2$ (methyl-Imidazole)	His, Trp
Resonance nitrogen	N_R	0	Donor	CH_3CONH_2	Asn, Gln, backbone
	N_RP	+	Both	$CH3NHC(NH_2)_2^+$	Arg
sp^3 nitrogen	N_3	0	Both	CH_3NH_2	Neutral Lys

	N_3P	+	Donor	$CH_3NH_3^+$	Lys
sp^3 sulfur	S_3	0	Both	CH_3SH	Cys, Met

Table 3-6 DREIDING atoms types that are used in proteins. (1): Amide was picked because ethers do not naturally occur in proteins.

3.4.2 VDW Parameters

Unlike other force fields[15,66], we do not adjust the VDW radii of individual atoms involved in hydrogen bond. As mentioned above, we only modify the hydrogen bonding parameters and keep the same VDW parameters for those atom subtypes.

Atom Type 1	Atom Type 2	D_0	R_0
O_3, O_3W, O_3M, O_2, O_2M	H__A	0.0001 kcal/mol	3.1566Å
N_R, N_RP, N_A, N_3, N_3P		0.0001 kcal/mol	3.2738Å
S_3		0.0001 kcal/mol	3.4343Å

Table 3-7 Off-diagonal VDW terms for hydrogen bond acceptors and the hydrogen bonding hydrogen (H___A). R_0 values are derived from geometric mean of heavy atom VDW radii and H___A VDW radii.

As mentioned in Section 3.3.2.3 during the discussion of the water dimer, however, the off-diagonal VDW interaction between hydrogen bond acceptor oxygen and the polar hydrogen is reduced in order for the two atoms to better approach each other. The same reduced off-diagonal approach is done for each pair of hydrogen bond acceptor heavy atom and polar hydrogen. The off-diagonal well depth, D_0, is set to 0.0001 kcal/mol. The off-diagonal R_0 is obtained by using the usual geometric mean combination rule. These values are listed in Table 3-7.

3.4.3 Hydrogen Bond Parameterization Methodology

Model compounds of atom types to be parameterized are shown in Figure 3-6 and Figure 3-7. For each hydrogen bond acceptor-donor pair interaction, a corresponding pair of model compounds is constructed. The positioning of the pair of model compounds is such that a hydrogen bond is clearly formed between them (much like the hydrogen bonding structure of the water dimer, see Figure 3-3). These structures are optimized at the X3LYP/aug-cc-pvtz(-f) level. Because of polarization, Mulliken charges are assigned each molecule individually, at the X3LYP/6-31G** level. For a discussion of charge assignment, see Section 3.3.2.2.

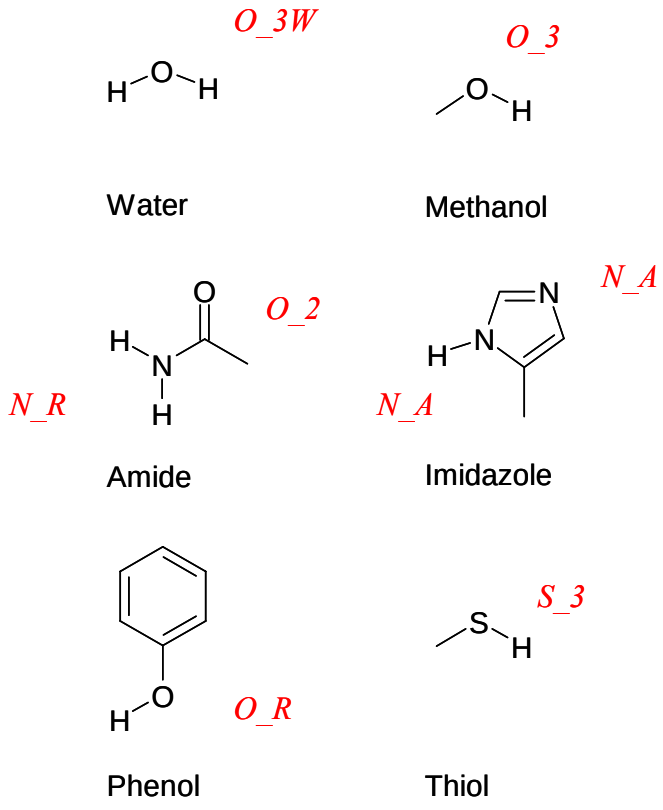


Figure 3-6 Samll molecule model compounds used in parameterization of hydrogen bonds common in proteins.

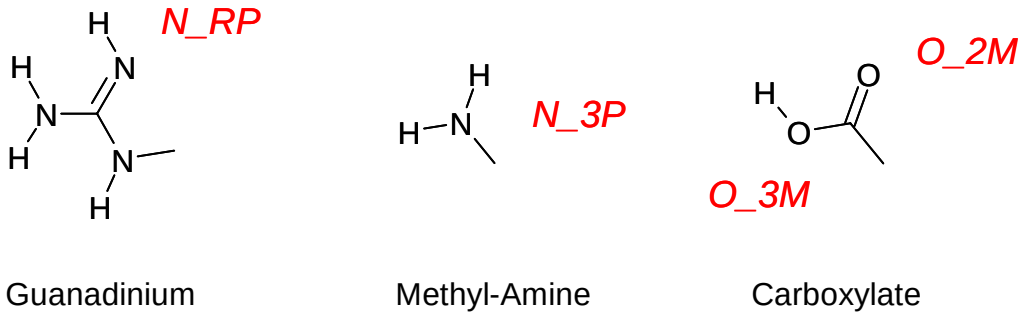


Figure 3-7 Small molecule model compounds for neutralized form of charged amino acids.

We use two pieces of data resulting from these quantum mechanics calculations for parameterization purposes. The first piece of data is the QM binding energy for the optimized structures. In our force field, the binding energy only involves non-valence terms and we have the following relation:

$$BE_{QM} \overset{Fitting}{\approx} BE_{FF} = BE_{vdw} + BE_{Coulomb} + BE_{HB}$$

where BE_{FF} is the binding energy as calculated from the force field (i.e. DREIDING), BE_{vdw} the binding energy component of VDW interactions, $BE_{Coulomb}$ the binding energy component of electrostatic interactions, and BE_{HB} the binding energy component of the DREIDING hydrogen bond term.

The second piece of data we use is the QM donor-acceptor heavy atom distance (R_{QM}). For water dimer, this value is 2.91Å. The force field donor-acceptor heavy atom distance is denoted R_{FF}.

Since we do not vary the parameters in VDW terms and 6-31G** Mulliken charges are used for charge assignment, only the parameters from the BE_{HB} term can be adjusted. Given our hydrogen bond expression (see Section 3.3.2.5) and the

decision to fix the γ parameter at 9.70, the only terms we can modify are the D_{HB} and R_{HB} terms. Modifying these terms will result in different force-field-optimized configurations, with different BE_{FF} and R_{FF} values.

Our fitting target is to achieve $BE_{FF} \approx BE_{QM}$ and $R_{FF} \approx R_{QM}$, where BE_{FF} and R_{FF} are values from the force field *after* structure optimization in that force field. Conveniently, we have two adjustable parameters D_{HB} and R_{HB} to fit to the two constraints. An exact fit can be therefore obtained, although because of the constraints imposed by not varying VDW and electrostatics the fit will on occasion not be exact. Outside of these exceptional cases, fitting is performed to within 0.1kcal/mol and 0.1Å of quantum mechanics values.

3.4.4 Parameterization of Neutral Hydrogen Bond Atom Types

From our set of possible protein atom types, there are 30 possible pairs of donor-acceptor hydrogen bond terms. Presented in Table 3-8 are the quantum R_{QM}, BE_{QM} values and the fitted R_{HB} and D_{HB} parameters.

Donor Atom Type	Acceptor Atom	Model Compounds (Donor-Acceptor)	BE_{QM} (kcal/mol) / R_{QM} (Å)	D_{HB} (kcal/mol) / R_{HB} (Å)
O_3W	O_3W	H_2O – H_2O	5.00 / 2.91	1.3 / 2.95
	O_3	H_2O – CH_3OH	5.38 / 2.88	0.6 / 2.85
	O_2	H_2O - Amide	7.11 / 2.81	1.7 / 2.675
	O_R	H_2O – C_6H_5OH	3.63 / 2.94	0.45 / 2.91
	N_A	H_2O – Me-Im	6.98 / 2.87	2.80 / 2.77
	S_3	H_2O – CH_3SH	3.81 / 3.34	2.45 / 3.29
O_3	O_3W	CH_3OH - H_2O	4.77 / 2.93	1.5 / 2.925
	O_3	CH_3OH -	5.16 / 2.90	0.8 / 2.85

		CH_3OH		
	O_2	CH_3OH – Amide	6.35 / 2.83	1.3 / 2.75
	O_R	CH_3OH – C_6H_5OH	3.434 / 3.00	0.40 / 3.09
	N_A	CH_3OH – Me-Im	6.33 / 2.89	2.70 / 2.79
	S_3	CH_3OH – CH_3SH	3.79 / 3.40	2.50 / 3.25
O_R	O_3W	C_6H_5OH – H_2O	6.21 / 2.86	2.10 / 2.85
	O_3	C_6H_5OH – CH_3OH	6.68 / 2.85	1.40 / 2.81
	O_2	C_6H_5OH – Amide	9.40 / 2.67	3.05 / 2.59
	O_R	C_6H_5OH – C_6H_5OH	4.54 / 2.92	0.0 / 0.0
	N_A	C_6H_5OH – Me-Im	8.82 / 2.80	3.90 / 2.70
	S_3	C_6H_5OH – CH_3SH	3.79 / 3.47	2.15 / 3.53
N_A	O_3W	Me-Im – H_2O	4.78 / 3.06	1.90 / 3.05
	O_3	Me-Im – CH_3OH	4.97 / 3.11	1.40 / 3.13
	O_2	Me-Im – Amide	6.54 / 3.04	2.60 / 3.08
	O_R	Me-Im – C_6H_5OH	2.99 / 3.22	0.0 / 0.0 [(2)]
	N_A	Me-Im – Me-	6.97 / 3.08	3.10 / 3.00

		Im		
	S_3	Me-Im – CH_3SH	2.06 / 3.92	0.80 / 3.93
N_R	O_3W	Amide – H_2O	4.64 / 3.07	1.5 / 3.1
	O_3	Amide – CH_3OH	5.61 / 2.95	1.1 / 2.825
	O_2	Amide – Amide	13.85 / 2.87[(1)]	2.75 / 2.77
	O_R	Amide – C_6H_5OH	4.07 / 3.10	0.60 / 3.15
	N_A	Amide – Me-Im	8.52 / 3.02	6.30 / 3.00
	S_3	Amide – CH_3SH	2.54 / 3.79	1.40 / 3.84
S_3	O_3W	CH_3SH – H_2O	1.86 / 3.76	1.20 / 3.82
	O_3	CH_3SH – CH_3OH	1.98 / 3.70	0.70 / 3.79
	O_2	CH_3SH – Amide	2.48 / 3.57	1.35 / 3.67
	O_R	CH_3SH – C_6H_5OH	0.96 / 3.60	0.0 / 0.0[(2)]
	N_A	CH_3SH – Me-Im	1.98 / 3.58	0.55 / 3.60
	S_3	CH_3SH – CH_3SH	1.1 / 4.32	0.0 / 0.0[(2)]

Table 3-8 (continued) Fitting parameters for all atom types that are present in neutral residues in proteins. The accuracy fitting is within 0.1kcal/mol in overall binding energies and 0.1Å in the equilibrium distance between hydrogen bond donor and acceptor atoms. Me-Im: Methyl Imidazole at the delta position, $CH_3C_3H_4N_2$. Amide: Methyl-Amide, CH_3CONH_2. (1) Involves two hydrogen bonds. (2)No hydrogen bond term necessary, since electrostatics is sufficient to account for the polar interaction.

Exact fit was achieved for all but 3 of these cases. In all of these 3 cases, electrostatic forces over-accounts for the actual QM interaction between the molecules and therefore no additional hydrogen bond term is needed. The error for two of these three cases, a imidazole-phenol hydrogen bond and thiol-thiol hydrogen bond, had the energy overestimated by less than 0.2 kcal/mol and donor-acceptor distance underestimates the bond distance by 0.2Å, which are very good agreements with QM nonetheless. The thiol-phenol interaction was overestimated by about 1.0 kcal/mol and the donor-bond distance underestimated by 0.3Å. This is largely due to the VDW parameters for sulfur being not large enough and difficult charge assignment for the sulfur atom[67]. Despite this, we accept a < 1kcal/mol error.

3.4.5 Parameterization for Neutralized Form of Charged Residues

A separate set of parameterization was carried out for the salt bridges. Our treatment allows the movement of a proton from the positively charged species to the negatively charged species, with the resulting interaction being stronger than a typical neutral-neutral polar interaction. As mentioned earlier, new atom types are introduced for certain atoms on ASP, GLU, ARG and LYS (See Figure 3-7.)

Donor Atom Type	Acceptor Atom	Model Compounds (Donor-Acceptor)	QM Binding Energy (kcal/mol)	/QM donor-acceptor distance (Å)	Fitted Parameters: D_{HB} (kcal/mol) / R_{HB} (Å)
N_RP	O_2M	$CH_3NHCN_2H_3$ – CH_3COOH	18.65	2.84	4.60 / 2.80
O_3M	N_RP			2.68	7.40 / 2.60
N_3P	O_2M	CH_3NH_3 – CH_3COOH	11.08	3.02	3.40 / 2.90
O_3M	N_3P			2.71	6.20 / 2.65

Table 3-9 Fitted parameters for interaction between salt bridges, allowing for proton transfer.

Parameters are reported in Table 3-9. There are two hydrogen bond parameters that need to be derived in these systems and we obtain the parameters by an iteration procedure. We start with an arbitrary value for one of the two hydrogen bonds D_{HB}/R_{HB} pair and optimize the parameters for the other D_{HB}/R_{HB} pair. We then fix the one set of optimized parameters and modify the original D_{HB}/R_{HB} pair. This procedure is repeated until we obtain a fit of within 0.1 kcal/mol in binding energy and 0.1Å for both hydrogen bond donor-acceptor distances. We use the parameters reported in Section 3.4.4 for interaction between these atom types and the usual neutral atom types.

3.4.6 Validation

The performance of DREIDING with our updated hydrogen bond is illustrated here with a number of test cases. The hydrogen bond pairs are made up of model compounds illustrated in Figure 3-8, including model compounds for the amino acid tryptophan (1), the amide protein backbone (2), ε-protonated histidine (3) (the

model compound for δ-protonated histidine was used in parameterization), methionine (4), and nucleotides (11-14). A few common organics (5-10) are also included in the set. The assignment of atom types in these molecules follow the rules described Section 3.4.1. 20 pairs of hydrogen bond forming entities are constructed from this test set and the set from Figure 3-6, with at least one molecule always from this test set. The total number of total hydrogen bonds vary from a single hydrogen bond up to a maximum of three (interaction between guanine and cytosine).

We follow the same procedure for all quantum mechanical calculations and Mulliken charge assignments as previously (X3LYP/aug-cc-pvtz(-f) level for optimization and binding energy, X3LYP/6-31G** for charge assignment). Minimization in force field is done on the structures before comparisons are made.

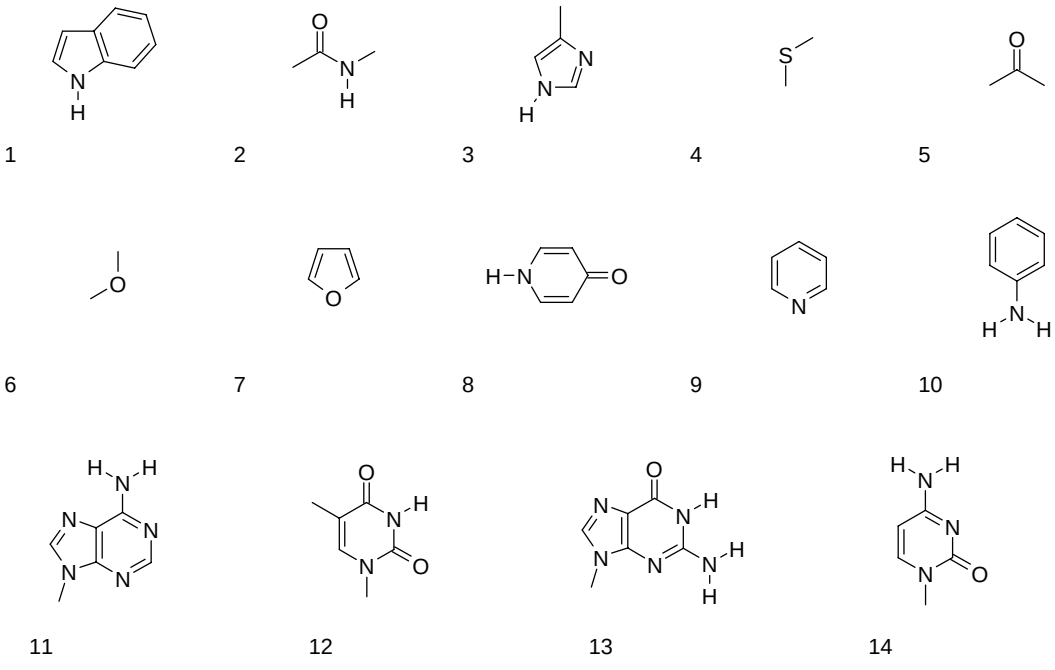


Figure 3-8 The 14 small molecules chosen in our test set.

Test set of 20 hydrogen forming pairs	Average Error (vs. QM)
Binding Energy	0.86 ± 0.59 kcal/mol
Donor/Acceptor Distance	0.08 ± 0.06 Å

Table 3-10 Average error in force field binding energies and donor/acceptor distances compared to quantum mechanics for the test set of 20 hydrogen bond forming pairs.

Results are reported in Table 3-10. A source of outlier error originates from the guanine-cytosine interaction, where the QM binding energy is 28.2 kcal/mol whereas we predicted 30.56 kcal/mol, for an absolute difference of 2.3 kcal/mol, even if the actual percentage difference for this interaction is small.

The results are very encouraging, especially after taking into the fact that the number of atom types we used here is only 7, in comparison to 16 in AMBER[67] and 19 in CHARMM[13]. Clearly, results can be improved upon simply by introducing more atom types for special chemical groups.

3.5 Applications

3.5.1 Protein Structure Preparation

In protein design, the starting structure used to perform mutations is usually prepared from a crystal structure. The preparation process involves minimization of the crystal structure in any given force field to remove bad contacts. Here, we demonstrate the benefits of minimization using our updated DREIDING force field with a neutralized system. We make comparisons to performing minimization with all charged residues.

3.5.1.1 Methods and Material

A test set of high resolution crystal structures [38] (1bpi, 1isu, 1ptx, 1xnb, 2erl, 2hbf, 2ihl, 256b, 5rxn, 9rnt, all with resolution better than 1.7Å) is obtained from the PDB database. Water and solvents, if present, are removed from the crystal structures. WHATIF[48] is used to check for the protein health of PDB files and correct for mistakes when present. The assignment of DREIDING atom types are carried out according to the atom typing rules in Table 3-6. For minimization done with all residues neutralized, the procedure is carried out as described in Section 3.2.2. For minimization with charges, counter-ions are added by placing Na^+ or Cl^- next to exposed charged without salt bridging partners. The resulting system has a net charge of zero. The DREIDING force field parameters (including the modified VDW and hydrogen bond terms) reported in this presentation are used. CHARMM charges are assigned on the protein sidechains and backbone.

Conjugate gradient minimization of both neutralized proteins and charged proteins in the molecular mechanics code CMDF[68] to within 0.2 kcal/mol/Å force root mean square.

3.5.1.2 Comparison of RMSD

The purpose of minimizing a crystal complex is to merely remove bad clashes for proper calculations in a subsequent step. Thus, we do not wish to perturb the structure too much. Reported in Table 3-11 are the heavy atom coordinate root mean square deviation (RMSD) values for the minimized structures compared to crystal values over the test set of proteins after structural alignment. The RMSD difference for neutralized and charged protein minimizations are both within 0.5Å of the original crystal structure. The neutralized system has a statistically significant (to one standard deviation) smaller RMSD, indicating that the minimized structure is less perturbed than the one described using charged residues.

	Heavy Atom RMSD from Crystal Structure (Å)
Charged	0.493 ± 0.048
Neutralized	0.440 ± 0.033

Table 3-11 RMSD comparison of minimization of charged and neutralized proteins to original crystal coordinates.

We note that the salt bridges in the minimization using neutral residues are preserved because the missing Coulombic interaction has been compensated by our parameterized hydrogen bonds. On the other hand, we illustrate in Figure 3-9 one example in which the charged formulation creates an artificial salt bridge upon minimization. The protein is 1isu, where and an aspartate and a lysine in the charged protein minimization forms a salt bridge when there previously was none. In the crystal structure, the two closest atoms, $O_{\delta 1}$ on Asp56 and N_{η} on Lys17 were 5.97Å apart and were not forming a salt bridge. In the minimized charged structure, Asp56 rotated towards Lys17 and those two atoms minimized to 2.81Å of one another, forming a salt bridge. This artificial salt bridge formation is due entirely to the strong Coulombic interaction between these two residues in the charged formulation of our force field.

In the minimized neutral formulation, the two residues remained far apart as in the crystal structure. The fundamental change brought forth by minimization violates the premise that minimization was done merely to alleviate short contacts, i.e. "cleaning up" structures, and not to add properties that were not present in the crystal structure.

Combined with RMSD data, we conclude that structure preparation using the neutral formulation preserve crystal structure integrity better than using the charged

formulation. In particular, we have shown a danger in using charged system during minimization.

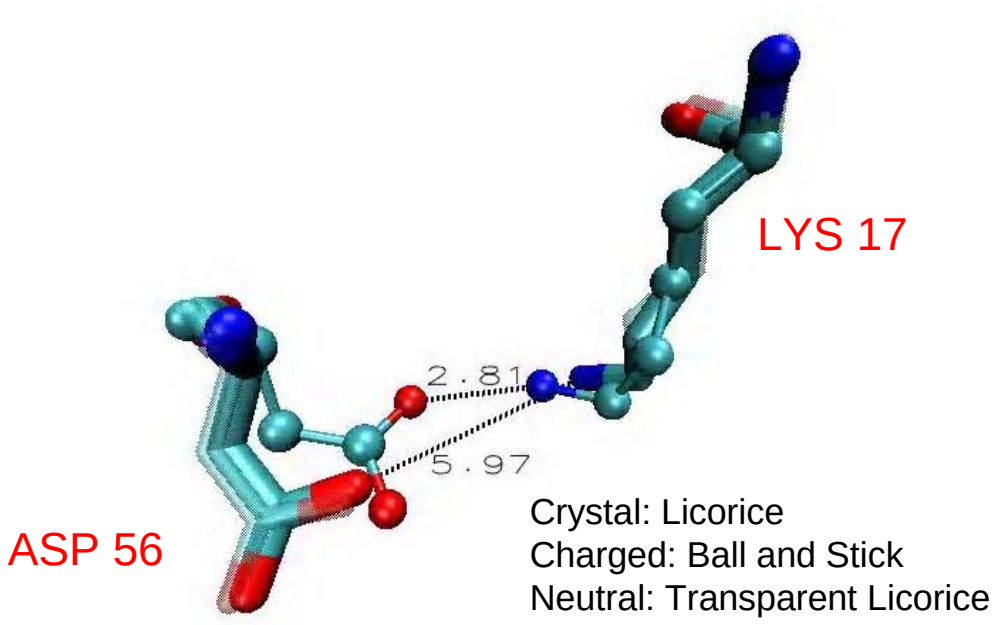


Figure 3-9 A salt bridge pair that was minimized into existence using charged residues. The protein shown is 1isu. Shown are the final minimized structure of the charged protein, neutralized protein, and the initial crystal structure of the residues Lysine 17 and Aspartate 56.

3.5.2 Effect on Molecular Dynamics: An Example

Next, we show that molecular dynamics simulation is stable after neutralization of all charged residues and that salt bridges are well-preserved. Such a simulation also has the property of being able to reach equilibrium in shorter simulation time.

3.5.2.1 Materials and Method

The crystal structure of protein human Interferon-beta (IFN-β) with pdb ascension code 1au1 is obtained from the PDB database. WHATIF[48] is used to check for protein health and correct for mistake if present. Atom typing, structure preparation and minimization are carried out as described in 3.5.1.1. Two structures were prepared: one with charged residues, the other with neutral residues

The molecular simulation code NAMD[69] is used to carry out the NPT simulation at 300K. Waters are added to both the neutral protein and the charged protein, counter-ions were added to only the charged system. Including hydrogens, the two systems both roughly have 20,000 atoms. 500 ps of simulation were carried out at 1 fs time step.

3.5.2.2 RMSD Comparison of Neutral and Charged System from Initial Structure

Without the presence of counter-ions, the neutral system should exhibit less fluctuation and take less time to equilibrate. Figure 3-10 shows the global RMSD values for the neutral system compared to initial structure vs. the charged system compared to initial structure. The neutral system took about 50 ps to equilibrate, whereas the charged system took about 200 ps. The equilibrated structure also has a smaller RMSD value from the initial structure for the neutral system, at about 1.3Å, than the charged system, at about 1.6Å. These observations agree with our assertion.

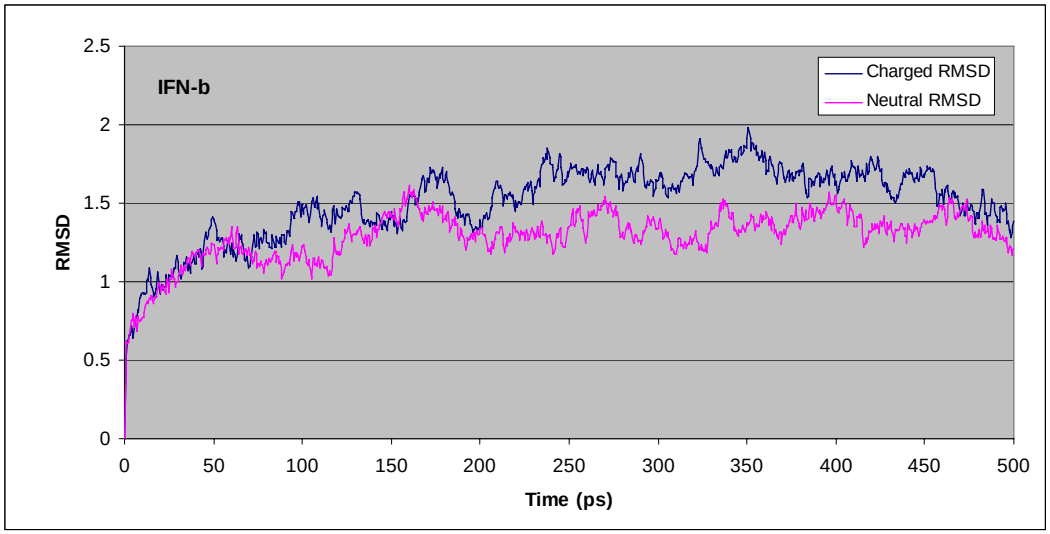


Figure 3-10 RMSD of protein heavy atoms for a charged and a neutral system, compared to initial simulation structure. 500 ps of simulation is shown.

We next examine the difference in dynamics of salt bridges in the neutral system compared to the charged system. Because of the difference in interaction energies between the charged description of a salt bridge and the neutral description, we expect major differences. Two salt bridges have been identified in the system: the Arg 27—Glu 29 pair (Figure 3-11) and the Arg 152—Glu 149 pair (Figure 3-13). The evolution of the distance between two atoms on the salt bridge is plotted in Figure 3-12 and Figure 3-14.

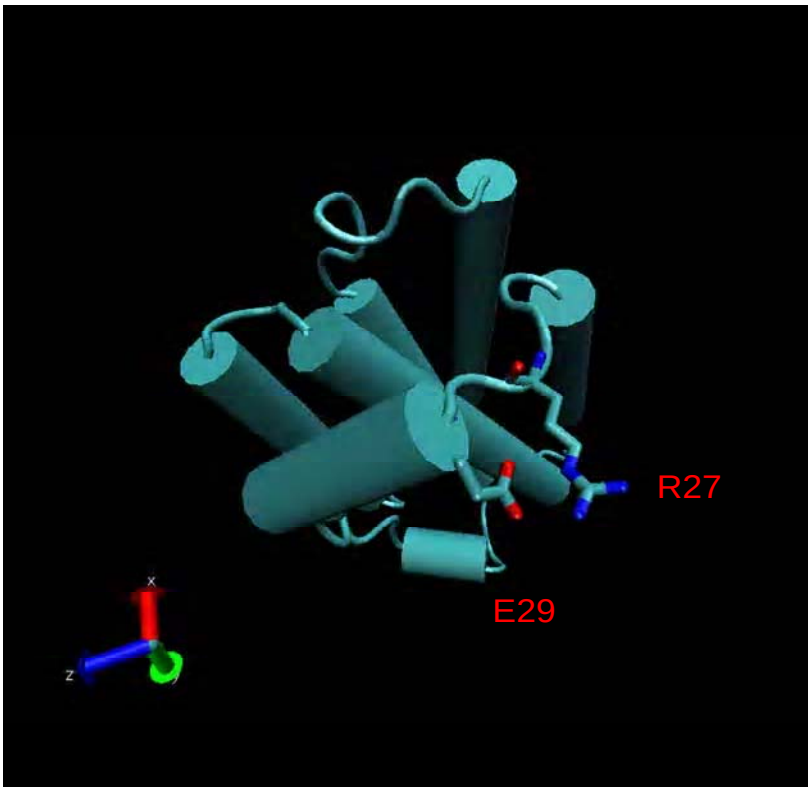


Figure 3-11 Residue R27 and E29 in IFN-β. The two residues are 3.7Å apart in the initial structure.

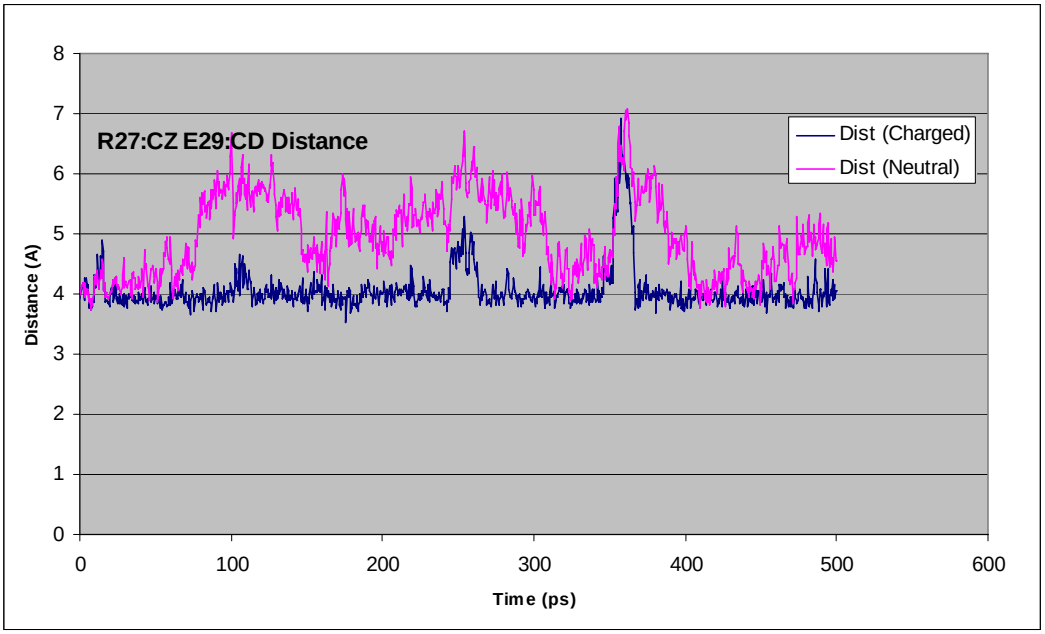


Figure 3-12 Evolution of distance between Cδ atom of E29 and Cζ atom on R27.

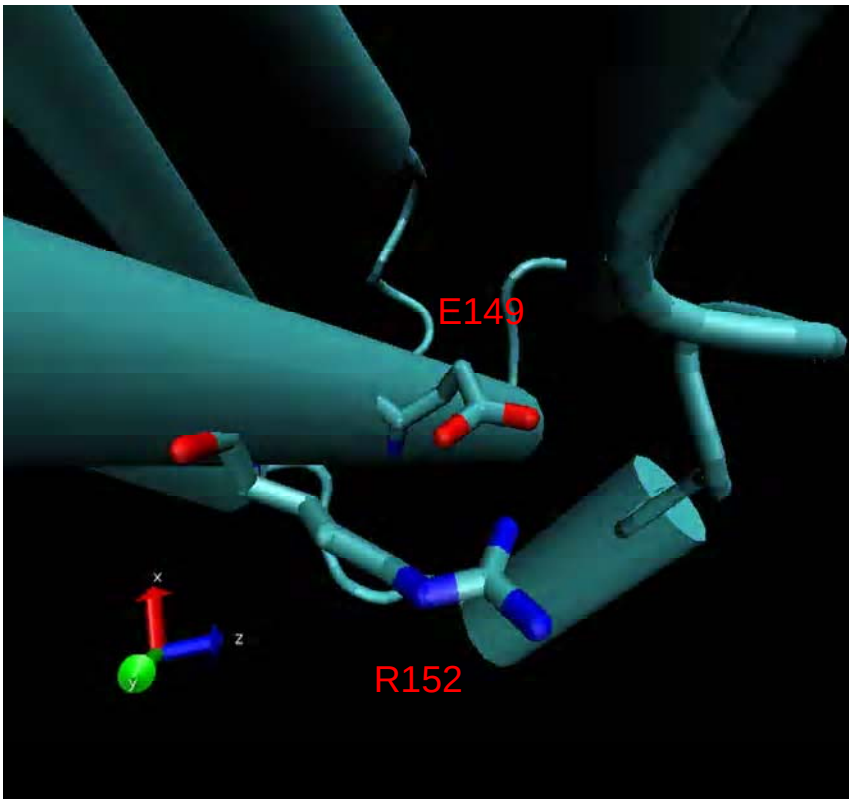


Figure 3-13 Residue R152 and E149 in IFN-β. The two residues are 4.3Å apart in the initial structure.

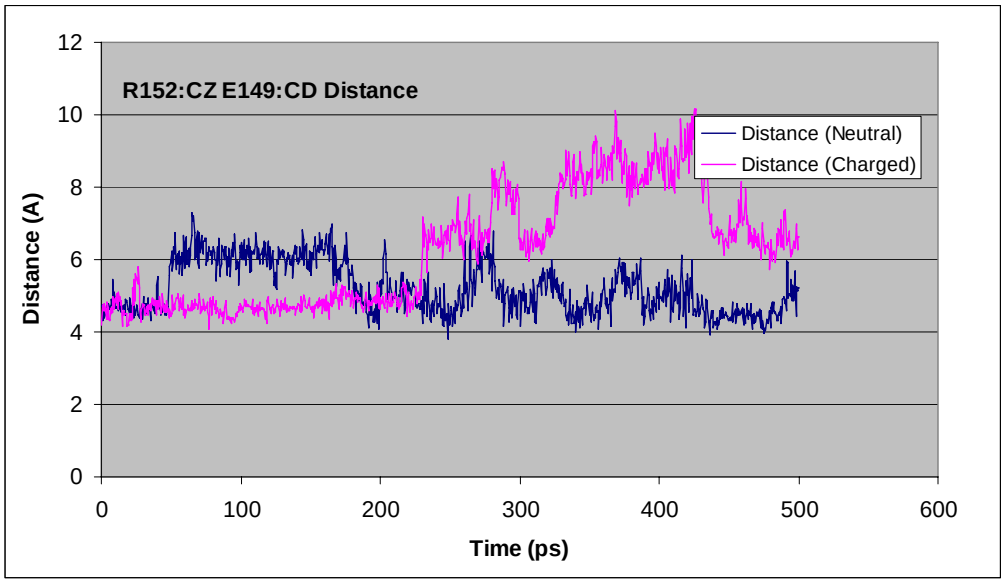


Figure 3-14 Evolution of distance between Cδ atom of E149 and Cζ atom on R152.

Both salt bridges experience bound and unbound states, with the unbound state typically caused by a mediating water molecule (visual inspection). In the R27-E29 case, the charged pair is essentially completely intact over the course of simulation, only becoming unbound essentially at around 370 ps and quickly returning to a bound state. The neutral pair experiences more fluctuation and is close to being in an unbound state on several occasions, but the salt bridge is never truly broken. In the R152-E149 salt bridge case, it is interesting to note that during the earlier portion of the simulation it is the neutral case that is bound, while the charged salt bridge is unbound. The situation reverses after about 200 ps, and the neutral salt bridge case is broken until it rematerializes closer to the end of the calculation. We note that surface exposed salt bridges are not stable species[63] and our simulation reflects this fact.

We are encouraged by the fact that the neutral protein maintains its stability and exhibits less fluctuation during simulation. In our examination of salt bridge formation and disruption we feel that using neutral system during MD can provide meaningful results.

3.5.3 Bovine Rhodopsin Helical bundle

3.5.3.1 Introduction

G-protein-coupled receptors (GPCRs) are membrane proteins that are involved in cell communication processes and mediate senses such as smell, taste and pain. Few 3D structures are present because of the difficulty of crystallization and therefore we have previously done work in the prediction of 3D structures of GPCRs by using MembStruk[21,70,71]. During the MembStruk procedure a difficulty is in the prediction of rotation angles of the 7 helices in a GPCR[21] (Figure 3-15).

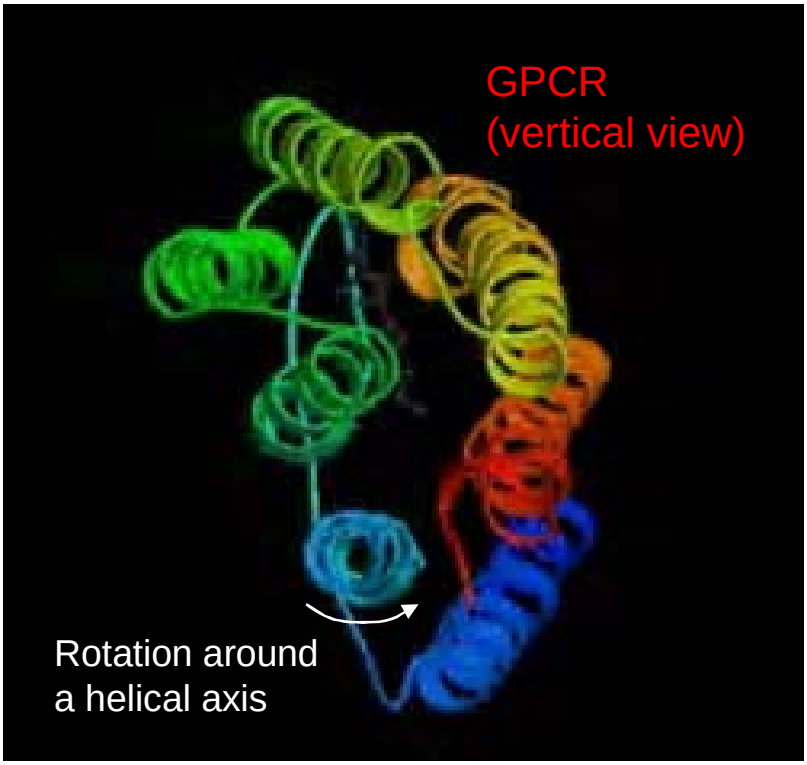


Figure 3-15 Top-down view of a GPCR with its 7 helixes. Each rotational axis of a helix is defined according to the MembStruk protocol.

A test for the neutralized protein and updated DREIDING force field would be to see whether it is capable of picking out the crystal structure from a set of decoys. Decoys are constructed by rotating each helix in the crystal structure individually at 5° increments.

3.5.3.2 Materials and Methods

The bovine rhodopsin structure (1u19, resolution 2.2Å) is taken from the PDB database and minimized using procedure described above. The centers and axis of rotation of the seven helices are defined according to the Membstruk[21] procedure. Inter-helical loops are removed.

A total of 7 x 72 = 504 structures are generated by 5° rotations of each individual helix, while keeping all other helices intact. Protein side chains are discarded and replaced by predictions from SCREAM with the rotamer library of 1.0Å diversity with an *s* value of 0.4, as described in Chapter 2. Both charged and neutral forms of systems are constructed based on the output from SCREAM and the resulting structures are then subject to 50 steps of conjugate gradient minimization to relieve short contacts. This final energy output is used to rank helix-angle-rotation structures.

3.5.3.3 Results and Discussion

We report results from using the neutralized system (Table 3-12) and using the charged system (Table 3-13). Only the top scoring structures are included. Both approaches achieved good results, with the desired selection in the top ranks. This is to be expected because large rotations away from the original structure would cause large VDW contacts, resulting in poor energies. We remark, however, that the system with no rotation scored the best among all 504 structures in the neutralized treatment, whereas it only ranked 5th in the charged system. We reason that electrostatics, because of its large variance, makes the prediction for the charged cases not as reliable as the neutral cases. For the top 100 results out of 504, the standard deviation in the electrostatic component of total energy is 7.9 kcal/mol for the neutral cases, 22.6 kcal/mol for the charged cases. Visual inspection of the electrostatic energies in Table 3-12 and Table 3-13 affirms the standard deviation statistic. A large standard deviation indicates a greater source of random errors, and this could be the cause of the zero rotation case in the charged formulation not having the best overall energy.

Rotation Structure	Overall Energy (kcal/mol)	Electrostatic (kcal/mol)
No Rotation	50.0	-161.9
H6_5	57.2	-156.8
H4_5	57.4	-167.7
H4_355	60.5	-158.0
H2_355	61.2	-154.8
H2_5	61.4	-148.6

Table 3-12 Top ranking structures using energies from neutralized system. The helix rotated is indicated in the second character, and the degree of rotation is indicated by the final 3 characters. Overall energy includes all valence terms and non-valence terms.

Rotation Structure	Overall Energy (kcal/mol)	Electrostatics (kcal/mol)
H1_5	144.5	-234.6
H5_355	145.3	-211.3
H6_355	148.2	-201.9
H2_355	149.4	-213.1
No Rotation	151.8	-213.4
H2_5	153.3	-200.9
H5_5	157.6	-234.3

Table 3-13 Top Ranking structures using energies from the charged system.

3.6 Conclusion

We have presented two related approaches to performing molecular mechanics simulation for proteins. First, we propose neutralizing all charged residues in order to obtain more reliable energy calculations, achieved through the reduction of electrostatic noise in calculations. Second, we modified the hydrogen bond in DREIDING and parameterized atom types commonly seen in proteins. Validation is carried out using our parameters on a test set of molecules involved in hydrogen bonding.

We believe that the neutralized approach along with the modified DREIDING force field gives us a solid first step in making predictions. Encouraging results were achieved from the test cases and applications we have performed. Further applications of our approach will be performed on other systems, including some protein design examples, as presented in Chapter 4.

4 Protein Design

In Chapter 2 and Chapter 3 we described two components of protein design: placement of sidechains and energy calculation. We present here examples of protein designs using those tools.

4.1 Introduction

The objective in computation protein design is to introduce new functionalities into currently existing protein by way of mutations. With the development of tools such as those we presented above, protein design is at a stage where computational predictions and results are emerging as a practical option in solving real-life problems[2,72].

We present two examples of protein design. The first case, design on the protein human interferon-β (IFN-β), involves mutation introduced at specific positions in a therapeutic protein, due to constraints in the manufacturing process. The second case involves trptophanyl-tRNA synthetase, an enzyme that we are interested in mutating residues of the active site so as to incorporate ligands with novel functional groups.

4.1.1 Methods and Material

Here, we outline procedures that are common to our examples presented in subsequent sections. Starting structures are crystal structures taken from PDB database. WHATIF[48] is used to check protein health of the structure and asparagines, glutamines and histidines are flipped when deemed appropriate by the program. Hydrogens are added by WHATIF, which optimizes the hydrogen bond network by using a genetic algorithm. CHARMM[13] charges are assigned to the atoms. When ligands are involved, Mulliken charges derived from quantum

mechanics at the 6-31G**/X3LYP level are used. Atom typing is performed using the procedure described in Section 3.4.1. Geometry optimization using conjugate gradient of the structure is then performed using MPSim[73] for 100 steps. Typically, the coordinate root mean square deviation (RMSD, in Å) between the optimized structure and the original crystal structure is less than 0.5Å.

4.1.2 The Design Protocol—Energy Excitation

Care and insight is necessary to identify pitfalls and opportunities in protein design. For instance, in active site redesign, a new ligand is introduced in place of an old one and it is necessary to generate a new geometry for the new ligand. Mutation of residues will proceed using the newly generated structure. Limiting mutations to certain type of amino acids (such as hydrophobic) would also be sensible in some cases. With that backdrop, we summarize our general protein design approach. We use a strategy called "*Energy Excitation*" to avoid combinatorial explosion of possible mutants, to be explained in detail as follows.

4.1.2.1 Identify Candidate Residues for Mutations

Identifying mutation candidates is an important step in protein design and is often driven by the problem at hand. In the IFN-β design case, methionines are to be mutated away for drug manufacturing reasons. In the TrpRS design case, the sites that are tagged for mutations are those that are in clashes with the newly introduced ligand. Mutations can be introduced to achieve a specific purpose, such as removing a potential hydrogen bond from forming. Visual inspection is also employed to ensure that residues that are selected to be mutation candidates are sensible.

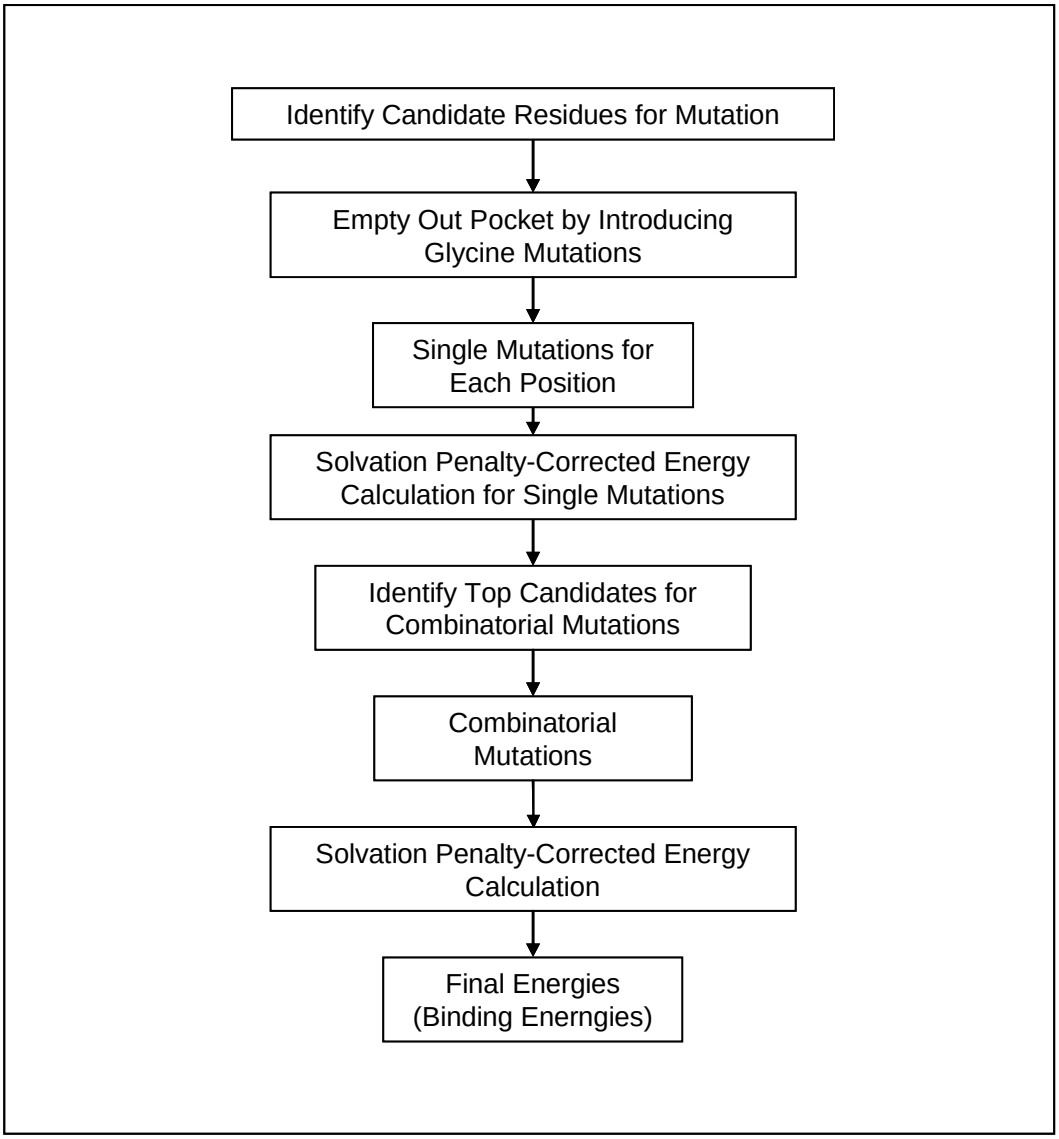


Figure 4-1 Flowchart for introducing mutations by using the "Singles Excitation" strategy.

4.1.2.2 Clearing Out Current Residues—Mutation to Glycine

After selecting the mutation sites, we remove the sidechains by introducing glycine mutations. This is done to completely eliminate any influence the previously existing sidechains may have on sidechains that are being mutated on.

In our design of TrpRS, ligands other than the one included in the crystal structure needs to be modeled into the binding site when performing mutations. In such cases, we perform a superposition of non-crystal ligand atoms with atoms on the crystal ligand. If the non-crystal ligand can take on multiple configurations (also termed rotamers), those configurations would need to be considered.

4.1.2.3 Single Mutations

SCREAM, as described in Chapter 2, is used for all sidechain placement purposes. "Single Mutation" refers to our method that mutations at each residue position are carried out individually at this stage. Since all other sidechains have been mutated into glycine at this stage, sidechains of each individual mutations only interact with the protein backbone, ligand, and other fixed sidechains. The list of potential mutations can include all 20 amino acids or a subset such as only hydrophobic residues.

4.1.2.4 Energy Calculation for Single Mutations

The overall energy of a mutation is calculated as below:

$$E_{Mutation} = E_{raw} - E_{InternalReferencee} + E_{SolvationPenalty}$$

Where $E_{Mutation}$ is the overall stability due to the mutation, E_{raw} is the force field interaction energy of the mutated sidechain with the rest of the protein after optimization, $E_{IntenalReference}$ is the reference energy for the internal energy terms, and $E_{SolvationPenalty}$ is the solvation penalty for this sidechain. These terms are explained in the following sections.

4.1.2.4.1 Structure Optimization for Single Mutations

Minimization of the sidechains of mutations is performed to relieve clashes after charged residues are neutralized as described in Chapter 3 to obtain more accurate electrostatics. Only atoms belonging to this sidechain are moveable during this minimization procedure. The final energy of atoms on the sidechain (from C_β on

out) and the rest of the protein is calculated from the optimized structure. This energy is the raw energy, E_{raw}, which includes the interaction energy between the sidechain and the rest of protein, and the internal energy of the sidechain.

4.1.2.4.2 *Internal Energies of Sidechains*

A reference point is needed in order to make the raw force field energy comparisons meaningful since different amino acid sidechains have different number of atoms. Internal energies of each sidechain include the valence terms (bond, angle, torsion, inversion) and non-valence terms (1-4 VDW and 1-4 electrostatic terms), and needs to be excluded by subtracting the reference internal energy for each of the 20 amino acids. This reference internal energy is pre-compiled as follows. The lowest energy rotamer of each amino acid side chain is placed in an extended Ala-X-Ala tri-peptide. This structure is then minimized (in vacuum) to 0.1 force RMSD. The force field energy of this structure involving just the amino acid sidechain atoms (starting from the C_β atom) is then calculated, and taken as the reference energy for this amino acid (Table 4-1).

Amino Acid	Reference Internal Energy (kcal/mol)	Solvation Energy[74] (kcal/mol)	Surface Area (Å^2)
Alanine	3.33212	1.94	74.4
Arginine	-92.2519	-19.92	253.94
Asparagine	-24.3287	-9.68	142.12
Aspartate	-23.4140	-10.95	123.41
Cysteine	5.2260	-1.24	108.41
Glutamine	-13.3623	-9.38	178.16
Glutamate	-10.0202	-10.20	160.09
Glycine	0.0000	2.39	n/a
Histidine	15.1799	-10.27	163.48

Isoleucine	12.6791	2.15	172.97
Leucine	3.4640	2.28	171.03
Lysine	6.2283	-9.52	186.43
Methionine	5.0126	-1.48	183.81
Phenylalanine	11.3280	-0.76	184.17
Serine	5.7298	-5.06	96.66
Threonine	-0.9482	-4.88	123.05
Trptophan	30.6985	-5.88	230.83
Tyrosine	6.6055	-6.11	208.74
Valine	7.6409	1.99	130.80

Table 4-1 Internal energies, experimental solvation energies and surface area for each of the 20 amino acids.

4.1.2.4.3 Solvation Penalty

Solvation is an important consideration in determining whether a mutation is favorable in its environment. Hydrophilic residues such as serine would be penalized for being buried inside a protein due to loss of solvation in water. Therefore, a good hydrogen bond network would be essential for hydrophilic residues to form with other sidechains and the protein backbone. A solvation penalty is assigned to buried sidechains according to experimental solvation energies[74] for the amino acids are listed in Table 4-1.

We take into account the degree of exposure of a sidechain. Solvation penalty for this mutation is scaled according to the following equation:

$$E_{SolvationPenalty} = \frac{ExposedSurfaceArea}{FullyExposedSurfaceArea} E_{SolvationEnergy}$$

Solvent accessible surface area is calculated by rolling a 1.4Å ball on the surface of a protein, with the radius of the atoms as defined as DREIDING. The fully exposed surface area is defined as the surface area of an amino acid sidechain in a fully extended tri-peptide Ala-X-Ala. The solvation penalty is then scaled by the ratio between the exposed surface area and the fully exposed surface area.

4.1.2.5 Determine Top Candidates for Combinatorial Mutations

Each mutation at a certain residue position has an energy calculated according to the previous section. To avoid combinatorial explosion, multiple mutations (Energy Excitation) are performed the following way:

1. Add up the individual energies for each mutation across all residues that had mutation introduced:

$$E_{total}(AA_1, AA_2, ..., AA_k) = \sum_i^k E_i(AA)$$

where *i* stands for the residue that was mutated, *AA* an amino acid mutation, $E_i(AA)$ the energy for mutation *AA* at position i obtained from section 4.1.2.4.

2. Sort $E_{total}(AA_1,AA_2,\ldots,AA_k)$ in ascending order. Perform these mutations.

When the total number of combinatorial possibilities is less than 200, we perform all possible mutations. Otherwise the total number of mutations we perform are capped at 200.

4.1.2.6 Multiple Mutations

Since the orientation of a side chain can be influenced by the placement of another, mutation sidechains are placed simultaneously at this stage. SCREAM as discussed in Chapter 1 is used for this purpose.

4.1.2.7 Energy Calculation for Multiple Mutations

Minimization is performed for all residues with atoms within a 5.0Å radius of atoms that belong to sidechains that have been introduced through mutations. Backbone atoms of these residues and ligand atoms are also minimized at this stage.

The correction for solvation penalty and internal energies are carried out in the manner as in Section 4.1.2.4.2

4.1.2.8 Binding Energies

The binding energy of a ligand bound to protein is calculated as:

$$\Delta\Delta E_{binding} = \Delta E_{complex} - \Delta E_{protein} - \Delta E_{ligand}$$

where $\Delta E_{complex}$ is the energy of the protein-ligand complex, $\Delta E_{protein}$ the free energy of the protein without the ligand, and ΔE_{ligand} the free energy of the ligand by itself (vacuum). Implicit solvation energy is included in the above calculations using the Poisson-Boltzmann solver DelPhi[75].

The above procedure selects for maximal stability from a set of mutations. Specificity is important in some cases and the differential binding energy is calculated as follows:

$$\Delta\Delta E_{differential} = \Delta\Delta E_{binding,1} - \Delta\Delta E_{binding,2}$$

Where $\Delta\Delta E_{binding,1}$ and $\Delta\Delta E_{binding,2}$ are binding energies for ligand 1 and ligand 2 in the same protein environment.

4.2 Design of Human Inteferon-β

4.2.1 Introduction

Interferons (IFNs) are proteins produced by the body's immune system as defense against foreign agents such as viruses, parasites and tumor cells. In particular, because of its effect in slowing progression of disability in multiple sclerosis, it has attracted attention in pharmaceutical companies. IFN-β based drugs already on the market include Avonex, Rebif and CinnoVex.

We are interested in mutating the methionines away in human IFN-β, because of a constraint in a manufacturing process. Our objective is to introduce mutations at Methionine positions while maintain the stability of the original protein. In human IFN-β, there are three methionines, Met 36, Met 62, and Met 117 (Figure 4-2), not including the methionine residue at the amino terminus of the protein. Since hese residues are far apart, we will treat them as independent mutations and carry out calculations as such. The crystal structure with pdb ascension code 1au1 is used for structure preparation as described in 4.1.1 after solvents and waters are removed.

4.2.2 Single Mutations

4.2.2.1 Met 36 and Met 117

We first perform single mutation of Met36 and Met117. These are both surface exposed positions, and therefore, we consider single mutations only for these two residues. The results of our calculations are shown in Table 4-2 and Table 4-3. Energies are shown using the wildtype methionine as a reference energy.

Mutation at Position 36	Energy (kcal/mol)
M	0.0
T	-2.96
V	0.102
Y	0.338
I	0.504
F	1.165

Table 4-2 Energies for top mutations introduced at position 36 of human IFN-β. Energy values are relative to the wildtype Methionine energy.

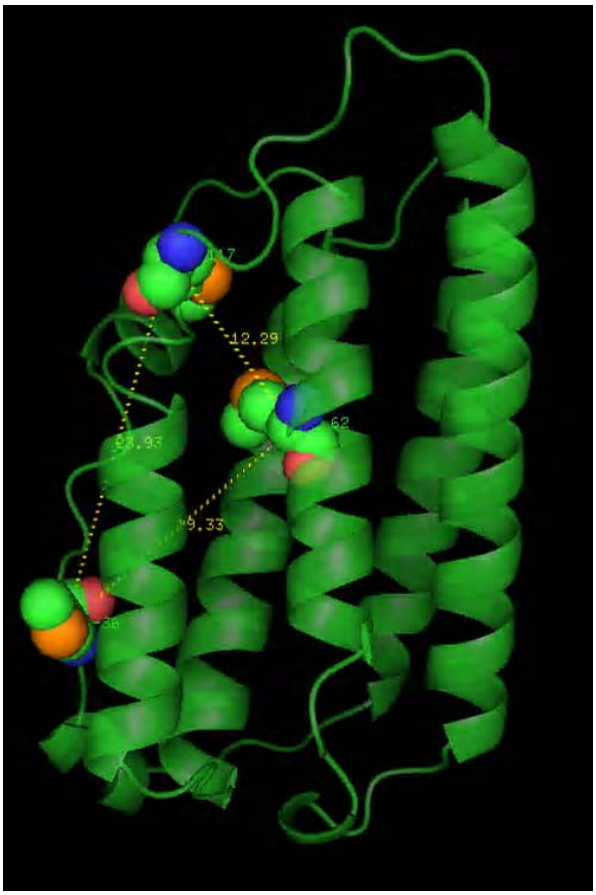


Figure 4-2 Human IFN-β (pdb code: 1au1). Methionines are shown in ball and stick format. Met 36 is the one on the bottom left hand corner, Met 117 top left hand corner, Met 62 the one positioned in the middle. Distances between the methionines are indicated on the picture.

Mutation at Position 117	Energy (kcal/mol)
M	0.0
L	-2.57
T	-1.55
S	-0.93
G	0.65
I	1.20

Table 4-3 Energies for top mutations introduced at position 117 of human IFN-β. Energy values are relative to the wildtype Methionine energy.

These energy differences were small compared to the wildtype Met, and therefore we do not expect that the structure would become unstable because of them. Our experimental collaborators did mutations on some of the amino acids. At position 36, Thr, Ile and Ala mutations experiments were done, and all mutations proved fully active. At position 117, Thr, Tyr, Ser and Gly mutations weredone. Again, all mutations led to active IFN-β. In fact, Thr and Tyr mutations led to mutants that were actually more active than the wildtype IFN-β. Our predictions are essentially consistent with the experimental results, even if it is just one mutation on the surface of the protein.

4.2.2.2 Met 62

Met62 is completely buried. We first performed single point mutations at this position, with the top mutation candidates shown in Table 4-4. The best mutation candidate is Thr for this position, but it is still 6.8 kcal/mol less favorable than the wildtype methionine. This is a large difference in stability to overcome and indeed, experimentally, no mutants at this position led to a stable protein.

Mutations at Position 62	Reference Energy (kcal/mol)
M	0.0
T	6.8
L	8.6
A	9.9

Table 4-4 Top mutation candidates introduced at position 62 of human IFN-β.

Therefore, we turn our attention to multiple mutations around the Met 62 neighborhood.

4.2.3 Combinatorial Mutations of Residues around Met 62

Since we were unable to find a mutation that worked well for Met 62, we proceeded to try out mutations around the Met 62 neighborhood to improve the packing. We observed (Figure 4-3) that Ile40 and Ile44 are also primary residues in determining the stability of the Met62 neighborhood. Therefore, we proceeded to mutate these residues in this neighborhood. Seeing that the residues are deeply buried in the protein and are hydrophobic, we picked only hydrophobic mutations for residue positions 40, 44 and 62. Results are shown in Table 4-5.

Energy (kcal/mol)	62	40	44
0.0	M	I	I
0.103	I	F	L
5.624	F	I	V
6.972	T	L	I
11.944	I	I	I
13.296	L	I	L
17.125	L	L	I

Table 4-5 Energies for various mutations in the hydrophobic pocket around Met 62. The wildtype energy is used as reference for other mutations.

The stability of the triple mutant I62, F40, L44 was very close to the wild type sequence according to our calculations, being only 0.1 kcal/mol apart. It is evident from Figure 4-3 that the three mutations lead to a very tight packing. Experimentally, this triple mutant proved to be a fully active IFN-β.

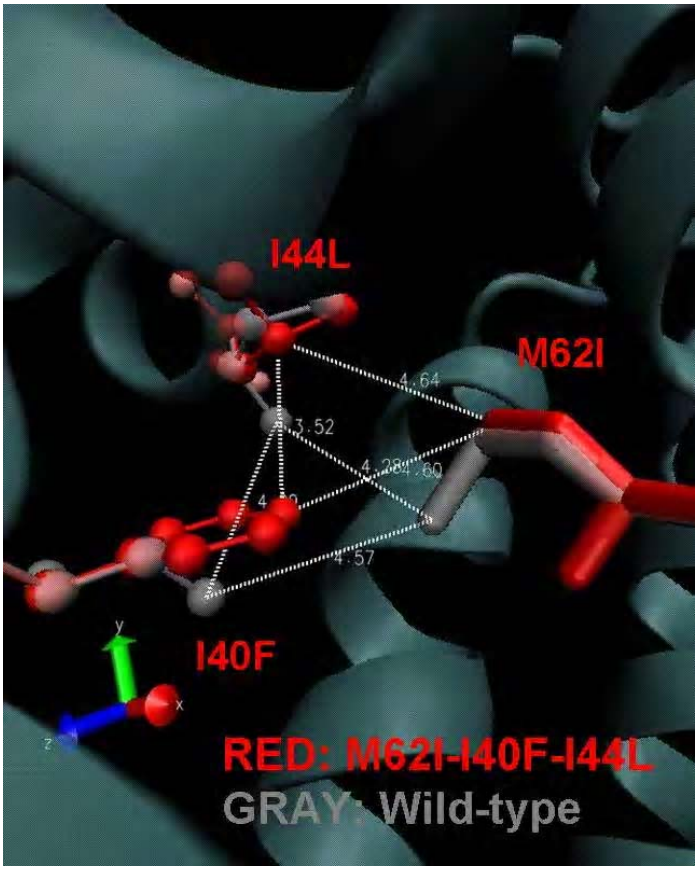


Figure 4-3 Residues near position 62 in human IFN-β. Ile40 and Ile44 are identified as residues that are in direct contact with Met62 and needs to be mutated when changes to Met62 is made.

4.3 Active Site Design of TrpRS

4.3.1 Introduction

4.3.1.1 Amino-acyl tRNA Synthetases

Naturally occurring proteins are made up of twenty common amino acids. To expand the functionalities and properties of proteins, bioengineers have incorporated non-natural amino acids into proteins[76-78]. One strategy for such incorporation is accomplished by tweaking the protein translational machinery—specifically, by introducing mutations to a class of proteins called amino-acyl tRNA synthetases (aaRS).

aaRSs are enzymes responsible for the transfer of amino acids onto transfer RNAs (tRNA) during the protein translation step[79,80]. In the translation process (Figure 4-4), the identity of each amino acid is determined by a codon in the messenger RNA (mRNA). Each codon is recognized by the anti-codon on a specific tRNA for that amino acid. Each tRNA, in turn, needs to have the accurate amino acid charged onto it by way of the corresponding aaRS. As a result, there are 20 aaRSs in the vast majority of organisms, one for each amino acid. Mistakes in the translation process could involve either the codon-anticodon interaction or the charging of amino acids on tRNAs by the aaRSs.

Mistakes can be turned into opportunities for protein engineers. Since each codon is represented by 3 nucleotides, there are a total of 4*4*4 = 64 possible distinct codons. There are only 20 natural occurring amino acids plus the start and stop codons, so there is redundancy in the genetic code. One common strategy to take advantage of this situation would be to introduce a tRNA with the proper anti-codon to interact with a redundant codon. An amber (UAG) stop codon is often used as the target for this purpose. This extra tRNA can be amino-acylated with a desired non-natural amino acid analogue, which would then be incorporated in the protein via the cell's protein manufacturing machinery. Indeed, this strategy has been used by protein engineers as opportunities for exploring novel chemistry[81,82].

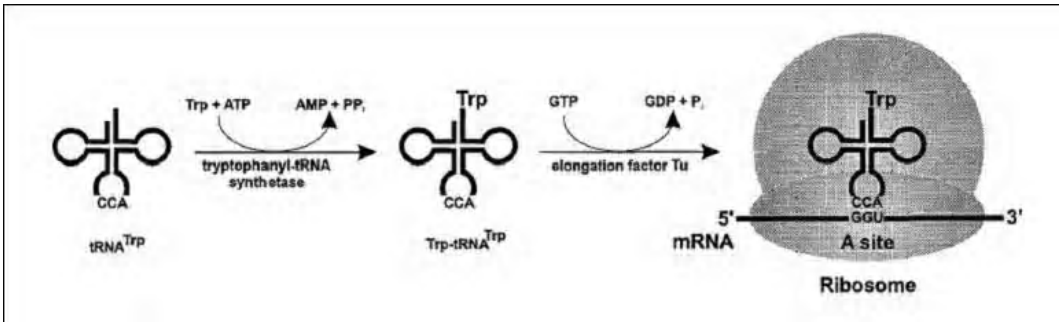


Figure 4-4 Cartoon depicting the role played by an amino-acyl tRNA synthetase. The amino acid tryptophan is being charged in this example. (Adapted from Ibba. et al[79]).

In this presentation, we focus on the role played by aaRSs. We mentioned that we could incorporate an amino-acid analogue via the amino-acylation of a tRNA that has an anti-codon paired with the UAG stop codon. There remains the problem of modifying the aaRS so that it would actually function as desired.

There is intense evolutionary pressure on aaRSs to achieve high fidelity in the amino-acylation of a tRNA to its proper amino acid. This is because charging an incorrect amino acid onto a tRNA would result in an incorrect protein sequence, which would very likely lead to inactive proteins and wasted resources. Thus, aaRSs are very efficient in its recognition of the cognate amino acid. Indeed, the error rate is approximately 1 in 3,000[79]. By introducing mutations in the active site and disrupting the recognition of the cognate amino acid, we can induce the enzyme to charge non-natural amino acid analogues[3,83,84] onto a tRNA.

PheRS (phenylalanine tRNA synthetase) and TyrRS (tyrosyl tRNA synthetase) have been used to incorporate unnatural amino acid analogues[81]. However, many interesting amino acid analogues are a bigger size than what the binding pockets of PheRS and TryRS can accommodate. Trptophan is the largest of the 20 common amino acids, and its corresponding tryptophanyl-tRNA synthetase (TrpRS), as a

result, has the potential to accommodate the biggest ligands. As a result, we carry out *in silico* redesigns of *Bacillus Stearothermophilus* TrpRS and human TrpRS to recognize non-natural amino acid analogues.

4.3.1.2 Click Chemistry

Click chemistry[85,86] refers to a type of reactions with simple reaction conditions, high yield, and modular building blocks. Of specific interest is the azide-alkyne Huisgen cycloaddition[87], in which an azide reacts with an alkyne to form a triazole. This reaction is highly exothermic. We performed quantum mechanics at the 6-31G**/X3LYP level and found that the reaction is exothermic by 71 kcal/mol when R1=trp ring and R2 = methyl. The activation barrier of this reaction is 13 kcal/mol in the presence of a copper (Cu^{2+}) catalyst.

We are interested in incorporating an amino acid analogue that contains an alkyne or an azido functional group for the purpose of performing click chemistry involving proteins. For example, by using an amino acid analogue with an ethynyl group in place of a natural protein, we can covalently link the protein to another molecule which has an azido group by performing click chemistry. This molecule can be a florescent tag or a protective chain to facilitate drug delivery.

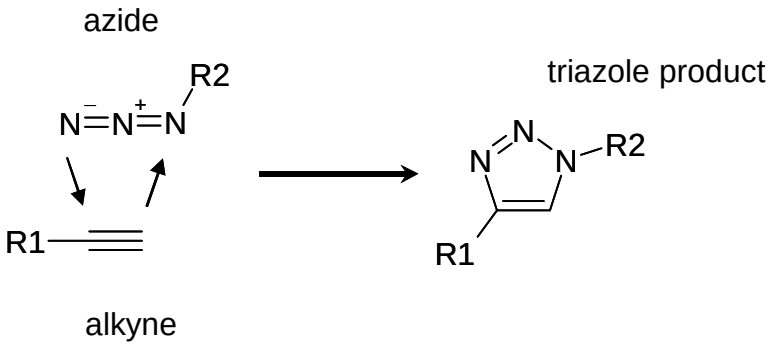


Figure 4-5 Illustration of an uncatalzyed azide-alkyne 1,3-cycloaddition.

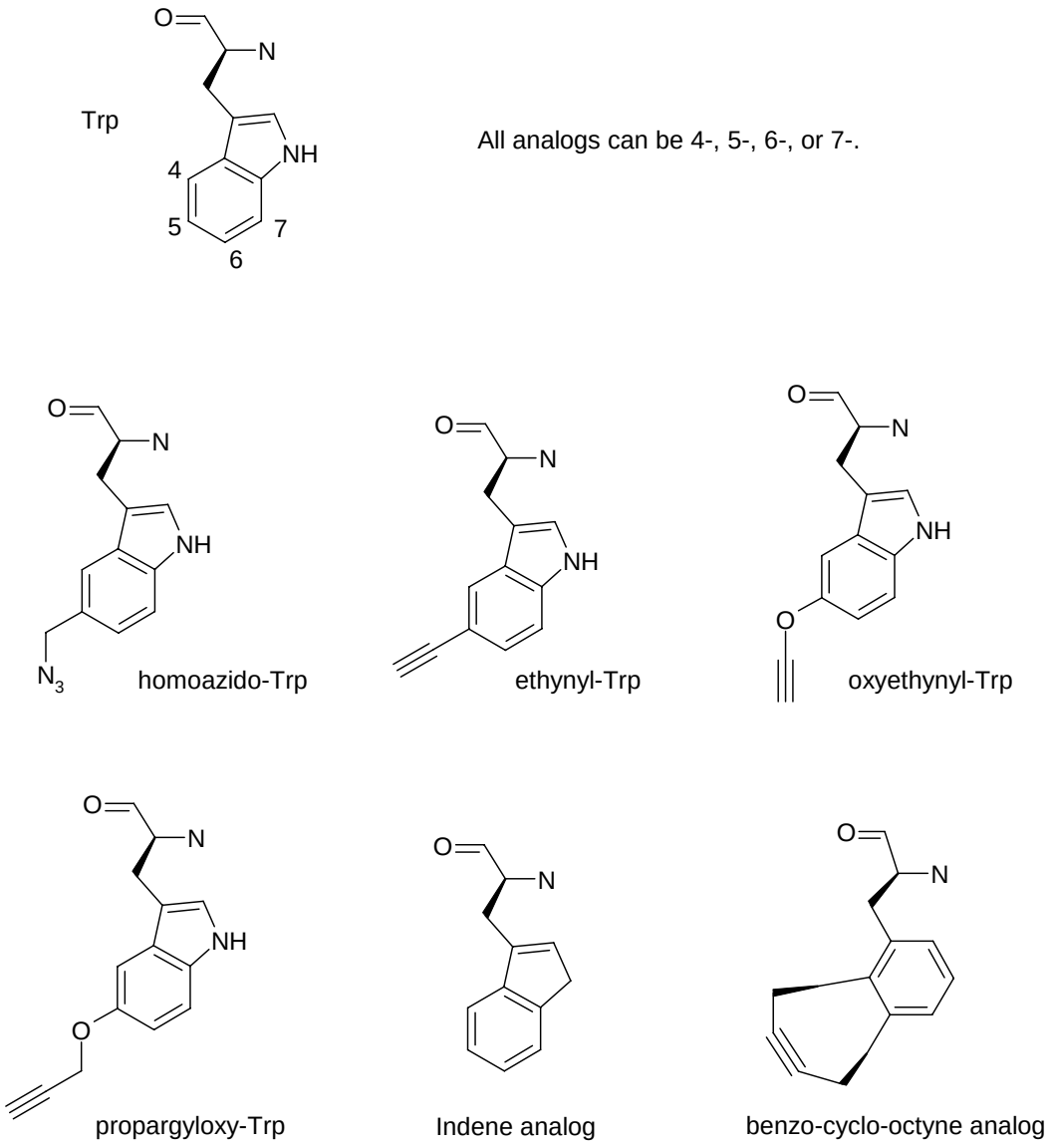


Figure 4-6 Non-natural amino acid analogs of Trptophan with functional groups that are relevant to click chemistry.

We show a list of non-natural amino acid analogues that we are interested in incorporating in Figure 4-6. With the exception of the Indene amino acid analogue, all these molecule can participate in click chemistry. The indene analog is chosen as a model compound for design purposes, to be explained below.

4.3.1.3 Preparation of Non-natural Amino Acid Structures

We have no preferences regarding the position of the ethynyl or azido functional groups with respect to the trp-ring. All non-natural amino acids are built and then optimized at the X3LYP/6-31G** level. Mulliken charges are also assigned at the X3LYP/6-31G** level. An atom-for-atom alignment with the cognate Trp ligand is used for the placement of non-natural analogue into a TrpRS active site.

4.3.2 *Bacillus Stearothermophilus* TrpRS Active Site Design

4.3.2.1 Introduction

B. Stearothermophilus TrpRS crystal, 1mb2[88], is obtained from the PDB database. Out of the two chains that are involved in the crystal packing, we keep only chain A from the crystal structure. Structure preparation of the structure is then done as described in Section 4.1.1. The ligand in the structure is the cognate tryptophan amino acid. The active site of the optimized structure is shown in Figure 4-7.

The binding pocket of TrpRS contains mainly hydrophobic residues. Phe5, Ile133, Val141 and Val143 line up the active site that recognizes the tryptophan ring, with all of those residues at about 4Å away from the cognate ligand. Clearly, mutations of these residues will be crucial if non-natural tryptophan analogs as those in Figure 4-6 were to be incorporated by TrpRS. In addition, the HN group on tryptophan forms a hydrogen bond with Asp132, a critical residue in the activity of this TrpRS as shown by mutagenesis studies[88].

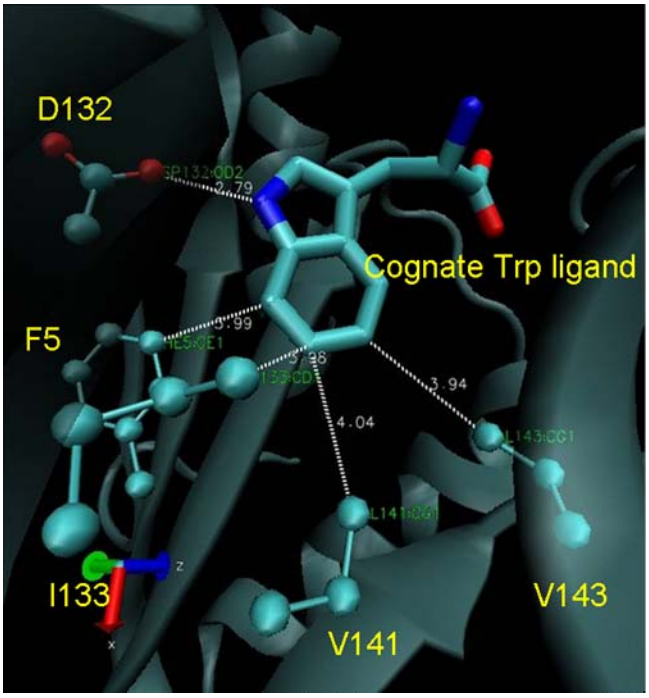


Figure 4-7 The active site of B. Stearothermophilus TrpRS with cognate Trp ligand. The ligand is shown in licorice representation, whereas the residues interacting with the ligand is shown in ball-and-stick format.

4.3.2.2 Design for Ethynyl-Trp

Ethynyl-Trp is shown in Figure 4-6. To determine which of the 4-, 5-, 6- and 7-position of Ethynyl-Trp would best be accommodated in the active site, we built all four ligands and placed the ligands into the crystal structure. The ethynyl group of 4-, 6- and 7- ethynyl-Trp would clash with the protein backbone and are therefore eliminated from further consideration. As a result, 5-ethynyl-Trp is the only configuration used in our design effort.

V143 is the position we identified as a target for mutation. We follow the protein design protocol as described in Section 4.1.2. Since the ligand is highly hydrophobic, we allow only hydrophobic mutations at position 143. The results are shown in Table 4-6. Only an alanine or glycine mutation worked favorably in the presence of 5-Ethynyl-Trp.

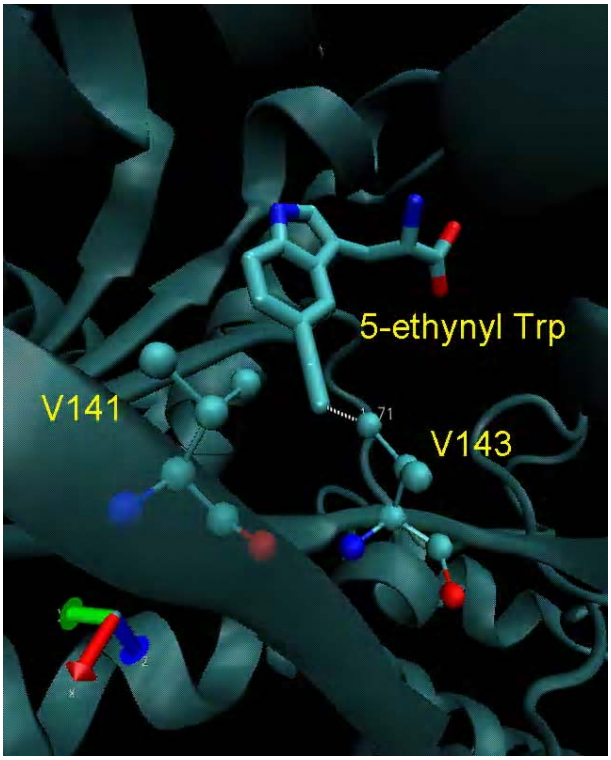


Figure 4-8 The active site of *B. Stearothermophilus* TrpRS with 5-ethynyl-Trp ligand. The ligand is shown in licorice representation. V141 and V143, two residues potentially interacting with the ligand, are shown in ball-and-stick format.

Mutation at Position 143	Energy (kcal/mol)
G	0.00
A	3.61
M	8.45
L	9.43
V	17.93
I	19.95
F	94.21

Table 4-6 Interaction energies for single mutations at position 143 in the presence of 5-Ethynyl Trp in place of cognate Trp ligand. The energy of the best mutation is used as a reference energy.

Any mutation that introduces an amino acid larger than Alanine would not fit in the binding pocket. Subsequently, we calculate the differential binding energy of 5-Ethynyl-Trp and cognate ligand for the Alanine and Glycine mutations at position 143:

Interaction Energy	V143A	V143G
Cognate Trp	-42.4 kcal/mol	-42.3 kcal/mol
5-Ethynyl-Trp	-47.2 kcal/mol	-45.7 kcal/mol
Differential Binding	4.8 kcal/mol	3.4 kcal/mol

Table 4-7 Binding energies and differential binding of V143A and V143G mutations of 5-Ethynyl-Trp compared to the cognate Trp ligand.

Differential binding energies of 4.8 kcal/mol for the V143A mutation and 3.4 kcal/mol for the V143G mutation were calculated. Thus, we predict that the mutations should be sufficient to distinguish cognate Trp from the non-natural 5-Ethynyl-Trp amino acid. Experimental results (private correspondence) for these two mutations show that 5-Ethynyl-Trp is indeed incorporated by introducing those mutations, but not incorporated when no mutations are included. However, the cognate ligand was also incorporated for those mutations, at a level equivalent to the target ligand.

While our predictions on the incorporation was correct, the differential binding energy was not sufficient for the target ligand to be selected over the cognate ligand. In the subsequent section, we describe introducing further mutations to improve the differential binding of non-natural amino acid over cognate ligand.

4.3.2.3 Mutating Away Asp132 to Reduce Cognate Trptophan Recognition

The *B. Stearothermophilus* TrpRS D132 residue recognizes the Trptophan via the amide group at the 3- position of the Trp ring. This residue is conserved across all Bacteria and Archaea organisms, and its importance in recognizing Trp can be seen

in Figure 4-7. It is clear that if we can mutate away D132, the recognition of the cognate Trp would decrease by the strength of a hydrogen bond. The amide is not a necessity for our purposes since we desire only the functionality of an ethynyl group. Therefore, in order to achieve higher differential binding between the cognate ligand Trp and other non-natural amino acid analogs, we pursue the strategy of mutating away D132.

The model amino acid analog with indene group (having a CH_2 instead of HN at the 3- position, see Figure 4-9) is used as the model ligand, which is isosteric to trptophan. Functional groups such as ethynyl or azido can be branched from indene at the 4-, 5-, 6- or 7- position in the same manner as trptophan analogs. We treat the Asp132 mutation independent of mutations among residues that line the hydrophobic pocket.

Our first effort is to mutate D132 alone, following the procedures outlined in 4.1.2. The top scoring candidates are presented in Table 4-8.

Mutation at Position 132	Energy (kcal/mol)
N	-13.25
S	1.88
C	2.14
A	8.20
G	10.86
T	12.36
V	17.82

Table 4-8 Mutation at position 132 of *B. Stearothermophilus* TrpRS. The indole ring of the cognate Trp ligand has been replaced by an indene ring.

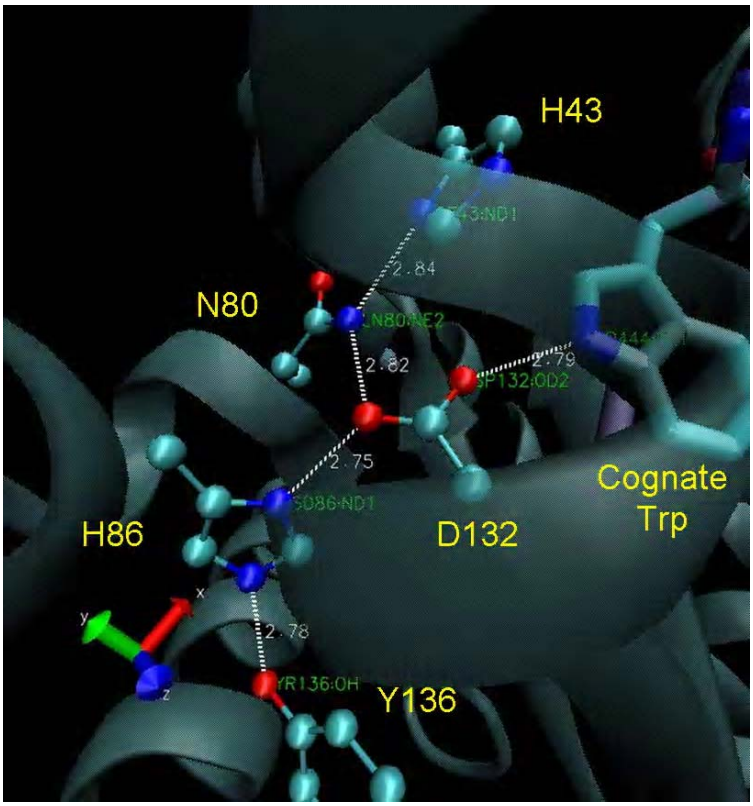


Figure 4-9 Residues around D132 in *B. Stearothermophilus* TrpRS. One of the oxygens of D132 serves as the hydrogen bond acceptor of the HN group of Trp.

There is extensive hydrogen bonding between D132 and its neighboring residues. There is experimental evidence that it is rare for hydrogen bond donors not to be satisfied in protein structures, and thus it would be prudent to also consider mutations only to hydrophobic residues. This means we need to introduce more than just one mutation so that we can create a hydrophobic pocket in place of the hydrogen bond network. Thus, we decide to include a neighboring hydrogen bond partner, N80, in our mutation search. H43 and H86, while also forming part of the hydrogen bond network with D132, are kept as is because the histidines can act as both donor and acceptors depending on the δ- or ε- states, lending flexibility to the hydrogen network that they make.

The results for double mutations at position D132 and N80 are presented in Table 4-9. A Methionine mutation at position 80 fits snugly in the pocket, as it occupies the space left empty by the missing atoms from the Aspartate on position 132, which have been mutated to small residues, Alanine and Glycine.

Position 132	Position 80	Interaction Energy (kcal/mol)
G	M	-9.538
A	M	-8.596
M	G	-5.641
G	G	-4.780
A	G	-4.005
G	A	-3.014
M	A	-2.668
A	A	-2.257
G	L	2.478

Table 4-9 Double mutations at position 132 and 80 of *B. Stearothermophilus* TrpRS.

The mutation of a charged amino acid in an active site to a non-charged amino acid brings about a concern regarding whether the mutated protein will fold. Whether a protein folds after mutations have been introduced is not something calculations can reliably predict at this stage.

4.3.3 Human TrpRS Active Site Design

4.3.3.1 Introduction

In preparing for the possibility that the protein may not fold, we performed sequence alignment of various TrpRS to determine if there are any species whose TrpRS sequence does not have an Aspartic acid at the HN recognition site. We

identified the human TrpRS as such a candidate, with an asparagine (N194) in place of the aspartate (D132) at the equivalent position (Figure 4-10). This structure has also been crystallized, with pdb code 2dr2. Our hope is that, experimentally, introducing mutations at this residue would not lead to inactive proteins.

Our strategy is to reduce the complexity of the design problem by proceeding along two tracks. One task is to introduce mutations around N194 to remove recognition of the HN group on the cognate tryptophan ring. The other task would be to introduce mutations for residues that line the hydrophobic pocket. Figure 4-10 shows that the two groups of residues are independent of each other. Try159, Gln284 recognize the HN group of the cognate Trp ligand, whereas Ile307 and Cys309 are residues that independently recognize the Trp ring.

Again, the position (4-, 5-, 6- or 7-) of the functional group with regard to the Trp ring is of no significant consequence. The ligand structures are prepared and charges on the ligands are assigned according to Mulliken charges at X3LYP/6-31G**.

Crystal structure of human TrpRS (pdb ascension code 2dr2) is used for structure preparation, following the steps outlined in Section 4.1.1.

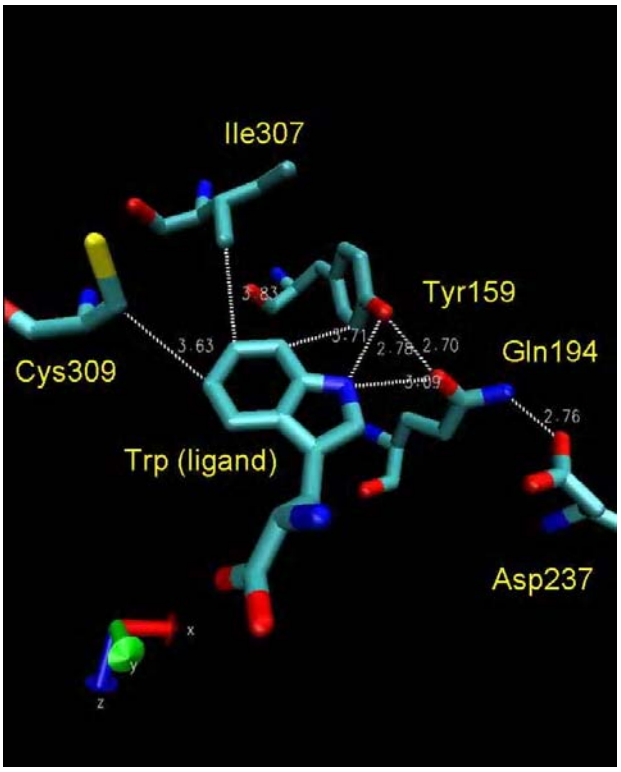


Figure 4-10 Binding site of human TrpRS (pdb code 2dr2). Cys309 and Ile307 can be mutated to accommodate larger ligand in the binding pocket.

4.3.3.2 Mutating Away the Recognition of HN Group

An indene analog (Figure 4-6) is used in place of the crystal Trp ligand. In addition to Y159 and N194, we decided to also mutate away D237, even though it is a charged amino acid. It is, however, further away from the active site so its mutations should not play a role in the activation of enzyme.

Since we are trying to mutate away the recognition of the HN group, we allow only hydrophobic groups as mutation candidates at positions 159, 194 and 237. Following the procedures in Section 4.1.2, we report candidates for good mutations in Table 4-10. The picture of the active site with the top candidates is shown in Figure 4-11.

Position 159	Position 194	Position 237	Overall Energy (kcal/mol)
L	L	G	0.0
L	G	G	0.61
L	G	M	4.82
G	L	G	6.64
L	G	A	6.68
L	L	A	7.42
L	A	G	9.58

Table 4-10 Top mutation candidates for positions 159, 194 and 237 when the ligand in place is an indene amino acid analog.

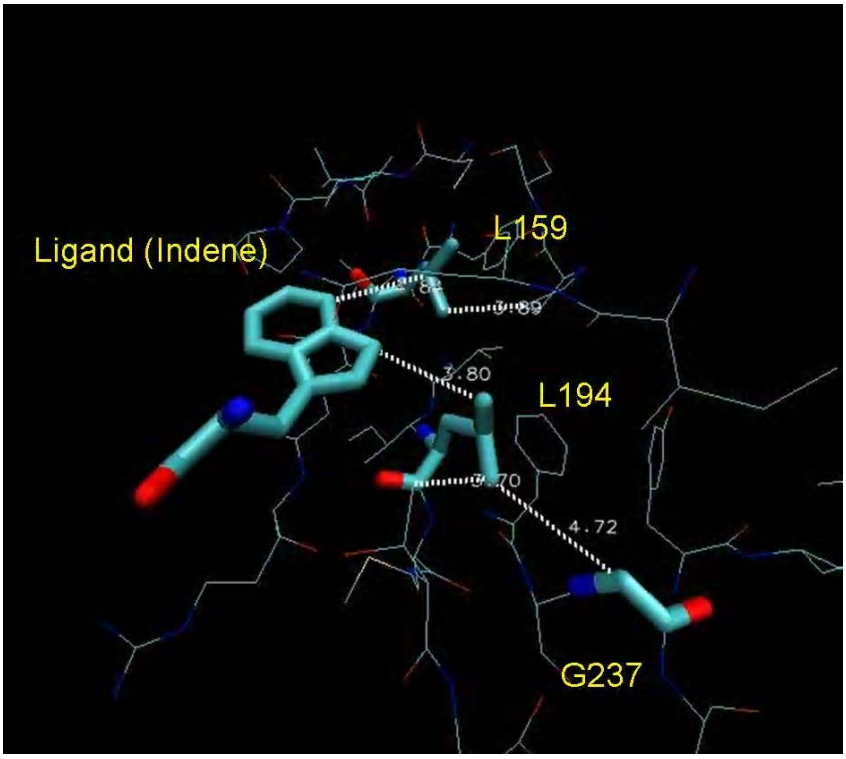


Figure 4-11 Binding site of 2dr2 with indene amino acid analog and top mutation candidates at positions 159, 194 and 237.

Differential binding energy between the indene analog and the cognate Trp ligand in the presence of the top L159-L194-G237 mutation is reported in Table 4-11. The final differential energy of 3.05 kcal/mol is close to the energy of a hydrogen bond, fulfilling the purpose of performing the mutation as described.

Ligand	Binding Energy (kcal/mol)
Cognate Trp	20.88
Indene-Analog	23.93
Differential Binding	3.05

Table 4-11 Binding energy of cognate ligand and indene analog with top mutation candidate at positions 159, 194, 237.

In the following sections we introduce mutations at the hydrophobic lining pocket that stabilize various amino acid analogs.

4.3.3.3 Ethynyl-Trp

Ethynyl-Trp is the minimal non-natural amino acid analog to have a basic tryptophan structure and an ethynyl functional group for Click chemistry. The ligand has no degrees of freedom on the ethynyl group, however, and unlike in the *B. Stearothermophilus* TrpRS case, there is some backbone clash involving the ethynyl atoms on the ligand. Therefore, we choose not to proceed with designs involving this non-natural amino acid.

4.3.3.4 Oxyethynyl-Trp

Oxyethynyl-Trp is slightly bigger than ethynyl-Trp, with an ether oxygen inserted between the main tryptophan ring and the ethynyl functional group. The presence of this ether oxygen gives the ethynyl group some rotational freedom. This freedom is restricted to just two local minima, however, because the ether oxygen is weakly in resonance with the tryptophan ring, lending it sp^2 character. The O_R

atom type is assigned to this oxygen. The two local minima place the ethynyl on the same plane as the tryptophan group.

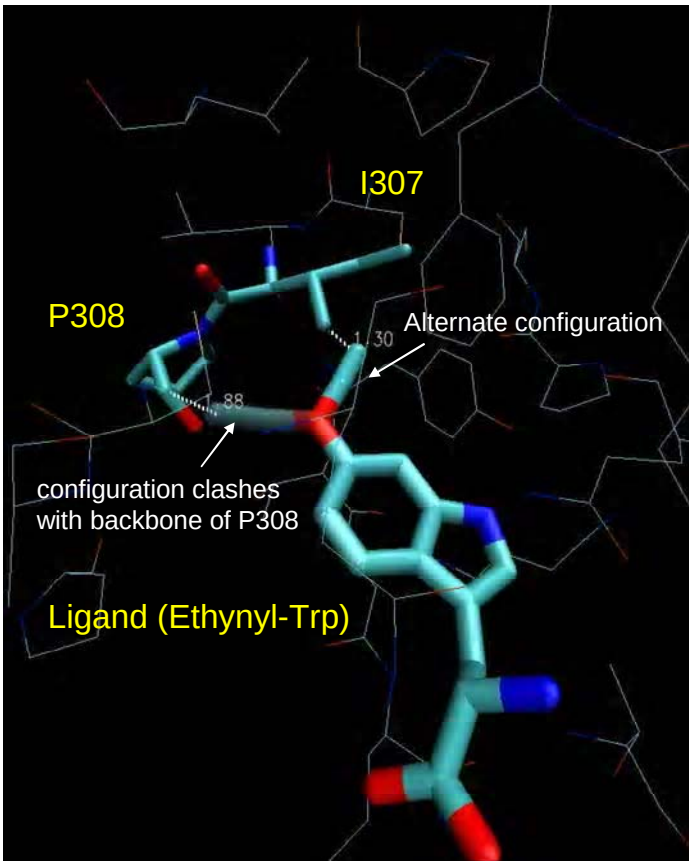


Figure 4-12 Two possible configurations for the oxyethynyl groups.

In addition to Ile307, Cys309 is also included as a position to introduce mutations. Following the procedure in Section 4.1.2, we report top mutation candidates as in Table 4-12.

Position 307	Position 309	Energy (kcal/mol)
G	G	0.0
G	A	1.17
A	G	7.11
A	A	7.38
C	A	9.02

Table 4-12 Top mutation candidates for Oxyethynyl-Trp for human TrpRS incorporation.

The top two mutations essentially allows room for the ether oxygen to reside in the binding pocket. Serines and threonines were also mutated at position 309, but no configuration could form a hydrogen bond with the oxygen on the ligand.

4.3.3.5 Propargyloxy-Trp

Again, like the previous cases, only the 5- position propargyloxy-Trp can be added to the human TrpRS binding site. This ligand adds another degree of rotational freedom onto the ligand, with an ether oxygen and a sp^3 carbon inserted between the trp ring and the ethynyl group. We manually generated $2 \cdot 3 = 6$ configurations for this ligand: 3 positions for the ethynyl group for each of the 2 positions of sp^3 carbon. We found that one of these configurations have a good potential as being the binding mode after we mutate away the clashing residues I307, F317, L334 and C309. Following procedures outlined in Section 4.1.2, we report the top mutation candidates in Table 4-13.

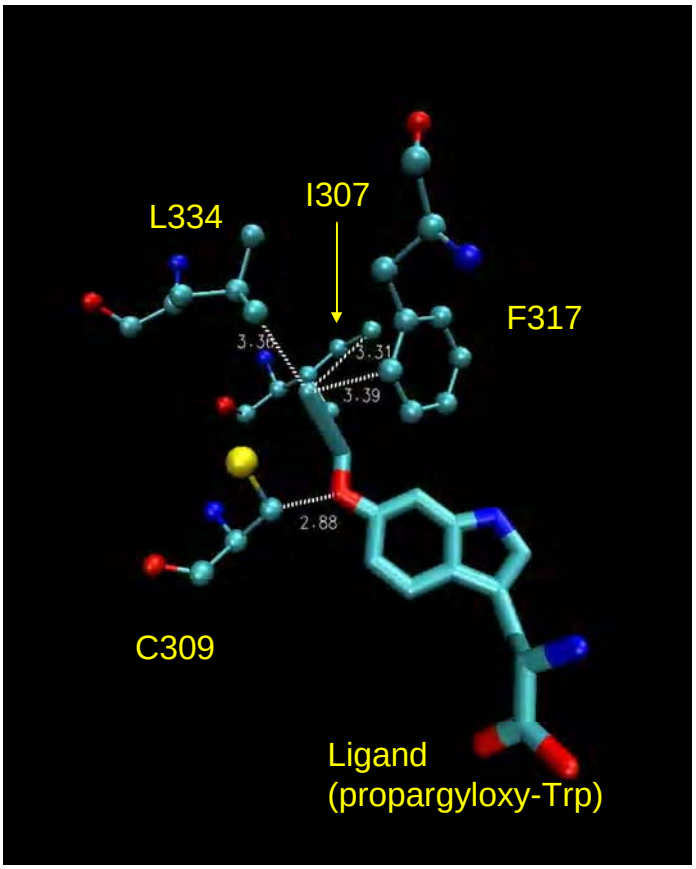


Figure 4-13 Starting configuration for the unnatural amino acid analog propargyloxy-Trp. The ethynyl group is off the Trp ring plane by 60°.

Position 317	Position 334	Position 307	Position 309	Energy (kcal/mol)
V	A	G	G	0.0
V	L	G	G	2.934
V	C	G	G	4.840
L	A	G	G	5.601
I	A	G	G	6.407

Table 4-13 Top mutation candidates for propargyloxy-Trp as ligand in human TrpRS.

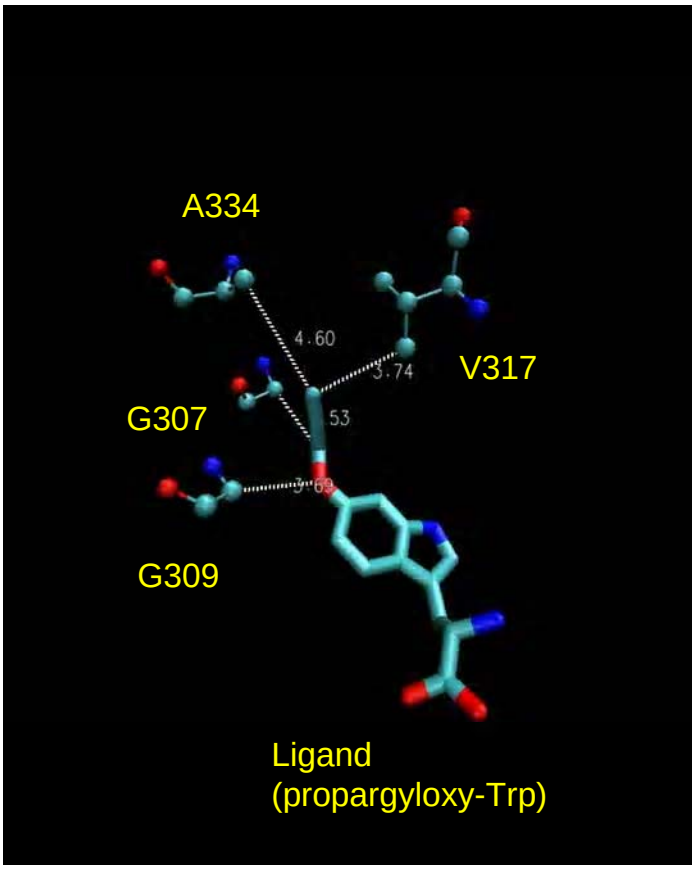


Figure 4-14 Top mutation candidate for propargyloxy-Trp as ligand in human TrpRS.

Shown in Figure 4-14 is the top mutation candidate V317-A334-G307-G309 for the non-natural amino acid propargyloxy-Trp. The glycine mutation at position 309 is necessary since the C_β atom would otherwise clash with the ether oxygen on the ligand. The case is similar at position 307 where the C_β atom would otherwise clash with the sp^3 carbon of the ligand. There are more design opportunities at positions 317 and 334, with two bulky residue, a phenylalanine and a leucine respectively, in the crystal structure. The mutation to an alanine and a valine allows space for the ethynyl functional group to reside. As can be seen from Table 4-13, many other hydrophobic mutations could also work out in this region.

4.3.3.6 Homoazido-Trp

The unnatural amino-acid analog homoazido-Trp has a similar rotational freedom profile as propargyloxy-Trp, but the sp^3 carbon between the azido group and the trp ring allows more freedom for positioning the azido group. We manually generated 6 rotamers for this ligand and found that the rotamer shown in Figure 4-15 is the only rotamer that does not have bad contacts with the protein backbone. This ligand configuration is used for the purpose of this calculation.

Using the same procedure as before, we found that the only good mutation for this ligand is a double glycine mutation at both position 307 and 309.

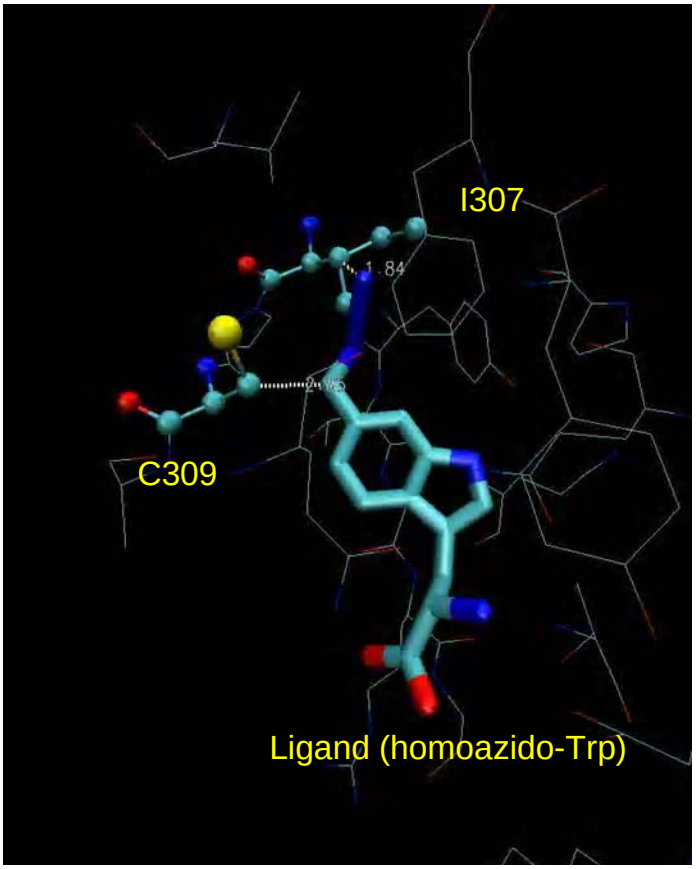


Figure 4-15 Homoazido-Trp as ligand. Clashes with C309 and I307 are shown.

4.3.3.7 Benzo-cyclo-octyne Analog

The benzo-cyclo-octyne amino acid analogue can react with an azido group without the need of a catalyst, with an activation barrier of just 7.6 kcal/mol. This is due to the tremendous strain in the cyclo-octyne ring.

The cyclo-octyne group can be fused with the benzyl ring at two different positions, as shown in Figure 4-16. In addition, the ethyne group will be out-of-plane with the benzyl ring, leading to two distinct configurations. Thus, we have a total of 4 possible configurations to consider for this amino acid analogue.

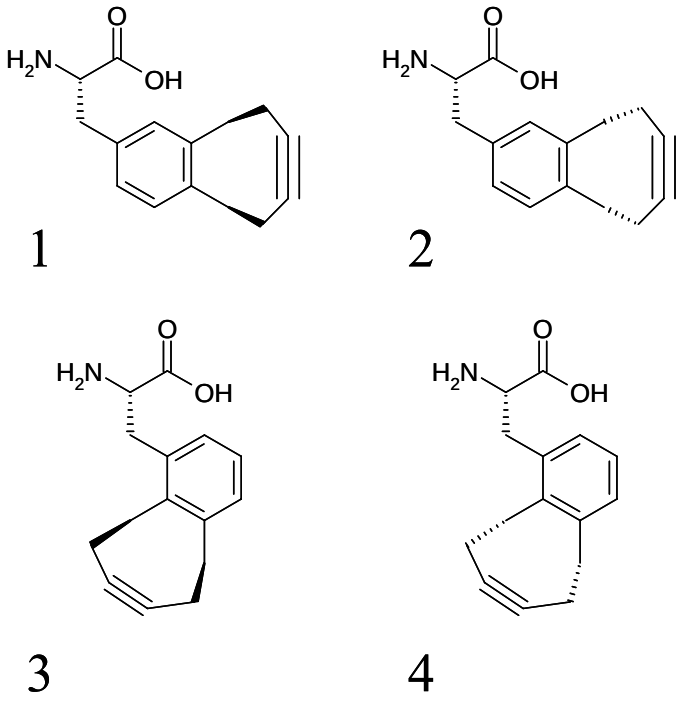


Figure 4-16 Possible configurations of benzo-cyclo-octyne amion acid analog.

We constructed all 4 configurations and placed them into the human TrpRS crystal structure by a superposition with the cognate tryptophan. As shown in Figure 4-17, there is only one configuration that does not clash with either a protein backbone or

a crucial amino acid. Asn 313 is involved in the recognition of the backbone of the amino acid. Since TrpRS catalyzes the acylation of amino acids to the tRNA, mutation of this Asn 313 will likely lead to a loss of this function and is therefore not considered a mutation candidate.

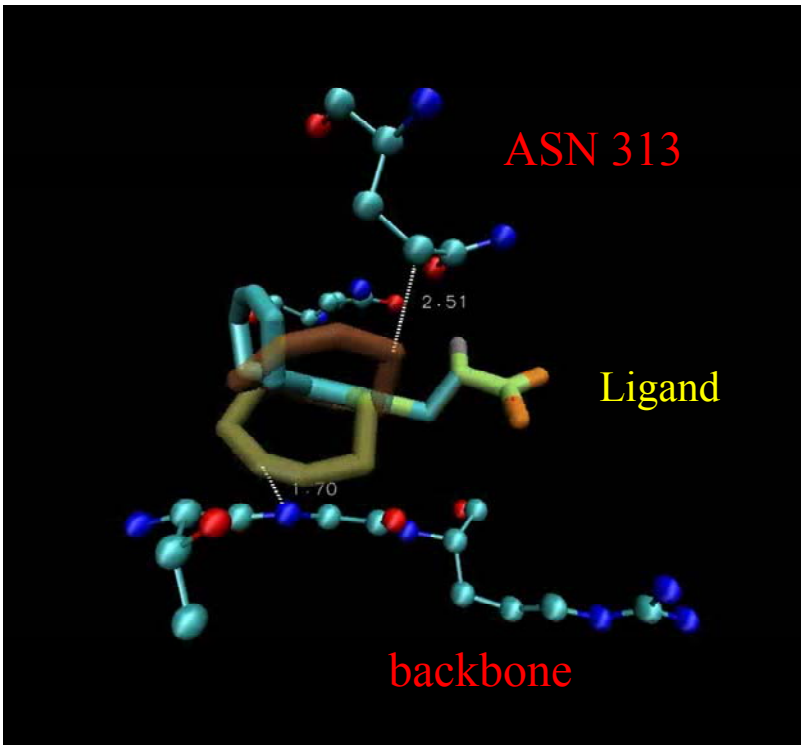


Figure 4-17 Three out of four possible configurations of benzo-cyclo-octyne amino acid analogue are shown in the human TrpRS crystal structure active site.

Following the procedure from Section 4.1.2, we identified six residues to be mutated from the active site. These six residues are divided into two groups. The first is the Y159-N194-D239 group which in the crystal structure forms a hydrogen bond network and recognizes the HN on the cognate trptophan ligand. The second group comprises of F317-I309-C307, which lines the hydrophobic pocket. These residues are shown in ball-and-stick representation in Figure 4-18.

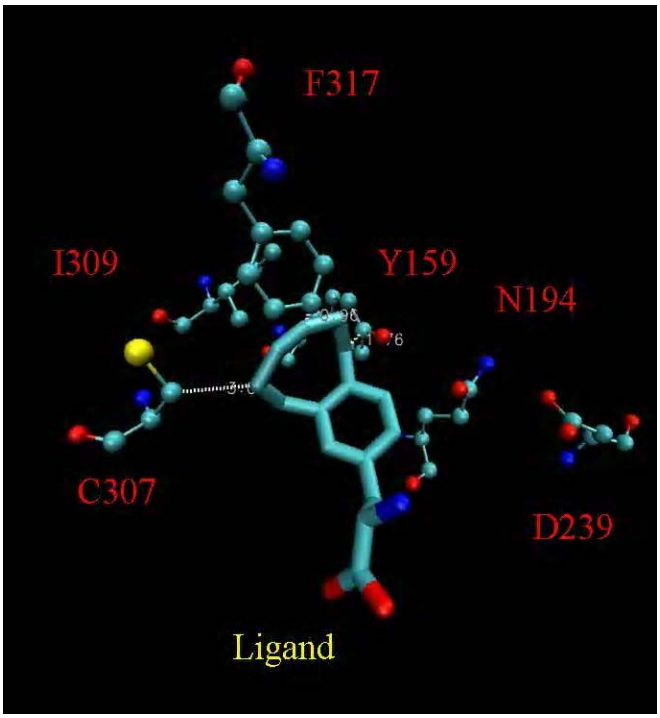


Figure 4-18 Mutation candidates for TrpRS for incorporation of benzo-cyclo-octyne amino acid analogue.

Position 237	Position 159	Position 194	Energy (kcal/mol)
A	L	M	0.0
A	L	L	6.0
C	L	M	7.8
A	M	M	9.0
C	L	L	9.6
A	L	C	10.0

Table 4-14 Top mutation candidates for positions 237, 159 and 194 in human TrpRS in the presence of benzo-cyclo-octyne amino acid analog.

Position 317	Position 307	Position 309	Energy (kcal/mol)
L	L	A	0.0
A	L	A	2.0
L	V	M	2.7
C	L	A	4.1
M	L	C	5.7
L	V	L	6.0

Table 4-15 Top mutation candidates for positions 317, 307 and 309 in human TrpRS in the presence of benzo-cyclo-octyne amino acid analog.

Reported in Table 4-14 and Table 4-15 are the top mutation candidates as predicted using the procedure outlined in 4.1.2.

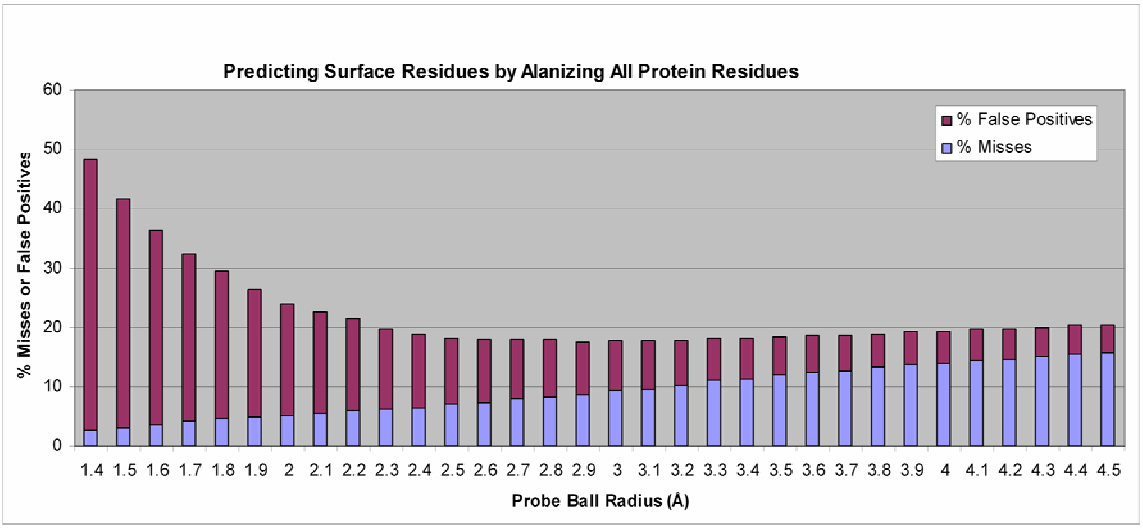
4.3.3.8 Conclusions

In this section, we have applied our methods on a few protein design cases. We have succeeded in predicting mutants that contribute to the stability of IFN-β. Predictions have also been made for ligands that participate in click chemistry for TrpRS incorporation. We have submitted our predictions to our collaborators, and would be interested to find out the performance of our prediction.

Appendix A SCREAM Supplementary Material

A.1 Prediction of Surface Residues Prior to Sidechain Assignment

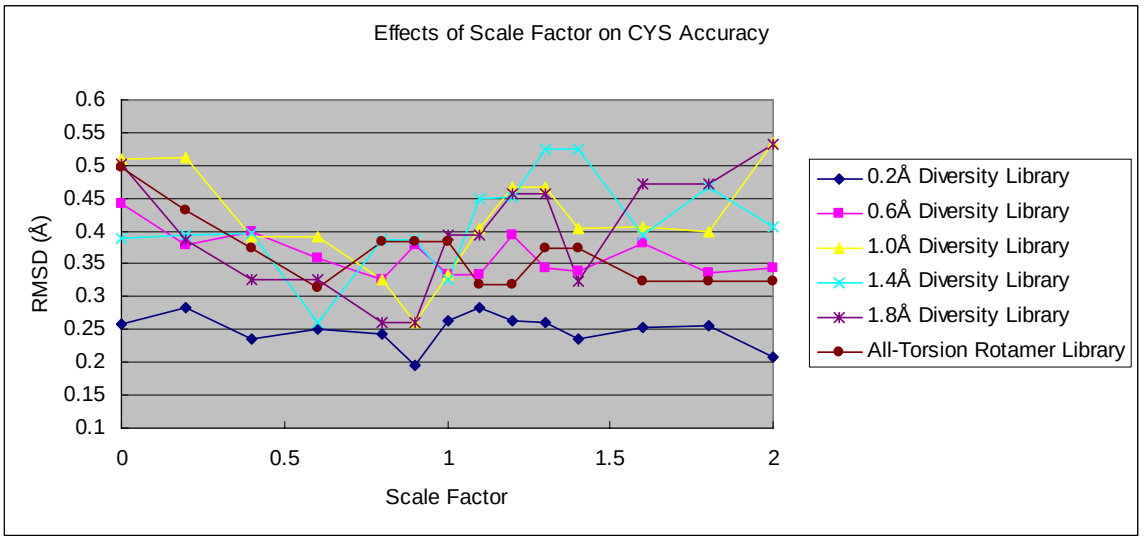
SCREAM does not currently distinguish between surface and bulk residues in its calculation. In order to predict the surface residues prior to assigning the sidechains, we recommend using the alanized protein and calculating the solvent exposed surface area (SASA) by rolling a ball of 2.9 Å instead of the standard 1.4 Å, as shown in Figure A. A "miss" is defined when the algorithm does not find an exposed residue as in the original crystal structure. A "false positive" is defined when the algorithm assigns a residue as exposed but is in fact buried in the original crystal structure. The usual 20% exposed surface area criterion is used for determining whether a residue or its alanized is exposed or buried. Based on the results, using a probe ball radius of 2.9 Å minimizes the sum of false positive and misses. The Xiang set of proteins is used for testing.

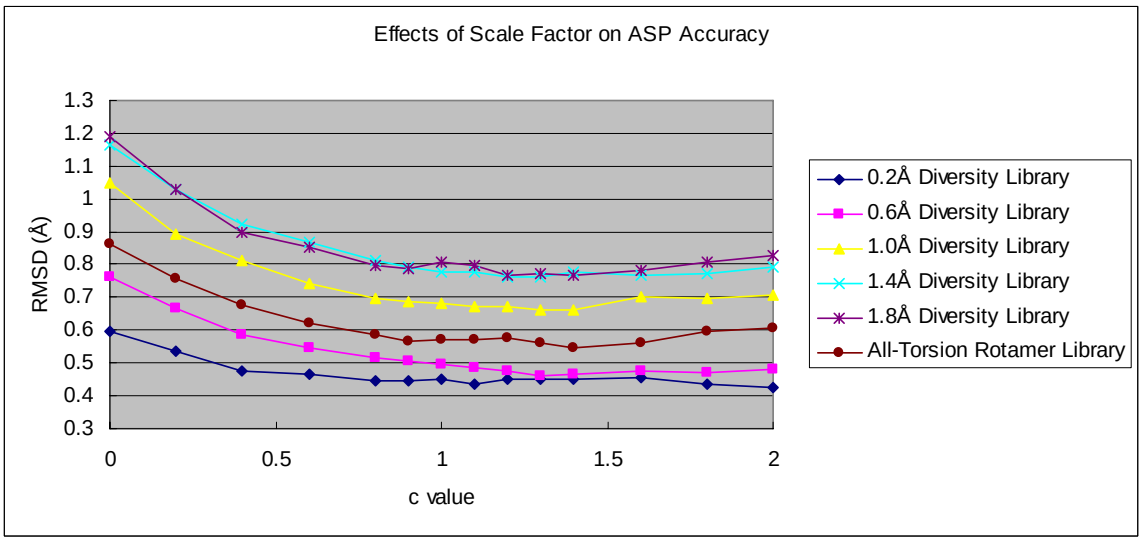


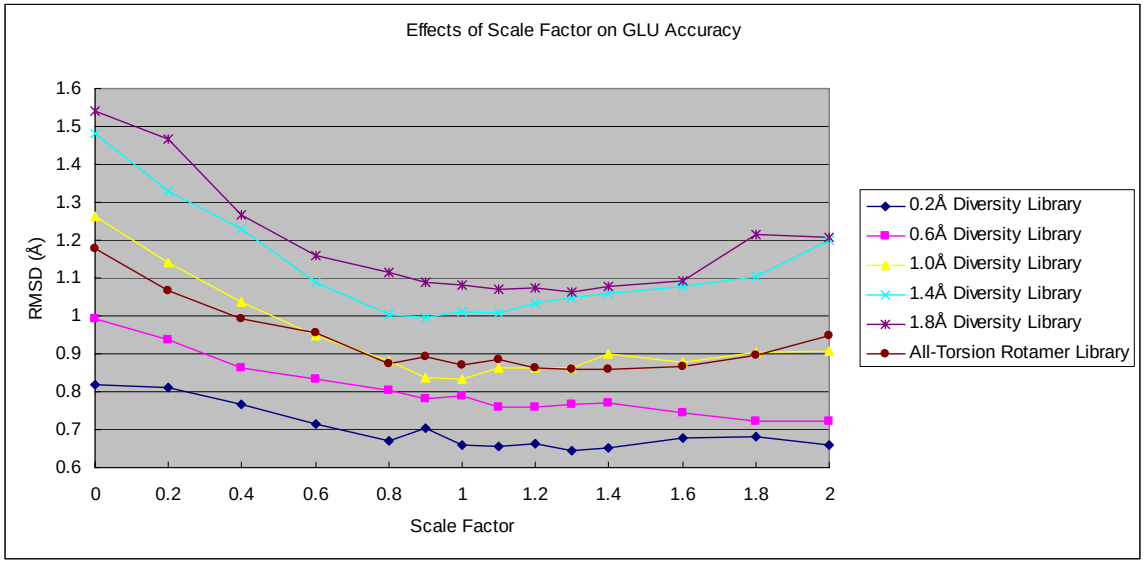
The previous figure shows the accuracy of predicting surface residues using the method described. The percentage is calculated based on the number of exposed residues in the crystal structure (after removing waters and solvents).

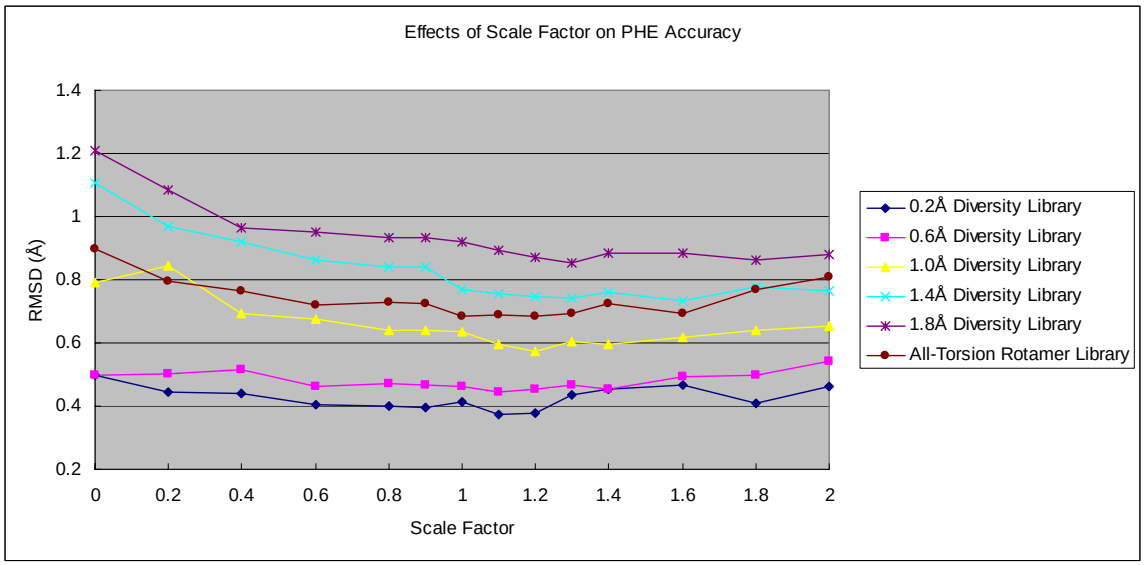
A.2 Impact of Scaling Factor *s* in Combinatorial Placement:

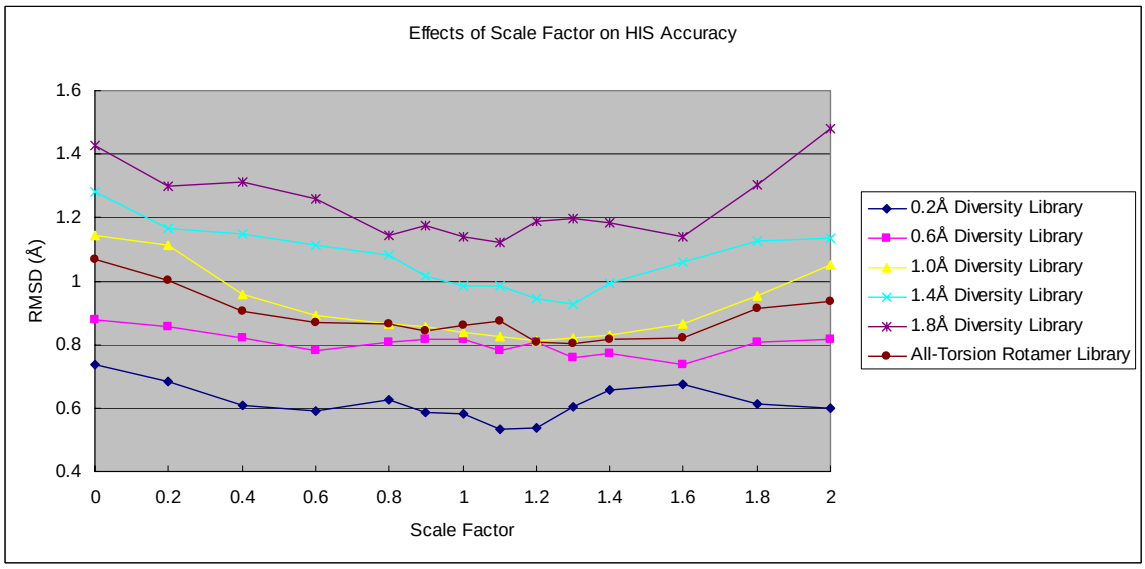
Here we include plots for all 17 sidechains. RMSD is plotted against the Scale factor *s*. The Xiang set of proteins are used for testing.

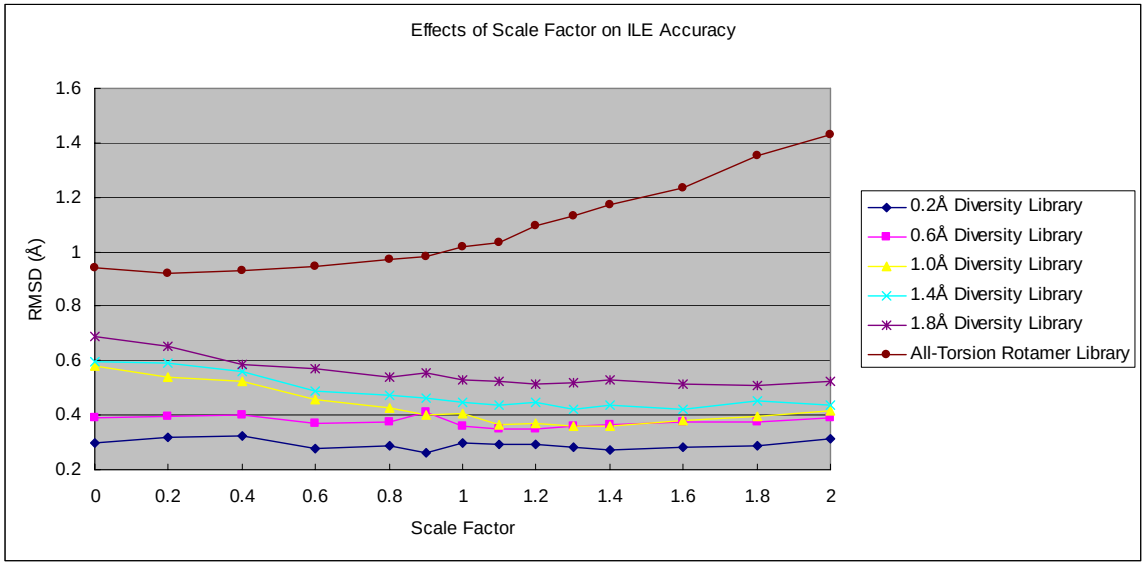


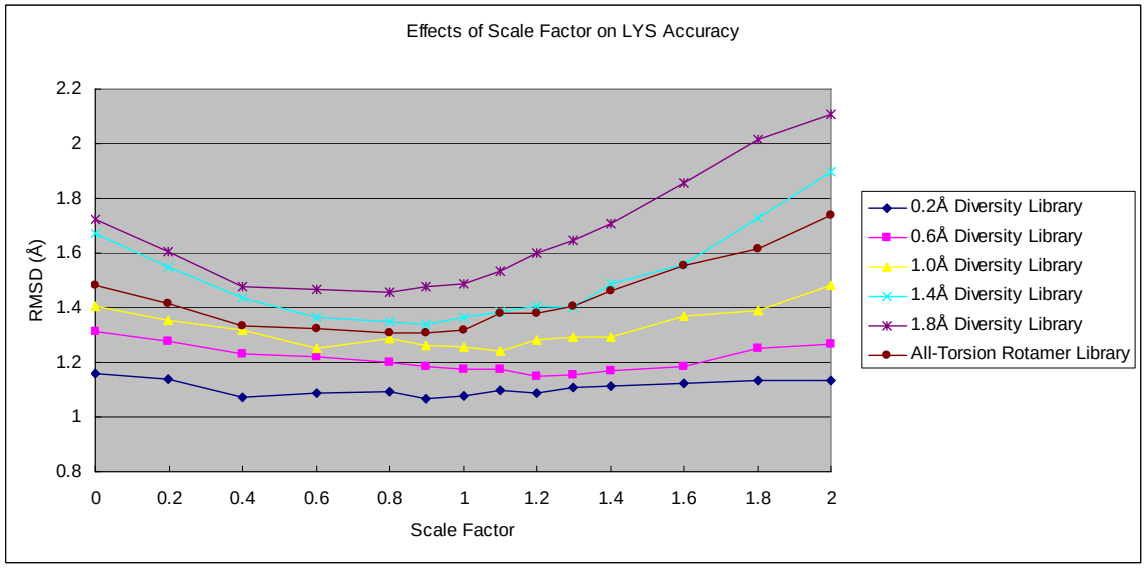


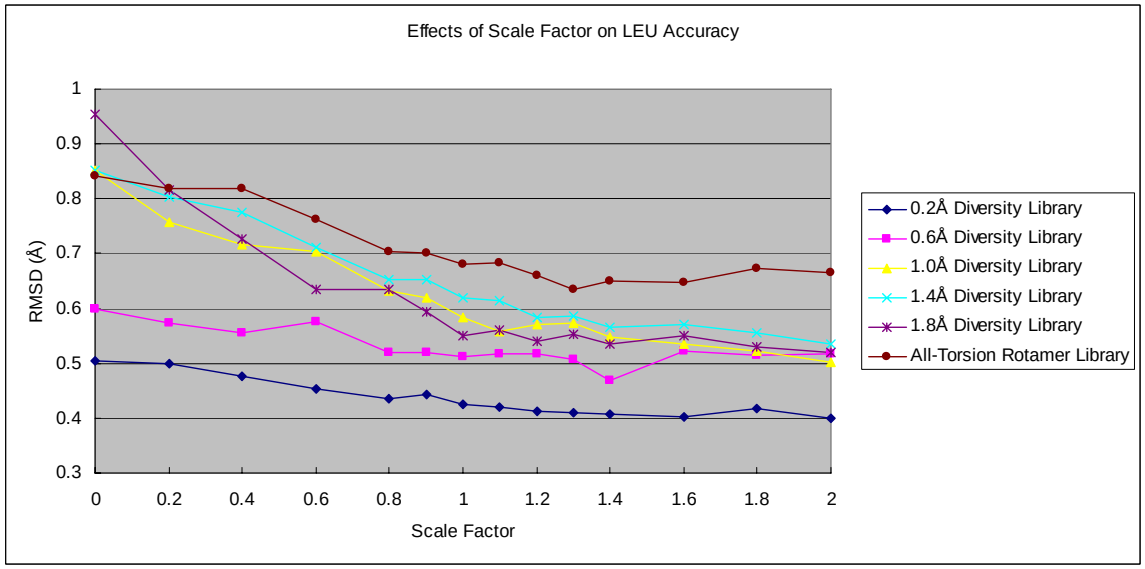

Effects of Scale Factor on GLU Accuracy
RMSD (Å)
1.6
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Scale Factor
0.2Å Diversity Library
0.6Å Diversity Library
1.0Å Diversity Library
1.4Å Diversity Library
1.8Å Diversity Library
All-Torsion Rotamer Library

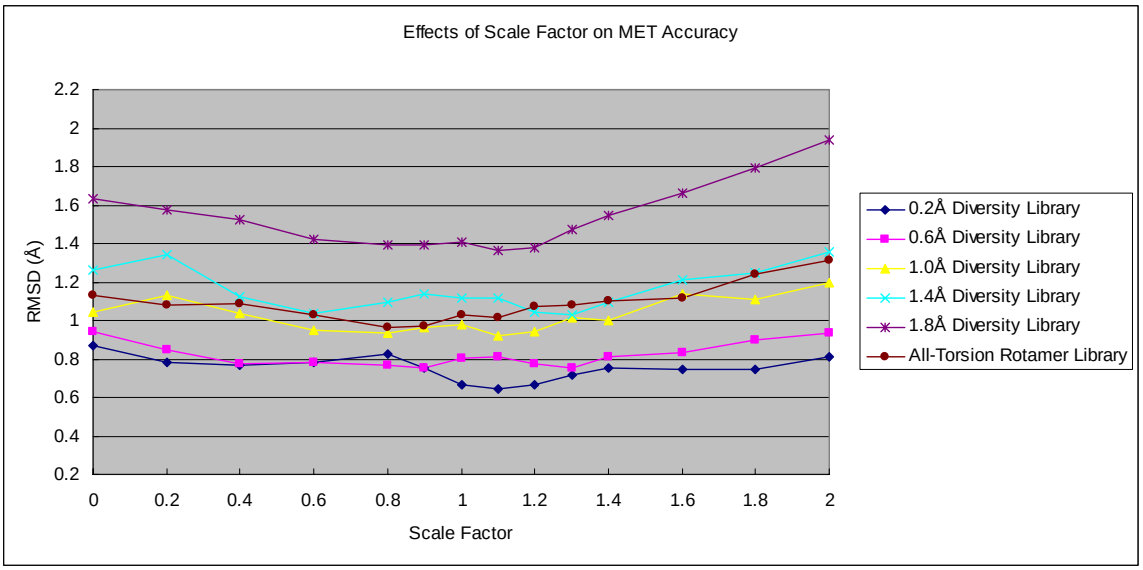

Effects of Scale Factor on PHE Accuracy
RMSD (Å)
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Scale Factor
0.2Å Diversity Library
0.6Å Diversity Library
1.0Å Diversity Library
1.4Å Diversity Library
1.8Å Diversity Library
All-Torsion Rotamer Library

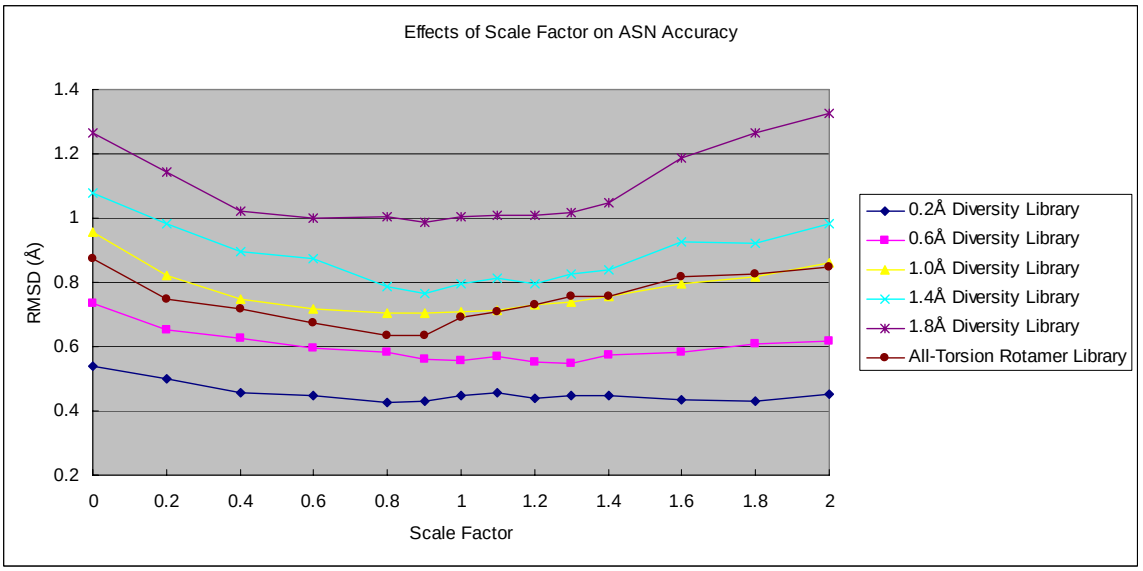

Effects of Scale Factor on HIS Accuracy
RMSD (Å)
1.6
1.4
1.2
1
0.8
0.6
0.4
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Scale Factor
0.2Å Diversity Library
0.6Å Diversity Library
1.0Å Diversity Library
1.4Å Diversity Library
1.8Å Diversity Library
All-Torsion Rotamer Library

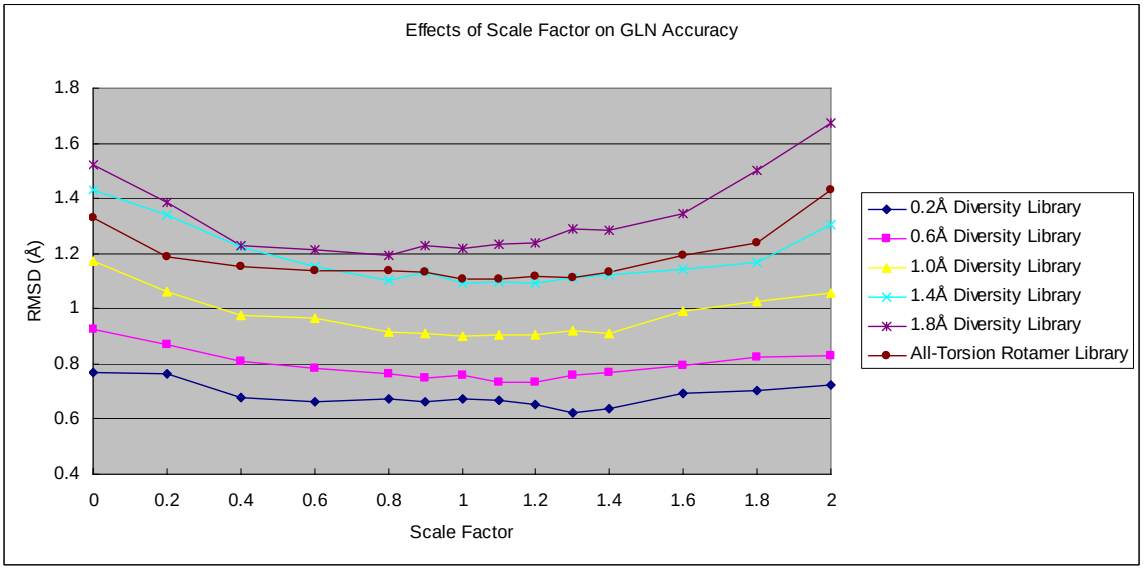

Effects of Scale Factor on ILE Accuracy
RMSD (Å)
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Scale Factor
0.2Å Diversity Library
0.6Å Diversity Library
1.0Å Diversity Library
1.4Å Diversity Library
1.8Å Diversity Library
All-Torsion Rotamer Library

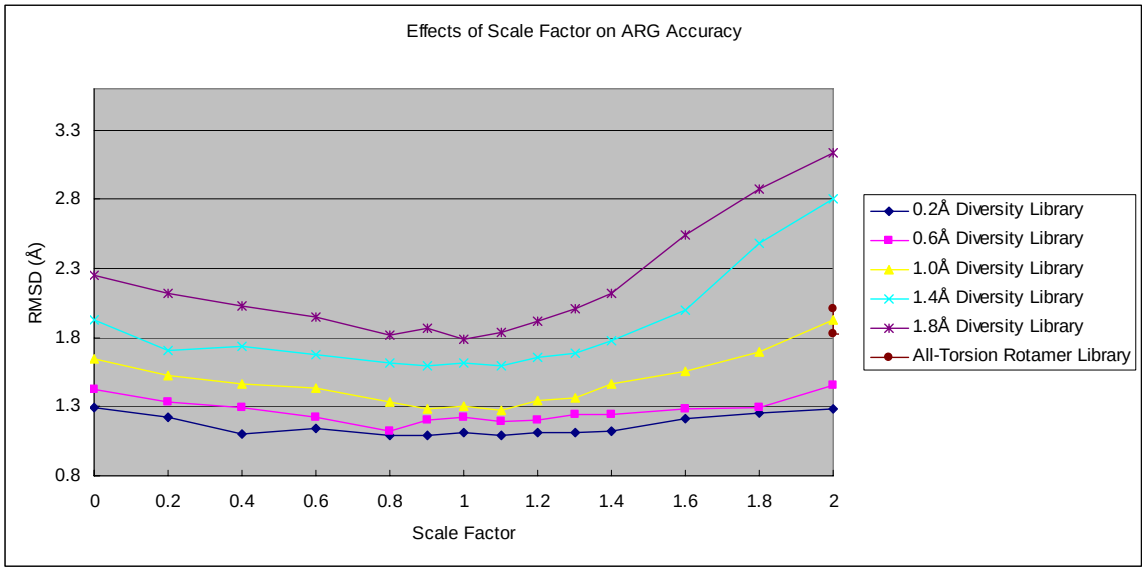

Effects of Scale Factor on LYS Accuracy
RMSD (Å)
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Scale Factor
0.2Å Diversity Library
0.6Å Diversity Library
1.0Å Diversity Library
1.4Å Diversity Library
1.8Å Diversity Library
All-Torsion Rotamer Library

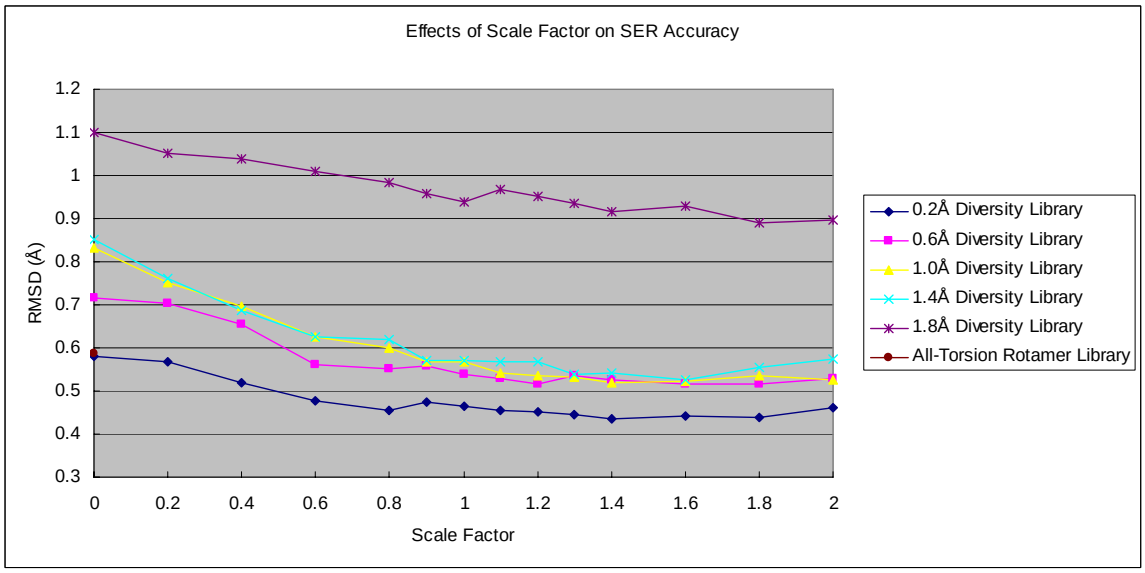

Effects of Scale Factor on LEU Accuracy
RMSD (Å)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Scale Factor
0.2Å Diversity Library
0.6Å Diversity Library
1.0Å Diversity Library
1.4Å Diversity Library
1.8Å Diversity Library
All-Torsion Rotamer Library

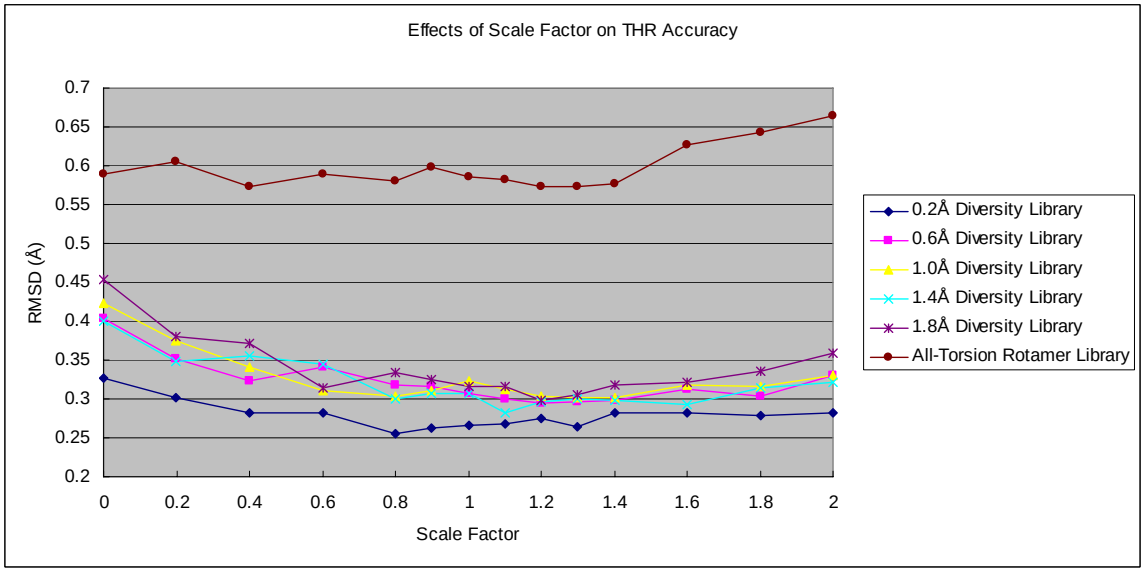

Effects of Scale Factor on MET Accuracy
RMSD (Å)
Scale Factor
0.2Å Diversity Library
0.6Å Diversity Library
1.0Å Diversity Library
1.4Å Diversity Library
1.8Å Diversity Library
All-Torsion Rotamer Library

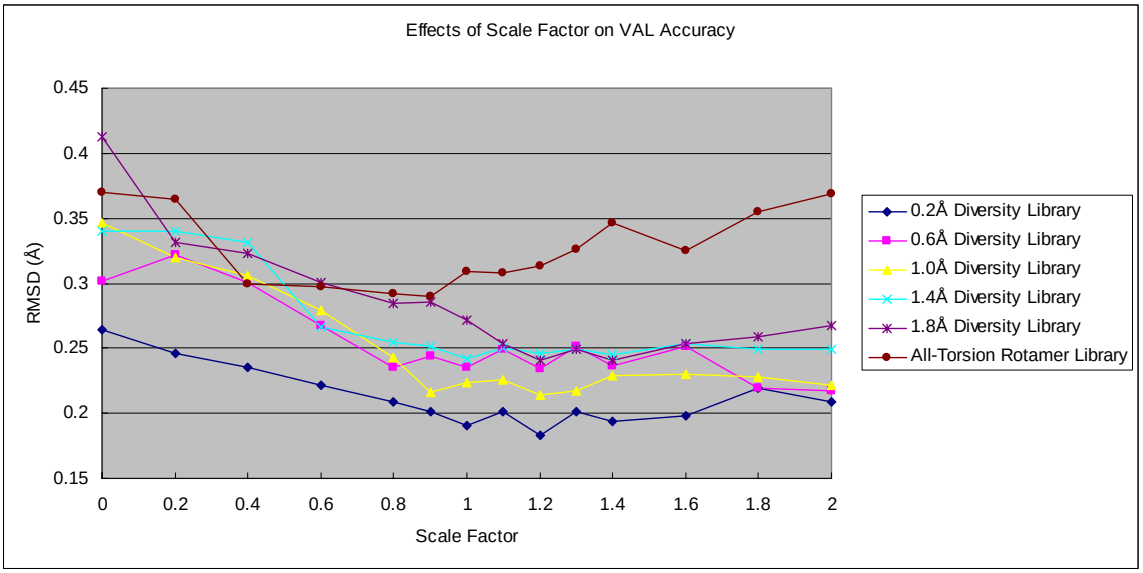

Effects of Scale Factor on ASN Accuracy
RMSD (Å)
Scale Factor
0.2Å Diversity Library
0.6Å Diversity Library
1.0Å Diversity Library
1.4Å Diversity Library
1.8Å Diversity Library
All-Torsion Rotamer Library

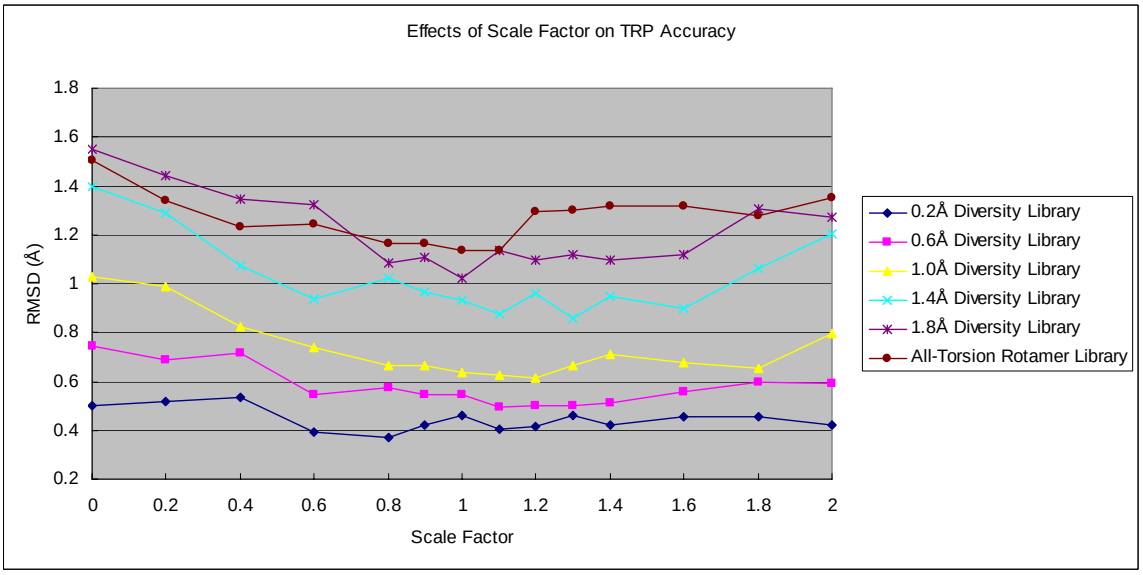

Effects of Scale Factor on GLN Accuracy
RMSD (Å)
Scale Factor
0.2Å Diversity Library
0.6Å Diversity Library
1.0Å Diversity Library
1.4Å Diversity Library
1.8Å Diversity Library
All-Torsion Rotamer Library

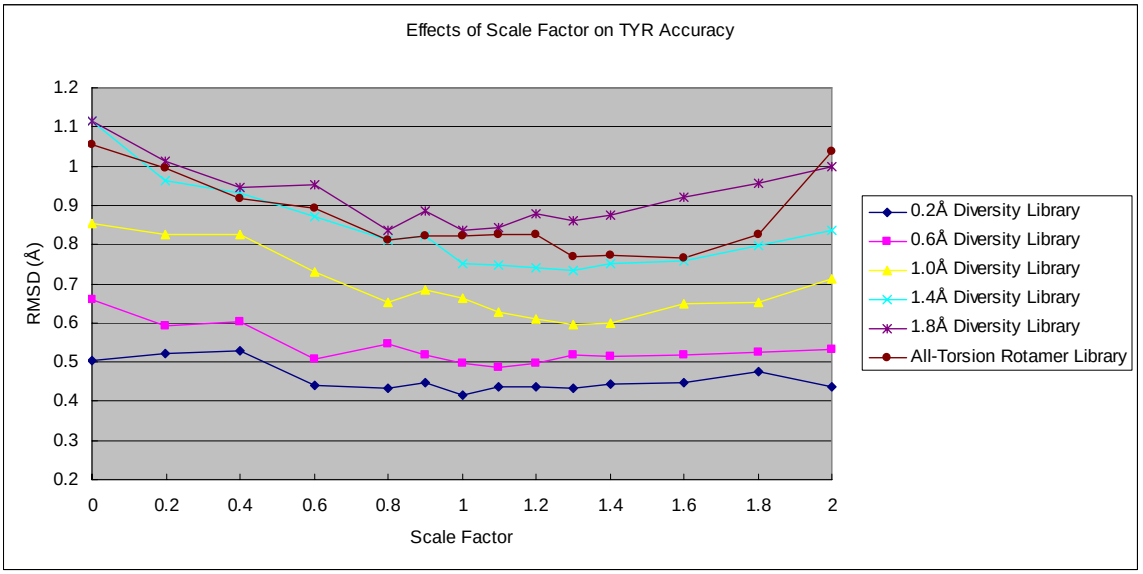

Effects of Scale Factor on ARG Accuracy
RMSD (Å)
3.3
2.8
2.3
1.8
1.3
0.8
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Scale Factor
0.2Å Diversity Library
0.6Å Diversity Library
1.0Å Diversity Library
1.4Å Diversity Library
1.8Å Diversity Library
All-Torsion Rotamer Library


Effects of Scale Factor on SER Accuracy
RMSD (Å)
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Scale Factor
0.2Å Diversity Library
0.6Å Diversity Library
1.0Å Diversity Library
1.4Å Diversity Library
1.8Å Diversity Library
All-Torsion Rotamer Library


Effects of Scale Factor on THR Accuracy
RMSD (Å)
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Scale Factor
0.2Å Diversity Library
0.6Å Diversity Library
1.0Å Diversity Library
1.4Å Diversity Library
1.8Å Diversity Library
All-Torsion Rotamer Library


Effects of Scale Factor on VAL Accuracy
RMSD (Å)
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Scale Factor
0.2Å Diversity Library
0.6Å Diversity Library
1.0Å Diversity Library
1.4Å Diversity Library
1.8Å Diversity Library
All-Torsion Rotamer Library


Effects of Scale Factor on TRP Accuracy
RMSD (Å)
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Scale Factor
0.2Å Diversity Library
0.6Å Diversity Library
1.0Å Diversity Library
1.4Å Diversity Library
1.8Å Diversity Library
All-Torsion Rotamer Library


Effects of Scale Factor on TYR Accuracy
RMSD (Å)
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Scale Factor
0.2Å Diversity Library
0.6Å Diversity Library
1.0Å Diversity Library
1.4Å Diversity Library
1.8Å Diversity Library
All-Torsion Rotamer Library

Appendix B Algorithms Used in Protein Neutralization

B.1 Flow Network

Neutralizing charged residues means adding or removing protons from the residue. Because of this procedure, in order to maximize the total number of hydrogen bonds in the system, sometimes the hydrogen positions of other polar groups need to be modified as well. Histidine with two possible spots for the polar hydrogen, and water with multiple locations for its two polar hydrogens provide such examples.

Graph theory algorithms are used to ensure that the maximum number of hydrogen bonds is retained after the neutralization procedure. Since hydrogens must always be removed from positively charged groups and always be added to negatively charged groups, the movement of protons can be modeled as a flow network. The basic idea is that maximizing the number of hydrogen bonds in the protein would correspond to maximizing the flow in the network, a well-studied problem in computer science known as Maximum-Flow.

Using nonmenclature from Cormen[89], a flow network G=(V,E) is a directed graph in which each edge (u,v) $\in$ E has a nonnegative capacity c(u,v) ≥ 0. Two vertices are special: a *source* s and a *sink* t. A flow, formally, is a function $f : V \times V \to \Re$ that satisfies:

Capacity constraint: $f(u,v) \leq c(u,v) \qquad \forall u, v \in V$

Skew symmetry: $f(u,v) = -f(u,v) \qquad \forall u, v \in V$

Flow conservation: $\sum_{v \in V} f(u,v) = 0 \qquad \forall u \in V - \{s,t\}$

The correspondence of these terms to our hydrogen network is as follows:

Flow Network Terms	**Hydrogen Bond Network**
Vertex	Residue/group (or virtual group)
Directed Edge	Hydrogen bond, and the donor/acceptor pair
Capacity	Maximum number of hydrogen bonds from donor to acceptors
Sink/source	Positively charged residues/negatively charged residues
Flow	Transfer of protons

For charged groups, only one vertex is needed. For non-charged groups, special treatment is necessary, since we need to account for the fact that they can "bridge" only a limited number of oppositely charged groups. For instance, Histidine can only accommodate the protein transfer between one pair of oppositely charged groups. To do this, these bridging groups are turned into two virtual groups, and we assign the capacity of the edge between these two virtual groups to be the maximum number of proton transfers it can handle. For Histidine, this number would be one. The first virtual group handles just the accepting of protons, whereas the second virtual group handles just the donating of protons. See the remark later in this Appendix for a proof of the correctness of this procedure.

Since there can be multiple positively residues (source vertices) and negatively charged residues (sink vertices), we add a supersource and a supersink. These special nodes connect to regular sink and sources, respectively, by constructing directed edges. The capacity of these edges is infinite, so as to push the maximum amount of flow through the

network, which corresponds to maximizing the number of hydrogen bonds made due to the reassignment of protons.

The problem is now formally reduced to finding the maximum flow of this graph we have constructed, i.e. between the supersource and the supersink vertices. The Ford-Fulkerson method is used to find the maximum flow. The number of hydrogen bonds each residue/group can accept or donor is typically a small integer (not greater than two), so efficiency is not an issue. A potential hydrogen bond is defined as any two acceptor/donor pair that comes within 3.5Å. This is a value that can be adjusted. For groups/vertices that are not solved in this flow network (they are typically solvent exposed and do not interact with the rest of the protein), the hydrogen adding and removal rules below are used:

1. If there are multiple hydrogens that can be removed (e.g. NH_3^+), pick any one that does not remove existing hydrogen bond with other atoms.

2. If there are multiple positions the hydrogen can be added on (e.g. either oxygen on CO_2^-), pick any one that makes a hydrogen bond with other atoms.

If Rule 1 and Rule 2 do not apply, creation or removal of hydrogen is arbitrary.

The above rules are also applied to decide tie-breaks on the flow network algorithm.

B.2 Remark on Correctness

A proof is provided for ensuring that the introduction of bridging vertices does not affect change the number of protons that can be transferred.

The introduction of bridging vertices does indeed constrain the capacity of the network to the capacity of the bridging groups.

Proof: Apply the Max-flow min-cut theorem. Cut the graph in two, separating the first of all the virtual bridging vertices and the second of all virtual bridging vertices. This cut then

has value less than or equal to the sum of the capacity of all the bridging groups (since it might not be possible to find a cut that achieves equality). However, by max-flow min-cut theorem, the value of any flow in a flow network is bounded from above by the capacity of any cut. Thus, the introduction of the bridging vertices has the intended effect.

References

1. Adams, C. P.; Brantner, V. V. Health Aff 2006, 25(2), 420-428.
2. Bolon, D. N.; Grant, R. A.; Baker, T. A.; Sauer, R. T. Proceedings of the National Academy of Sciences of the United States of America 2005, 102(36), 12724-12729.
3. Zhang, D. Q.; Vaidehi, N.; Goddard, W. A.; Danzer, J. F.; Debe, D. Proceedings of the National Academy of Sciences of the United States of America 2002, 99(10), 6579-6584.
4. Desmet, J.; Demaeyer, M.; Hazes, B.; Lasters, I. Nature 1992, 356(6369), 539-542.
5. Dunbrack, R. L.; Karplus, M. Journal of Molecular Biology 1993, 230(2), 543-574.
6. Koehl, P.; Delarue, M. Journal of Molecular Biology 1994, 239(2), 249-275.
7. Vasquez, M. Biopolymers 1995, 36(1), 53-70.
8. Petrella, R. J.; Lazaridis, T.; Karplus, M. Folding & Design 1998, 3(5), 353-377.
9. Desmet, J.; Spriet, J.; Lasters, I. Proteins-Structure Function and Genetics 2002, 48(1), 31-43.
10. Ponder, J. W.; Richards, F. M. Journal of Molecular Biology 1987, 193(4), 775-791.
11. Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. Proteins-Structure Function and Genetics 2000, 40(3), 389-408.
12. Levitt, M.; Gerstein, M.; Huang, E.; Subbiah, S.; Tsai, J. Annual Review of Biochemistry 1997, 66, 549-579.
13. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. Journal of Computational Chemistry 1983, 4(2), 187-217.
14. Mayo, S. L.; Olafson, B. D.; Goddard, W. A. Journal of Physical Chemistry 1990, 94(26), 8897-8909.
15. Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. Comput Phys Commun 1995, 91(1-3), 1-41.
16. Shea, J. E.; Brooks, C. L. Annual Review of Physical Chemistry 2001, 52, 499-535.
17. Eaton, W. A.; Munoz, V.; Hagen, S. J.; Jas, G. S.; Lapidus, L. J.; Henry, E. R.; Hofrichter, J. Annual Review of Biophysics and Biomolecular Structure 2000, 29, 327-359.
18. Halperin, I.; Ma, B. Y.; Wolfson, H.; Nussinov, R. Proteins-Structure Function and Genetics 2002, 47(4), 409-443.
19. Schueler-Furman, O.; Wang, C.; Bradley, P.; Misura, K.; Baker, D. Science 2005, 310(5748), 638-642.
20. Lazaridis, T.; Karplus, M. Current Opinion in Structural Biology 2000, 10(2), 139-145.
21. Trabanino, R. J.; Hall, S. E.; Vaidehi, N.; Floriano, W. B.; Kam, V. W. T.; Goddard, W. A. Biophysical Journal 2004, 86(4), 1904-1921.
22. Vaidehi, N.; Kalani, Y. S.; Hall, S. E.; Freddolino, P. L.; Trabanino, R. J.; Floriano, W. B.; Spijker, P.; Goddard, W. A. Biophysical Journal 2005, 88(1), 357A-357A.

23. Vaidehi, N.; Floriano, W. B.; Trabanino, R.; Hall, S. E.; Freddolino, P.; Choi, E. J.; Zamanakos, G.; Goddard, W. A. Proceedings of the National Academy of Sciences of the United States of America 2002, 99(20), 12622-12627.
24. Malakauskas, S. M.; Mayo, S. L. Nature Structural Biology 1998, 5(6), 470-475.
25. Kraemer-Pecore, C. M.; Lecomte, J. T. J.; Desjarlais, J. R. Protein Science 2003, 12(10), 2194-2205.
26. Dwyer, M. A.; Looger, L. L.; Hellinga, H. W. Science 2004, 304(5679), 1967-1971.
27. Brooijmans, N.; Kuntz, I. D. Annual Review of Biophysics and Biomolecular Structure 2003, 32, 335-373.
28. Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A. Proteins-Structure Function and Bioinformatics 2004, 55(2), 351-367.
29. Al-Lazikani, B.; Jung, J.; Xiang, Z. X.; Honig, B. Current Opinion in Chemical Biology 2001, 5(1), 51-56.
30. Pierce, N. A.; Winfree, E. Protein Engineering 2002, 15(10), 779-782.
31. Mendes, J.; Soares, C. M.; Carrondo, M. A. Biopolymers 1999, 50(2), 111-131.
32. Kussell, E.; Shimada, J.; Shakhnovich, E. I. Journal of Molecular Biology 2001, 311(1), 183-193.
33. Lasters, I.; Demaeyer, M.; Desmet, J. Protein Engineering 1995, 8(8), 815-822.
34. Pierce, N. A.; Spriet, J. A.; Desmet, J.; Mayo, S. L. Journal of Computational Chemistry 2000, 21(11), 999-1009.
35. Looger, L. L.; Hellinga, H. W. Journal of Molecular Biology 2001, 307(1), 429-445.
36. Eisenmenger, F.; Argos, P.; Abagyan, R. Journal of Molecular Biology 1993, 231(3), 849-860.
37. Xiang, Z. X.; Honig, B. Journal of Molecular Biology 2001, 311(2), 421-430.
38. Liang, S. D.; Grishin, N. V. Protein Sci 2002, 11(2), 322-331.
39. Peterson, R. W.; Dutton, P. L.; Wand, A. J. Protein Science 2004, 13(3), 735-751.
40. DeMaeyer, M.; Desmet, J.; Lasters, I. Folding & Design 1997, 2(1), 53-66.
41. Dahiyat, B. I.; Mayo, S. L. Proceedings of the National Academy of Sciences of the United States of America 1997, 94(19), 10172-10177.
42. Kuhlman, B.; Baker, D. Proceedings of the National Academy of Sciences of the United States of America 2000, 97(19), 10383-10388.
43. Desjarlais, J. R.; Handel, T. M. Protein Science 1995, 4(10), 2006-2018.
44. Wernisch, L.; Hery, S.; Wodak, S. J. Journal of Molecular Biology 2000, 301(3), 713-736.
45. Mendes, J.; Baptista, A. M.; Carrondo, M. A.; Soares, C. M. Proteins-Structure Function and Genetics 1999, 37(4), 530-543.
46. Canutescu, A. A.; Shelenkov, A. A.; Dunbrack, R. L. Protein Science 2003, 12(9), 2001-2014.
47. Jain, T.; Cerutti, D. S.; McCammon, J. A. Protein Science 2006, 15(9), 2029-2039.
48. Vriend, G. J Mol Graph 1990, 8(1), 52-&.
49. Schuttelkopf, A. W.; van Aalten, D. M. F. Acta Crystallographica Section D-Biological Crystallography 2004, 60, 1355-1363.

50. Holm, L.; Sander, C. Proteins-Structure Function and Genetics 1992, 14(2), 213-223.
51. Rappe, A. K.; Goddard, W. A. Journal of Physical Chemistry 1991, 95(8), 3358-3363.
52. Abrol, R.; Kam, V. W. T.; Jenelle, B.; Wienko, H.; Goddard, W. A. unpublished.
53. Grigoryan, G.; Ochoa, A.; Keating, A. E. Proteins-Structure Function and Bioinformatics 2007, 68(4), 863-878.
54. Lindahl, E.; Hess, B.; van der Spoel, D. J Mol Model 2001, 7(8), 306-317.
55. Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. J Am Chem Soc 1996, 118(45), 11225-11236.
56. Park, S.; Park, S. Current opinion in structural biology 2004, 14(4), 487-494.
57. Patel, S.; Brooks, C. L. Journal of Computational Chemistry 2004, 25(1), 1-15.
58. Yu, H. B.; Hansson, T.; van Gunsteren, W. F. J Chem Phys 2003, 118(1), 221-234.
59. Aqvist, J. Journal of Computational Chemistry 1996, 17(14), 1587-1597.
60. Wang, T.; Wade, R. C. Proteins-Structure Function and Genetics 2003, 50(1), 158-169.
61. Ferrara, P.; Apostolakis, J.; Caflisch, A. Proteins-Structure Function and Genetics 2002, 46(1), 24-33.
62. Lazaridis, T.; Karplus, M. Proteins-Structure Function and Genetics 1999, 35(2), 133-152.
63. Lazaridis, T.; Archontis, G.; Karplus, M. In Advances in Protein Chemistry, Vol 47, 1995, p 231-306.
64. Bone, S.; Pethig, R. Journal of Molecular Biology 1985, 181(2), 323-326.
65. Dunning, T. H. Journal of Physical Chemistry A 2000, 104(40), 9062-9080.
66. Mahoney, M. W.; Jorgensen, W. L. J Chem Phys 2000, 112(20), 8910-8922.
67. Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Journal of Computational Chemistry 2004, 25(9), 1157-1174.
68. Buehler, M. J.; J., D.; van Duin, A. C. T.; Meulbroek, P.; Goddard, W. A. Mat Res Soc Proceedings (Combinatorial Methods and Informatics in Materials Science) 2006, 894.
69. Phillips, J. 2005, - 26(- 16), - 1802.
70. Kalani, M. Y. S.; Vaidehi, N.; Hall, S. E.; Trabanino, R. J.; Freddolino, P. L.; Kalani, M. A.; Floriano, W. B.; Kam, V. W. T.; Goddard, W. A. Proceedings of the National Academy of Sciences of the United States of America 2004, 101(11), 3815-3820.
71. Freddolino, P. L.; Kalani, M. Y. S.; Vaidehi, N.; Floriano, W. B.; Hall, S. E.; Trabanino, R. J.; Kam, V. W. T.; Goddard, W. A. Proceedings of the National Academy of Sciences of the United States of America 2004, 101(9), 2736-2741.
72. Lippow, S. M.; Tidor, B. Current Opinion in Biotechnology 2007, 18(4), 305-311.
73. Lim, K. T.; Brunett, S.; Iotov, M.; McClurg, R. B.; Vaidehi, N.; Dasgupta, S.; Taylor, S.; Goddard, W. A. Journal of Computational Chemistry 1997, 18(4), 501-521.
74. Wolfenden, R.; Andersson, L.; Cullis, P. M.; Southgate, C. C. B. Biochemistry 1981, 20(4), 849-855.
75. Nicholls, A. 1991, - 12(- 4), - 445.
76. Link, A. J.; Mock, M. L.; Tirrell, D. A. Current Opinion in Biotechnology 2003, 14(6), 603-609.
77. Kirshenbaum, K.; Carrico, I. S.; Tirrell, D. A. Chembiochem 2002, 3(2-3), 235-237.

78. van Hest, J. C. M.; Tirrell, D. A. Chemical Communications 2001(19), 1897-1904.
79. Ibba, M.; Soll, D. Annu Rev Biochem 2000, 69, 617-650.
80. Woese, C. R.; Olsen, G. J.; Ibba, M.; Soll, D. Microbiology and Molecular Biology Reviews 2000, 64(1), 202-+.
81. Wang, L.; Schultz, P. G. Angewandte Chemie-International Edition 2005, 44(1), 34-66.
82. Chin, J. W.; Cropp, T. A.; Anderson, J. C.; Mukherji, M.; Zhang, Z. W.; Schultz, P. G. Science 2003, 301(5635), 964-967.
83. Datta, D.; Wang, P.; Carrico, I. S.; Mayo, S. L.; Tirrell, D. A. J Am Chem Soc 2002, 124(20), 5652-5653.
84. Wang, P.; Vaidehi, N.; Tirrell, D. A.; Goddard, W. A. J Am Chem Soc 2002, 124(48), 14442-14449.
85. Himo, F.; Lovell, T.; Hilgraf, R.; Rostovtsev, V. V.; Noodleman, L.; Sharpless, K. B.; Fokin, V. V. J Am Chem Soc 2005, 127(1), 210-216.
86. Kolb, H. C.; Finn, M. G.; Sharpless, K. B. Angewandte Chemie-International Edition 2001, 40(11), 2004-+.
87. Huisgen, R. Proceedings of the Chemical Society of London 1961, 357.
88. Retailleau, P.; Huang, X.; Yin, Y. H.; Hu, M.; Weinreb, V.; Vachette, P.; Vonrhein, C.; Bricogne, G.; Roversi, P.; Ilyin, V.; Carter, C. W. Journal of Molecular Biology 2003, 325(1), 39-63.
89. Cormen, T. H.; Leiserson, C. E.; Rivest, R. L.; Stein, C. Introduction to Algorithms, 2001.