A Novel High Resolution C^{α}-C^{α} Distance Dependent Force Field Based On A High Quality Decoy Set

R. Rajgaria, S. R. McAllister, and C. A. Floudas*
Department of Chemical Engineering,
Princeton University,
Princeton, NJ 08544-5263, U.S.A.

June 27, 2006

Abstract

This work presents a novel C^{α}-C^{α} distance dependent force field which is successful in selecting native structures from an ensemble of high resolution near-native conformers. An enhanced and diverse protein set, along with an improved decoy generation technique, contributes to the effectiveness of this potential. High quality decoys (structures with low root mean square deviation with respect to the native; see Tables V-VIII) were generated for 1489 non-homologous proteins and used to train an optimization based linear programming formulation. The goal in developing a set of high resolution decoys was to develop a simple, distance-dependent force field that yields the

*Author to whom all correspondence should be addressed; Tel: +1-609-258-4595; Fax: +1-609-258-0211. *E-mail*: floudas@titan.princeton.edu

native structure as the lowest energy structure and assigns higher energies to decoy structures that are quite similar as well as those that are less similar. The model also includes a set of physical constraints that were based on experimentally observed physical behavior of the amino acids. The force field was tested on two sets of test decoys not in the training set and was found to excel on all the metrics that are widely used to measure the effectiveness of a force field. The high resolution (HR) force field was successful in correctly identifying 113 native structures out of 150 test cases and the average rank obtained for this test was 1.87. All the high resolution structures (training and testing) used for this work are available online and can be downloaded from http://titan.princeton.edu/HRDecoys.

Keywords: force field; potential model; high resolution decoys; protein structure prediction; linear optimization; protein design potential.

1 Introduction

Proteins are the most structurally advanced molecules known. Predicting the structure of these complex molecules is one of the most interesting and difficult problems of computational biology. The basic energetic model commonly used to solve this problem is based on Anfinsen's hypothesis[1], which says that for a given physiological set of conditions the native structure of a protein corresponds to the global Gibbs free energy minimum. Various components of the protein folding problem (e.g., fold recognition, ab initio prediction, comparative modeling and de novo design) make use of some kind of energy function for estimating the energy of native and non-native conformers. These energy functions are also referred as force fields.

A good force field should be able to distinguish between the native

and non-native conformers of a protein based on its energy estimates. Most generally, these potentials or force fields can be divided into two categories. The first class is the physics-based potential and the second class is the knowledge-based potential. An ideal physics-based force field should consider all types of interactions (for example, van der Waals interactions, hydrogen bonding, electrostatic interactions etc.) occurring between its atoms at the atomic level. This type of force field can be obtained by applying basic laws of physics and chemistry at the atomic level of a protein. Some of the well established force fields in this category are CHARMM[2], AMBER[3], ECEPP[4], ECEPP/3[5] and GROMOS[6]. It has been pointed out that even these types of potentials are sometimes not effective in capturing the correct energetics of a protein[7,8]. Hence a lot of effort has been invested in finding a simplified protein potential which is capable of differentiating native and non-native proteins without heavily increasing the computational load.

Knowledge-based potentials, as evident from their name, use information from the experimentally determined protein structures in the Protein Data Bank[9] to come up with the interaction energy parameters. Different approaches have been used to derive these potentials. Tanaka et al.[10], Finkelstein et al.[11], and Bryant and Lawrence[12] used the Boltzmann distribution to calculate knowledge-based force fields. The choice of the reference state used in these calculations was reviewed by Jernigan and Bahar[13]. Scheraga and coworkers developed a united residue representation (UNRES) of a polypeptide chain[14–17]. All atom force fields have been developed by several research groups[18–21]. Lu and Skolnick[18] developed a heavy atom distance dependent force field, increasing the number of residue centers from 20 (C^{α} based approach) to 167 (heavy atom approach). Samudrala and Moult[19] used an all atom based conditional probability approach for the force field estimation. Some of the other successful potentials are LKF[22], TE13[23], and

HL[24]. LKF and TE13 are distance dependent force fields, whereas HL is a contact based potential. A comprehensive, recent review on such potentials can be found in Floudas et al.[25].

As the efficacy of protein structure prediction tools increase[26], we need to move from low and medium resolution structure prediction to high resolution structure prediction. This prediction requires the ability to distinguish between very similar structures with low root mean square deviations (rmsds). The problem of high resolution structure prediction has recently received attention[27, 28]. The current work aims to address this problem by developing a high resolution energy function with the use of optimization based techniques.

This work presents a novel C^{α}-C^{α} distance dependent high resolution force field. The emphasis is on the high resolution, which would enable us to differentiate between native and non-native structures that are very similar to each other (rmsd $<$ 2 Å). The force field is called high resolution because it has been trained on a large set of high resolution decoys (small rmsd with respect to the native) and it intends to effectively distinguish high resolution decoys structures from the native structure. The basic framework used in this work is similar to the one developed by Loose et al.[22]. However, it has been improved and applied to a diverse and enhanced (both in terms of quantity and quantity) set of high resolution decoys. The new proposed model has resulted in remarkable improvements over the LKF potential. These high resolution decoys were generated using torsion angle dynamics in combination with restricted variations of the hydrophobic core within the native structure. This decoy set highly improves the quality of training and testing. The force field developed in this paper was tested by comparing the energy of the native fold to the energies of decoy structures for proteins separate from those used to train the model. Other leading force fields were also tested on this

high quality decoy set and the results were compared with the results of our high resolution potential. The comparison is presented in the Results section.

2 Theory and Modeling

In this model, amino acids are represented by the location of its C^{α} atom on the amino acid backbone. The conformation of a protein is represented by a coordinate vector, X, which includes the location of the C^{α} atoms of each amino acid. The native conformation is denoted as X_n, while the set $i = 1, \ldots, N$ is used to denote the decoy conformations X_i. Non-native decoys are generated for each of $p = 1, \ldots, P$ proteins and the energy of the native fold for each protein is forced to be lower than those of the decoy conformations (Anfinsen's hypothesis). This constraint is shown in the following equation:

$$E(X_{p,i}) - E(X_{p,n}) \; > \; \varepsilon \qquad p = 1, ..., P \quad i = 1, ..., N \tag{1}$$

Equation 1 requires the native conformer to be always lower in energy than its decoy. A small positive parameter ϵ is used to avoid the trivial solution in which all energies are set to zero. An additional constraint (Equation 2), is used to produce a nontrivial solution by constraining the sum of the differences in energies between decoy and native folds to be greater than a positive constant[29]. For the model presented in this paper, the values of ϵ and Γ were set to 0.01 and 1000, respectively.

$$\sum_{p} \sum_{i} [E(X_{p,i}) - E(X_{p,n})] \; > \; \Gamma \tag{2}$$

The energy of each conformation is taken as the arithmetic sum of pairwise interactions corresponding to each amino acid combination at

Table I: Distance dependent bin definition[22].

Bin ID	C^{α} Distance [Å]
1	3-4
2	4-5
3	5-5.5
4	5.5-6
5	6-6.5
6	6.5-7
7	7-8
8	8-9

a particular "contact" distance. A contact exists when the C^{α} carbons of two amino acids are within 9 Å of each other. So, the energy of each interaction is a function of the C^{α}-C^{α} distances and the identity of the interacting amino acids. To formulate the model, the energy of an interaction between a pair of amino acids, IC, within a distance bin, ID, was defined as $\theta_{IC,ID}$. The eight distance bins defined for the formulation are shown in Table I. The energy for any fold X, of decoy i, for a protein p, is given by Equation 3.

$$E(X_{p,i}) \; = \; \sum_{IC} \sum_{ID} \; N_{p,i,IC,ID} \; \theta_{IC,ID} \tag{3}$$

In this equation, $N_{p,i,IC,ID}$ is the number of interactions between an amino acid pair IC, at a C^{α}-C^{α} distance ID. The set IC ranges from 1 to 210 to account for the 210 unique combinations of the 20 naturally occurring amino acids. These bin definitions yield a total of 1680 interaction parameters to be determined by this model. To determine these parameters, a linear programming formulation is used

in which the energy of a native protein is compared with a large number of its decoys. The violations, in which a non-native fold has a lower energy than the natural conformation, are minimized by optimizing with respect to these interaction parameters.

Equation 1 can be rewritten in terms of $N_{p,i,IC,ID}$ as Equation 4, where the slack parameters, S_p, are positive variables (Equation 5) that represent the difference between the energies of the decoys and the native conformation of a given protein.

$$\sum_{IC} \sum_{ID} [N_{p,i,IC,ID} - N_{p,n,IC,ID}]\theta_{IC,ID} + S_p \ \geq \varepsilon \qquad (4)$$
$$p \ = \ 1, \ldots, P \quad i \ = \ 1, \ldots, N$$

$$S_p \geq \ 0 \quad p = \ 1, \ldots, P \qquad (5)$$

$$\min_{\theta(IC,ID)} \sum_{p} \ S_p \qquad (6)$$

The objective function for this formulation is to minimize the sum of the slack variables, S_p, written in the form of Equation 6. The relative magnitude of $\theta_{IC,ID}$ is meaningless because if all $\theta_{IC,ID}$ parameters are multiplied by a common factor then Equations 4 and 5 are still valid. In this formulation, $\theta_{IC,ID}$ values were bound between -25 and 25.

2.1 Physical Constraints

The above mentioned equations constitute the basic constraints needed to solve this model. However, this set does not guarantee a physically realistic solution. It is possible to come up with a set of parameters that can satisfy Equations 2-6 but would not reflect the actual interaction occurring between amino acids in a real system. To prohibit these

unrealistic cases, another set of constraints based on the physical properties of the amino acids was imposed. Statistical results presented in Bahar and Jernigan[30] were also incorporated through the introduction of hydrophilic and hydrophobic constraints.

2.1.1 General Constraints

This class of constraints was used to produce a smooth energy profile[23]. It is expected that when the distance changes from one bin to the next bin, the energy profile would change smoothly and it would not exhibit random jumps. In order to enforce this behavior, the difference in energy between two neighboring distance bins was limited to 8 units for the first two bins, and 4 units thereafter, as shown in Equations 7-10.

$$\theta_{IC,ID+1} - \theta_{IC,ID} \geq -8, \quad \forall IC; ID = 1 \tag{7}$$

$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq 8, \quad \forall IC; ID = 1 \tag{8}$$

$$\theta_{IC,ID+1} - \theta_{IC,ID} \geq -4, \quad \forall IC; ID = 2, 3, ...7 \tag{9}$$

$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq 4, \quad \forall IC; ID = 2, 3, ...7 \tag{10}$$

It is also intuitive that the effectiveness of interactions should decline at long distances, as force scales with the inverse of distance squared. This constraint was enforced by Equations 11-13.

$$\theta_{IC,ID} \leq 5, \quad \forall IC; ID = 7 \tag{11}$$

$$\theta_{IC,ID} \geq -4, \quad \forall IC; ID = 8 \tag{12}$$

$$\theta_{IC,ID} \leq 4, \quad \forall IC; ID = 8 \tag{13}$$

2.1.2 Hydrophobic-Hydrophobic Constraints

Hydrophobic-hydrophobic constraints were formulated to capture the specific interaction between certain types of amino acids. Amino acids can be classified as hydrophobic or hydrophillic, charged or uncharged. The classification used for this formulation is given in Table II[22].

The behavior of different classes of amino acids were studied by Bahar and Jernigan[30]. They established that the hydrophobic groups show favorable interactions at a distance of 4-6.5 Å. Also, these types of interactions tend to show an "energy well" at around 4.5 to 5.0 Å. These results are incorporated using Equations 14-18.

$$\theta_{IC,ID} \leq 0, \quad IC \in \{H,H\}; ID = 2,3,4,5 \tag{14}$$

$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq -4, \quad IC \in \{H,H\}; ID = 2 \tag{15}$$

$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq -2, \quad IC \in \{H,H\}; ID = 2,3 \tag{16}$$

$$\theta_{IC,ID+2} - \theta_{IC,ID} \geq 0, \quad IC \in \{H,H\}; ID = 4 \tag{17}$$

$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq 2, \quad IC \in \{H,H\}; ID = 4 \tag{18}$$

Alanine (ALA) shows a different kind of interaction with hydrophobic groups. Due to the small methyl side chain, it was observed that the steric effects were less dominant and alanine showed favorable interaction with hydrophobic residues at distances shorter than 4 Å. The interactions were still forced to be negative in the 4-6.5 Å range, but energy profiles were forced to increase rather than form energy wells based on previous studies[30] and the characteristics of alanine. These constraints are shown in Equations 19-22.

$$\theta_{IC,ID} \leq 0, \quad IC \in \{O,H\}; ID = 2,3,4 \tag{19}$$

$$\theta_{IC,ID+1} - \theta_{IC,ID} \geq 2, \quad IC \in \{O, H\}; ID = 2, 3 \tag{20}$$

$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq 4, \quad IC \in \{O, H\}; ID = 2 \tag{21}$$

$$\theta_{IC,ID+1} - \theta_{IC,ID} \geq 1, \quad IC \in \{O, H\}; ID = 2, 3 \tag{22}$$

Phenylalanine (PHE) residue interactions show some additional properties. Interactions between phenylalanine and other aromatic residues tend to remain favorable even at a longer distance. This may be due to their larger size, which would allow stronger interactions at distances greater than 6 Å when compared with smaller groups. These results were incorporated by using Equations 23-24.

$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq 0, \quad IC \in \{PHE, HA\}; ID = 4 \tag{23}$$

$$\theta_{IC,ID+2} - \theta_{IC,ID} \geq 0, \quad IC \in \{PHE, HA\}; ID = 5 \tag{24}$$

2.1.3 Charged Group Constraints

Charged group constraints are applied to charged amino acids. From the basic laws of physics, it is expected that a contact between two amino acids with the same charge should be very unfavorable at short distances, and become less unfavorable at longer distances. The opposite effect is expected for oppositely charged amino acids. This observation is written in form of Equations 25-28.

$$\theta_{IC,ID} \geq 0, \quad IC \in \{\{PP, PP\}, \{PN, PN\}\}; \forall ID \tag{25}$$

$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq -1, \quad IC \in \{\{PP, PP\}, \{PN, PN\}\}; ID = 2, 3, 4, 5 \tag{26}$$

$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq 0, \quad IC \in \{\{PP, PP\}, \{PN, PN\}\}; ID = 6, 7 \tag{27}$$

$$\theta_{IC,ID+1} \leq 0, \quad IC \in \{PP, PN\}; \quad ID = 1, 2, 3, 4 \tag{28}$$

It has also been observed that a histidine residue (HIS) shows favorable interactions with all groups, except other positively charged groups, because of its unique ionization properties[30]. This observation in written in form of Equations 29-30.

$$\theta_{IC,ID+1} - \theta_{IC,ID} \geq 1, \quad IC \in \{HIS, HIS \cup PP\}; ID = 2, 3 \tag{29}$$

$$\theta_{IC,ID} \leq 0, \quad IC \in \{HIS, PU \cup PN \cup HN \cup HA \cup O\}; \quad ID = 2, 3, 4 \tag{30}$$

2.1.4 Hydrophilic Group Constraints

Bahar and Jernigan[30] have also shown that hydrophilic groups exhibit very favorable interactions at a distance below 4 Å and this interaction decays as the distance increases. This finding was incorporated through Equation 31.

$$\theta_{IC,ID+1} - \theta_{IC,ID} \geq 4, \quad IC \in \{\{PU, PU \cup PP \cup PN\}, \{PP, PN\}\}; \quad ID = 1 \tag{31}$$

2.1.5 Hydrophilic-Hydrophobic Constraints

Hydrophilic-hydrophobic constraints were written to restrict the strength of interactions between certain types of amino acids. For example, based on the result of Bahar and Jernigan[30], it is not natural to expect the favorable interaction between two hydrophilic groups to be as strong as interactions between two hydrophobic groups or oppositely charged groups at distances longer than 4 Å. Also, no interactions

are expected to be as unlikely as those between two groups with the same charge at small distances. These results are incorporated through Equations 32-34.

$$\begin{aligned}\theta_{IC,ID} \geq \quad -6, \quad & IC \in \{\{PU, PU \cup PP \cup PN \cup HN \cup HA\}, \\ & \{PP, PP \cup HN \cup HA\}, \\ & \{PN, PN \cup HN \cup HA\}\}; ID = 2 \end{aligned} \tag{32}$$

$$\begin{aligned}\theta_{IC,ID} \geq \quad -4, \quad & IC \in \{\{PU, PU \cup PP \cup PN \cup HN \cup HA\}, \\ & \{PP, PP \cup HN \cup HA\}, \\ & \{PN, PN \cup HN \cup HA\}\}; ID = 3, \ldots, 8 \end{aligned} \tag{33}$$

$$\begin{aligned}\theta_{IC,ID} \leq \quad 10, \quad & IC \in \{\{PU \cup HN \cup HA, PU \cup PP \cup PN \cup HN \cup HA\}, \\ & \{PP, PN\}\}; ID = 2, 3, 4, 5, 6 \end{aligned} \tag{34}$$

2.1.6 Miscellaneous Constraints

Amino acids were grouped based on the work of Bahar and Jernigan[30], who developed two sets of hydrophobic ($H1$, $H2$) and two sets of hydrophilic groups ($P1$,$P2$), as shown in Table III[22]. Based on their work, interactions between a residue from $H1$ with another residue from $H1$ were required to be stronger than interactions with a residue from $H2$ and stronger than any interaction within $H2$. Additionally, interactions between a residue from $H1$ and a residue from $P1$ were forced to be stronger than an interaction with a residue from $P2$. These constraints are written in form of Equations 35-37.

$$\theta(H1, H1) < \theta(H1, H2) \tag{35}$$

$$\theta(H1, H1) < \theta(H2, H2) \tag{36}$$

$$\theta(H1, P1) < \theta(H1, P2) \tag{37}$$

Some additional constraints were incorporated using the results of their work. These constraints are shown in Equations 38-43.

$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq 0$$
$$IC \in \{PHE, HIS \cup ASN \cup GLU \cup ARG\}; ID = 2,3 \quad (38)$$
$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq -2$$
$$IC \in \{ILE, HN \cup HA\}; ID = 2,3 \quad (39)$$

$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq -2$$
$$IC \in \{LEU, GLN \cup ASN\}; ID = 2 \quad (40)$$
$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq -2$$
$$IC \in \{ASP, ILE \cup LEU \cup VAL \cup HA\}; ID = 2,3 \quad (41)$$
$$\theta_{IC,ID} \geq 0$$
$$IC \in \{ASP, ILE \cup LEU \cup VAL \cup HA\}; ID = 2,3 \quad (42)$$
$$\theta_{IC,ID} \leq 0$$
$$IC \in \{ASN, TYR \cup TRP\}; ID = 2,3,4 \quad (43)$$

The additional constraints of Equations 7-43, combined with the base model of Equations 4-6, complete the mathematical model of this formulation, a linear programming problem. Problems of this type can be readily solved using commercial solvers (e.g., CPLEX[31], Xpress[32]). The next section describes the method and approach used for high quality decoy generation.

2.2 Database Selection and Decoy Generation

Many advances have been made in the prediction of medium-resolution structures, both using discrete distance-dependent force fields as well as continuous, physically-based atomistic level force fields. Some of these force fields have done quite well distinguishing the native conformation of a protein from over thousands of its non-native conformers.

However, the important and challenging area of work is the prediction of high-resolution protein structures. The ultimate goal is to move the prediction barrier from low and medium resolution to high resolution. This calls for improvements in two areas: high quality decoy generation and enhanced training techniques.

The protein database selection is critical to force field training. The protein set should not be too large, as the training becomes computationally expensive and difficult with an increase in the size of the training set. At the same time, it should be large enough to adequately represent the PDB set. Previous work by Loose et al.[22] involved data set selection using PDBselect with protein lengths less than 150 amino acids. Tobi and Elber[23] used a training set of 572 proteins. A set of 1489 proteins developed by Zhang and Skolnick[33] was used for this work. This set has many distinct advantages over sets used by others. The length of these proteins varied from 41 to 200 amino acids, compared to a maximum of 150 in the previous work[22]. This improvement adds to the functional diversity of the protein set. Zhang and Skolnick[33] reported that this set contains 1,489 nonhomologous single domain proteins. The maximum pairwise sequence similarity reported for this set was 35 %. This set is also a well represented combination of α, β, and α/β proteins (448, 434 and 550 respectively). This set was used to generate decoys and then divided into a training set and a test set.

The decoy generation procedure was based on the hypothesis that high-quality decoy structures should preserve information about the distances within the hydrophobic core of the native structure of each protein. For this study, the hydrophobic core is defined as all residues within a β-strand and all hydrophobic residues within an α-helix. For native protein structures with little secondary structure (less than 25% of the amino acids within secondary structure elements) or less than

two secondary structure elements, the hydrophobic residues within the protein loops are considered part of the hydrophobic core as well.

Once the hydrophobic core for each protein is established, a number of distance constraints are introduced based on the hydrophobic-hydrophobic distances within the native structure. The proximity of the decoy structures can be controlled by varying the amount, which we call a slack value, that each of these pairwise distances is allowed to vary. Table IV shows the eight slack values used for the eight different sets of distance bounds. Subsequently, each set of distance bounds is used as input to a torsion angle dynamics program to establish a large number of protein decoy structures that satisfy the bounds. A program developed for NMR structure refinement, DYANA, is used to generate 200 structures for each slack value[34]. The DYANA run for a given set of distance constraints requires between 10 minutes and 2 hours on single 3.0 GHz Intel Xeon processor. A Beowulf cluster containing 80 nodes of dual 3.0 GHz Intel Xeon processors was used to serially distribute the work of these 8 runs for each of the 1489 proteins in the decoy set.

The selection of proteins for use in the training and testing sets is then based upon the minimum root mean squared deviation decoy structure. Tables V-VIII illustrate the distribution of minimum, maximum, median and mean rmsd of the decoy structure values across the entire set of proteins studied. Any protein with a minimum rmsd decoy structure of more than 2.0 Å is discarded, as it is incompatible with the goal of developing a force field to distinguish between high-resolution decoys and the native protein structure. For similar reasons, any individual decoy structure with an rmsd of more than 8.0 Å to the native structure is also discarded. The flowchart used for decoy generation is shown in Figure 1. In its final form, the high-resolution decoy set contains 1400 protein structures, with between 500 and 1600 decoy

structures for each protein. The entire set of protein decoy structures has been made available at http://titan.princeton.edu/HRDecoys/.

2.3 Training Set

Of the 1400 proteins used for decoy generation, 1250 were randomly selected for training and the rest were used for testing purposes. For every protein in the set, 500-1600 decoys were generated depending on the fraction of secondary structure present in the native structure of the protein (see Section 2.2). Table V shows the number of proteins in the training and testing set for each rmsd range. These decoys were sorted based on their C^{α} rmsd to the native structure and then 500 decoys were randomly selected to represent the whole rmsd range. This creates a training set of $500\times\ 1250 = 625{,}000$ decoys. However, because of computer memory limitations, it is not possible to include all of these decoys at the same time for training. Only 60,000 structures could be used at a time to solve the model. This memory problem has been previously addressed by Loose et al.[22] using the maximum feasibility heuristics[35]. A similar iterative scheme, "Rank and Drop", was employed to overcome the memory problem while effectively using all the high quality structures.

In the Rank and Drop scheme, a basic force field (FF_0) was developed using a subset of available decoys. All 500 decoys for each protein were ranked according to their C^{α} rmsds. Of these 500 decoys, the top 45 (lowest rmsd) were selected for each training set protein. These $45 \times 1250 = 56250$ ($< 60{,}000$) structures were then used to train the LP model and a force field FF_0 was developed. This force field was then used to rank all 500 decoys.Ranking of these 500 decoys would depend on the difference in the energetic landscape of the native and the non-native conformer. Equation 4 determines the slack value (difference in the energy of a native and non-native structure) for each of

these decoys. In general, a lower slack value would mean that there were fewer constraint violations and hence the decoy is a better and challenging structure. A high value of slack would mean there were lot of constraint violations and the structure is very different from the native structure of the protein.

It is of crucial importance to start off with a good FF_0 force field, as this force field further dictates the selection of decoys that are used for the next round of optimization. We used the top 45 structures (lowest rmsd) in the generation of FF_0 force field, as these were the most challenging structures in the set of 500.

After obtaining the rank ordered list of the 500 decoys, (based on their slack values) the top 45 decoys with the lowest slack values were selected while keeping a fraction of the decoys used in the previous iteration. The set of these 45 decoys for each of the 1250 training proteins defines the new training set. This set was further used and a new force field was developed. This process of force field generation and decoy ranking was repeated until there was no change or improvement in the ranking of the decoys. The final force field obtained by this iterative process was called the High-Resolution (HR) force field.

This force field model was solved using the GAMS modeling language coupled with the CPLEX linear programming package[31]. These calculations were performed on an Intel Pentium-4, 3.2 GHz workstation with 4 gigabytes of RAM.

2.4 Test Set

It is equally important to test a force field on a difficult and rigorous testing set to confirm its effectiveness. A number of interesting criteria that decide the severity of the tests have been pointed out by Park et al.[36]. They claim that the quality of a test set depends on factors like the structural proximity of the decoy with the native structure and

the diversity of the test set. These goals have been prioritized while designing the test set for this high resolution force field.

The test set was comprised of 150 randomly selected proteins (41-200 amino acids in length). For each of the 150 test proteins, 500 high resolution decoys were generated using the same technique that was used to generate training decoys. The minimum C^{α} based rmsds for these non-native structures were in the range of 0-2 Å (Table V). This range establishes the structural proximity of these decoys with their native counterparts. Since this work aims to address high resolution structure prediction, decoys with rmsd more than 5 Å were discarded from the test set.

This HR force field was also tested on another set of medium resolution decoys[22]. This set has 200 decoys for 151 proteins. The minimum RMSD of the decoys of this set ranged from 3-16 Å. This set, along with the high resolution decoy set, spans the practical range of possible protein structures that one might encounter during protein structure prediction.

3 Results and Discussion

A linear optimization problem was solved using information from 625,000 decoy structures and the values of all the energy parameters were obtained. The objective function of this formulation was to minimize the sum of the slack variables. A value of zero for the objective function would mean that there were no violations in which the non-native conformer had a lower energy than the corresponding native structure. However, a non-zero value for the objective function was obtained for this case. For 278 proteins (out of 1250 proteins), at least one constraint was violated. A similar fraction of violations was found for the test sets, indicating that the model was not overtrained. It is impor-

tant to realize that the problem at hand requires distinction between structures with an average C^{α} rmsd of 1-2 Å. It is difficult to find a set of parameters that would satisfy each and every inequality of this formulation for approximately 60,000 conformers in each run. Thus, a non-zero objective function (violation of Equation 1) does not reflect poorly on the efficacy of the HR force field.

The ability to distinguish between the native structure and native-like conformers is the most significant test for any force field. The HR force field was tested on 500 decoys of the 150 test proteins. In this testing, the relative position, or rank, of the native conformation among its decoys was calculated. An ideal force field should be able to assign rank 1 to the native structures of all the test proteins. It should be noted that the test set should not overlap with the training set as that would invalidate the force field assessment. This consideration was carefully incorporated in our test set and there was no overlap between the training and test set. The results of this testing are presented in Table IX. Other force fields like LKF[22], TE13[23], and HL[24] were also tested on this set of high resolution decoys. All these force fields are fundamentally different from each other in their methods of energy estimation. The LKF force field is a C^{α}-C^{α} distance dependent potential where the interaction distance range 3-9 Å is divided into 8 bins. The TE13 force field is also a distance dependent (13 bin) force field, but the the interaction distance is measured between the geometric centers of the side chain of two interacting residues. The HL force field is a simplified contact based force field, where a pair of amino acids contact when a non-hydrogen atom of a residue approaches within 4.5 Å of a non-hydrogen atom of another residue which is at least five residues apart from each other. Comparing the results obtained with these force fields aims to assess the fundamental utility of the HR force field. The comparison of the energy rankings obtained using different

force fields is presented in Table XI. Two of the high resolution test cases did not have the side chain coordinate information in the native files, so the TE13 force field was tested only on 148 test cases.

Table IX demonstrates that the HR force field is the most effective in identifying the native structures by rank. Assigning rank 1 to the native structure means that the force field is adept at finding the native structure from an array of its non-native configurations. The HR force field correctly identified the native folds of 113 proteins out of a set of 150 proteins, which compares favorably to a maximum of 92 (out of 148) by the TE13 force field.

Another analysis was carried out to evaluate the discrimination ability of these potentials. In this evaluation, all the decoys of the test set were ranked using these potentials. For each test protein, the C^{α} rmsd of the rank 1 conformer was calculated with respect to the native structure of that protein. The C^{α} rmsd would be zero for the cases in which a force field selects the native structure as rank 1. However, it will not be zero for all other cases in which a non-native conformer is assigned the top rank. The average of these rmsds represents the spatial separation of the decoys with respect to the native structure. The average rmsd value obtained for each of the force fields is shown in Table XI. It can be seen that the average C^{α} rmsd value is least for the HR force field. The average C^{α} rmsd value for the HR force field is 0.451 Å, which is much less compared to 1.721 Å by the LKF, and 0.813 Å by TE13 force field. This means that the structures predicted by the HR force fields have the least spatial deviation from their corresponding native structures.

The HR force field was also tested on the test set published by Loose et al.[22]. This is a medium resolution decoy set with the minimum C^{α} rmsd of the decoys varying in the range of 3-16 Å. This set has 200 (199 non-native and 1 native) structures for each of the 150 proteins. There

were 40 common proteins between this test set and the training set used in HR force field generation. These 40 common proteins were removed and the HR force field was tested only on the 110 non-homologous proteins. These test results are presented in Tables X. The TE13 and LKF potentials were also tested on this set and the results have been published in Loose et al.[22].

The summary of all the testing results, both on the high resolution decoys and on the medium resolution decoys have been comprehensively presented in Tables XI and XII. The proposed HR force field selected 113 native structures out of 150 test proteins, a very high success rate. Also, for the remaining 37 cases (in which it could not select the native as rank 1) it assigned a very high rank (<10, except 12 for 1g10A and 23 for 1g9pA) to the native structure, giving rise to a very low average rank and outperforming the other force fields. The effectiveness of the HR force field is further reinforced by its success on the medium resolution decoy test set. On the test set of 110 medium resolution decoys, it was capable of correctly identifying 78.2 % of the native structures, significantly more than other force fields.

The correlation between the energy and the rmsd was also calculated for all the high resolution test set proteins. An average correlation of R=0.80 was found for these test cases. This is important as a high energy-rmsd correlation suggests the usefulness of the HR potential to guide structure prediction from high rmsd regions to low rmsd regions. Figure 2 shows the energy-rmsd correlation for 4 test cases. Similar plots were generated for all 150 high resolution test cases and the histogram distribution of R (correlation coefficient) for all these cases is given in Figure 3.

The value of the interaction parameters, $\theta_{IC,ID}$, comprising the HR force field are given in Appendix A.

4 Conclusions

A new high resolution C^{α}-C^{α} distance dependent force field has been developed to address the problem of high resolution protein structure prediction. The force field was developed using an optimization based linear programming formulation, in which the model is trained using a diverse set of high quality decoys. The decoys were generated based on the premise that high quality decoy structures should preserve information about the distance within the hydrophobic core of the native structure of each protein. The set of interaction energy parameters obtained after solving the model were found to be of very good discriminatory capacity. This force field performed well on a set of independent, non-homologous high resolution decoys. It also showed good predictive capability when tested on a different medium resolution decoy set, while outperforming other leading force fields. This force field can become a powerful tool for fold recognition and de novo protein design. Further studies involve extending the C^{α} based approach to include the effect of the presence of amino acid side chains and evaluating the performance of these force fields on other decoys sets.

Acknowledgements

CAF gratefully acknowledges financial support from both the National Science Foundation and the National Institutes of Health (R01 GM52032; R24 GM069736). CAF also thanks Professor Skolinck and Professor Zhang for providing the large protein set and decoys.

References

[1] Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181.

[2] MacKerell Jr AD, Bashford D, Bellott M, Dunbrack Jr RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher III WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiórkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. Journal of Physical Chemistry B 1998;102:3586–3616.

[3] Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz Jr KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. Journal of the American Chemical Society 1995;117:5179–5197.

[4] Momany FA, McGuire RF, Burgess AW, Scheraga HA. Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. Journal of Physical Chemistry 1975;79:2361–2381.

[5] Némethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. Journal of Physical Chemistry 1992; 96:6472–6484.

[6] Scott WRP, Hunenberger PH, Trioni IG, Mark AE, Billeter SR, Fennen J, Torda AE, Huber T, Kruger P, VanGunsteren WF. The

GROMOS biomolecular simulation program package. Journal of Physical Chemistry A 1997;103:3596–3607.

[7] Novotny J, Rashin AA, Bruccoleri RE. Criteria that discriminate between native proteins and incorrectly folded models. Proteins: Structure, Function, and Bioinformatics 1984;177:788–818.

[8] Wang Y, Zhang H, Li W, Scott RA. Discriminating compact nonnative structures from the native structures of globular proteins. Proceedings of the National Academy of Sciences of the United States of America 1995;92:709–713.

[9] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Research 2000;.

[10] Tanaka S, Scheraga HA. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules 1976;9:945–950.

[11] Finkelstein AV, Badretdinov AY, Gutin AM. Why do proteins architectures have Boltzmann-like statistics? Proteins: Structure, Function, and Bioinformatics 1995;23:142–150.

[12] Bryant SH, Lawrence CE. The frequency of ion-pair substructures in proteins is quantitaively related to electrostatic potential. a statistical model for nonbonded interactions. Proteins: Structure, Function, and Bioinformatics 1991;9:108–119.

[13] Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. Current Opinon in Structural Biology 1996;6:195–209.

[14] Liwo A, Oldziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. A united-residue force field for off-lattice protein structure simulations. I. Functional forms and parameters of long-range

side-chain interaction potentials from protein crystal data. Journal of Computational Chemistry 1997a;18:849–873.

[15] Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Oldziej S, Scheraga HA. A united-residue force field for off-lattice protein structure simulations. II. Parameterization of short-range interactions and determination of weights of energy terms by z-score optimization. Journal of Computational Chemistry 1997b;18:874–887.

[16] Liwo A, Kazmierkiewicz R, Czaplewski C, Groth M, Oldziej S, Wawak RJ, Rackovsky S, Pincus MR, Scheraga HA. A united-residue force field for off-lattice protein structure simulations. III. Origin of backbone hydrogen bonding cooperativity in united residue potential. Journal of Computational Chemistry 1998; 19:259–276.

[17] Liwo A, Odziej S, Czaplewski C, Kozlowska U, Scheraga HA. Parametrization of backbone-electrostatic and multibody contributions to the UNRES force field for protein-structure prediction from ab initio energy surfaces of model systems. Journal of Physical Chemistry B 2004;108:9421–9438.

[18] Lu H, Skolnick J. A distance-dependent knowledge-based potential for improved protein structure selection. Proteins: Structure, Function, and Bioinformatics 2001;44:223–232.

[19] Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. Journal of Molecular Biology 1998;275:895–916.

[20] Subramaniam S, Tcheng DK, Fenton J. A knowledge-based method for protein structure refinement and prediction. In: D States, P Agarwal, T Gaasterland, L Hunter, R Smith, editors, Proceedings of the Fourth International Conference on Intelligent

Systems in Molecular Biology. AAAI Press, Boston, MA, 1996; pp. 218–229.

[21] DeBolt S, Skolnick J. Evaluation of atomic level mean force potentials via inverse folding and inverse refinement of proteins structures: atomic burial position and pairwise non-bonded interactions. Protein Engineering 1996;9:637–655.

[22] Loose C, Klepeis JL, Floudas CA. A new pairwise folding potential based on improved decoy generation and side-chain packing. Proteins: Structure, Function, and Bioinformatics 2004;54:303–314.

[23] Tobi D, Elber R. Distance-dependent, pair potential for protein folding: Results from linear optimization. Proteins: Structure, Function, and Bioinformatics 2000;41:40–46.

[24] Hinds DA, Levitt M. Exploring conformational space with a simple lattice model for protein structure. Journal of Molecular Biology 1994;243:668–682.

[25] Floudas CA, Fung HK, McAllister SR, Mönnigmann M, Rajgaria R. Advances in protein structure prediction and de novo protein design : A review. Chemical Engineering Science 2006;61:966–988.

[26] Č Venclovas, Zemla A, Fidelis K, Moult J. Assessment of progress over the CASP experiments. Proteins: Structure, Function, and Bioinformatics 2003;53:585–595.

[27] Misura KMS, Morozov AV, Baker D. Analysis of anisotropic side-chain packing in proteins and application to high-resolution structure prediction. Journal of Molecular Biology 2004;342:651–664.

[28] Bradley P, Misura KMS, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 2005;309:1868–1871.

[29] Tobi D, Shafran G, Linial N, Elber R. On the design and analysis of protein folding potentials. Proteins: Structure, Function, and Bioinformatics 2000;40:71–85.

[30] Bahar I, Jernigan RL. Inter-residue potential in globular proteins and the dominance of highly specific hydrophillic interactions at close separation. Journal of Molecular Biology 1997;266:195–214.

[31] ILOG CPLEX User's Manual 9.0. 2003.

[32] Dash Optimization. Xpress-MP: Getting Started, 2003.

[33] Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proceedings of the National Academy of Sciences of the United States of America 2004; 101:7594–7599.

[34] Güntert P, Mumenthaler C, Wüthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. Journal of Molecular Biology 1997;273:283–298.

[35] Meller J, Wagner M, Elber R. Maximum feasibility guideline in the design and analysis of protein folding potentials. Journal of Computational Chemistry 2002;23:111–118.

[36] Park B, Levitt M. Energy functions that discriminate x-ray and near native folds from well-constructed decoys. Journal of Molecular Biology 1996;258:367.

[37] Loose C, Klepeis JL, Floudas CA. A new pairwise folding potential based on improved decoy generation and side-chain packing. Proteins: Structure, Function, and Bioinformatics 2004;54:303–314.

Table II: Classification of Amino Acids[22].

Hydrophilic (Neut) $\{\mathcal{PU}\}$	Hydrophilic (Pos) $\{\mathcal{PP}\}$	Hydrophobic Non-Aromatic $\{\mathcal{HN}\}$	Hydrophobic Aromatic $\{\mathcal{HA}\}$
GLY	LYS	CYS	PHE
HIS	ARG	ILE	TYR
ASN	Hydrophobic	LEU	TRP
PRO	(Neg) $\{\mathcal{PN}\}$	MET	Other
GLN	ASP	THR	$\{\mathcal{O}\}$
SER	GLU	VAL	ALA

Table III: Additional Classification of Amino Acids into Hydrophobic and Hydrophilic sets.[22]

Class	Amino Acids
$\mathcal{H}1$	PHE, ILE, LEU, MET, VAL
$\mathcal{H}2$	TRP, TYR
$\mathcal{P}1$	HIS, ASN, GLN, SER, THR
$\mathcal{P}2$	LYS, GLU, ASP

Table IV: Slack values used within the pairwise hydrophobic distance bounds of each torsion angle dynamics run.

Run	1	2	3	4	5	6	7	8
Slack (Å)	0.5	1.0	1.5	2.0	2.5	3.0	4.0	5.0

Table V: Distribution of minimum rmsd decoy structures across the protein set.

C-alpha RMSD	Total Count	Training Set	Test Set
0.0-0.5	13	12	1
0.5-1.0	518	458	60
1.0-1.5	681	607	74
1.5-2.0	188	173	15
2.0+	89	—	—

Table VI: Distribution of maximum rmsd decoy structures across the protein set.

C-alpha RMSD	Total Count	Training Set	Test Set
0.0-4.0	447	391	56
4.0-5.0	438	372	48
5.0-6.0	146	119	10
6.0-7.0	78	62	9
7.0+	380	306	27

Table VII: Distribution of median rmsd decoy structures across the protein set.

C-alpha RMSD	Total Count	Training Set	Test Set
0.0-2.0	402	353	49
2.0-2.5	555	494	61
2.5-3.0	259	224	29
3.0-4.0	176	128	9
4.0+	97	51	2

Table VIII: Distribution of mean rmsd decoy structures across the protein set.

C-alpha RMSD	Total Count	Training Set	Test Set
0.0-2.0	246	212	34
2.0-2.5	571	504	67
2.5-3.0	323	285	35
3.0-4.0	199	151	9
4.0+	150	98	5

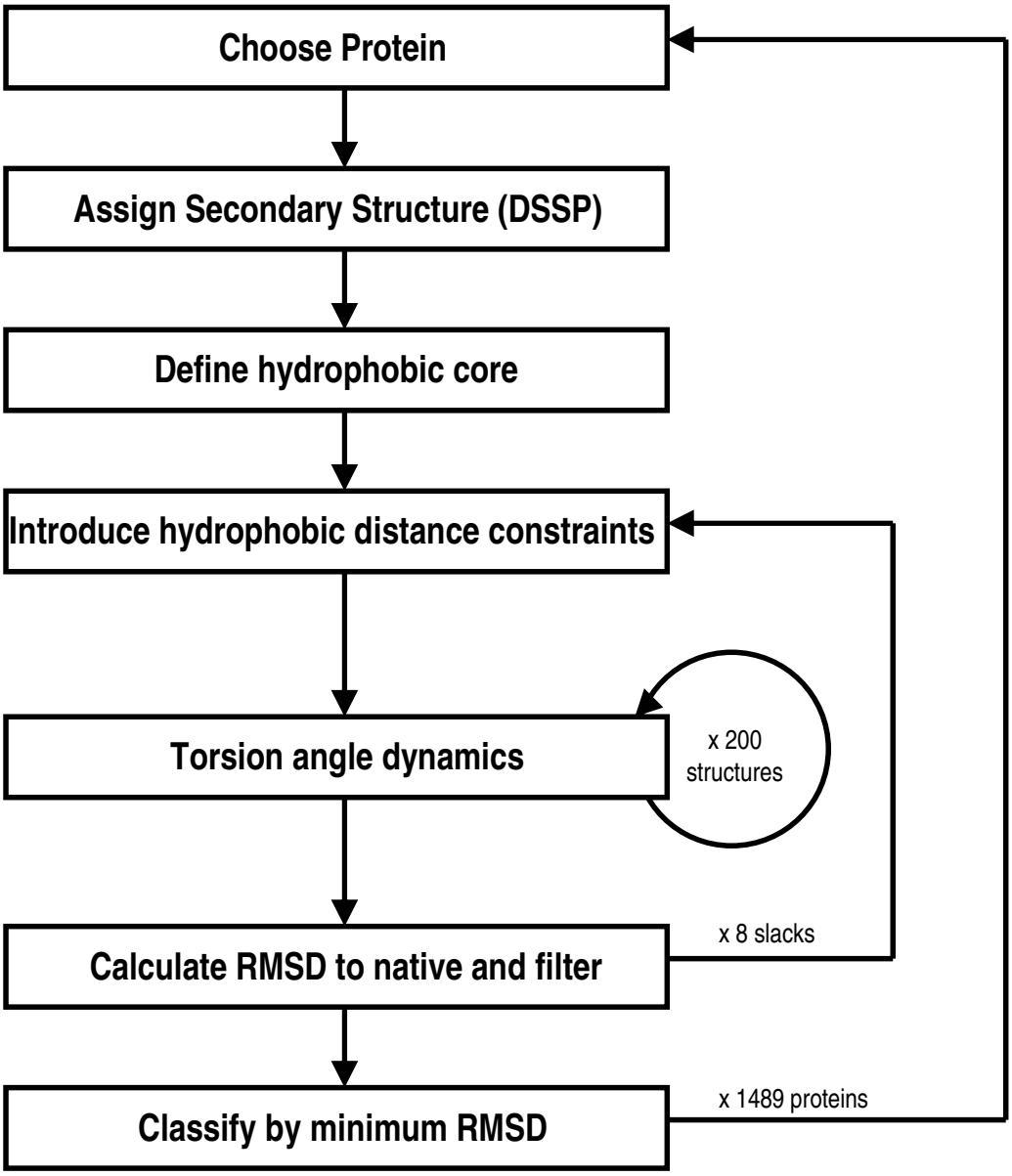

Figure 1: Flowchart showing the decoy generation process.

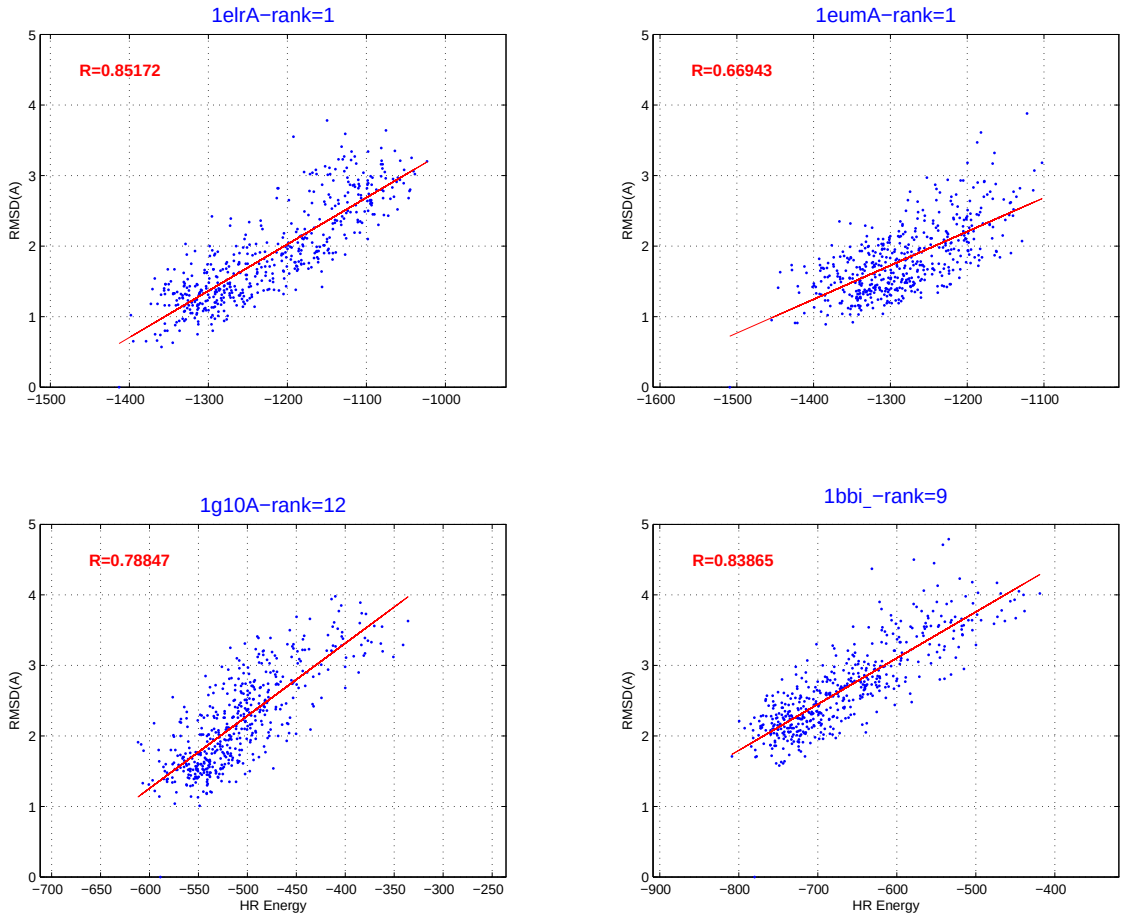

Figure 2: Energy-rmsd plot for 4 high resolution test cases.

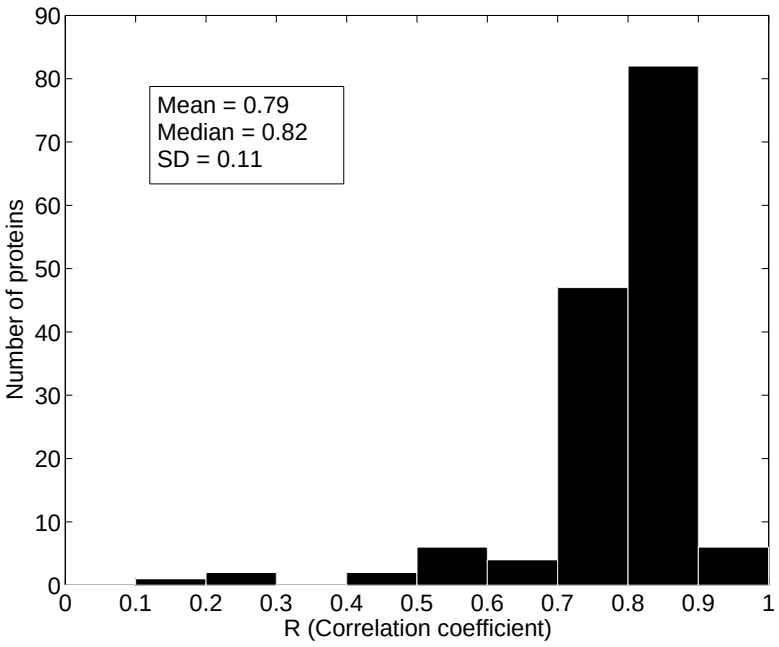

Figure 3: Distribution of R (correlation coefficient between energy and rmsd) for all 150 high resolution test cases.

Table IX: Rankings of the native conformations using the HR and TE-13 force field on the high resolution test set.

ID	HR	TE13	ID	HR	TE13	ID	HR	TE13	ID	HR	TE13
1elrA	1	2	1ash_	1	1	1qckA	1	1	1qkkA	1	1
1qe2A	2	2	1lis_	1	1	1std_	1	1	1eca_	2	1
1afi_	1	1	1b9lA	1	1	1vie_	2	1	1aueA	2	1
1h97A	1	1	1g2pA	1	1	1jmvA	1	1	2bopA	1	2
1jfuA	1	1	1gyzA	1	70	1kpsB	1	1	1d3bA	2	3
1hbiA	1	1	1eumA	1	1	1ezvH	1	9	1h6hA	1	1
1dbwA	7	1	1i4sA	1	1	1ten_	1	1	1vdrA	2	1
1am4A	4	1	1he1A	1	2	1nox_	1	1	1b4sA	1	1
1occJ	4	4	1pytA	2	1	1colA	1	2	1d02A	1	1
1ed7A	1	210	1n72A	1	1	1tiiD	1	12	121p_	1	1
1ae2_	1	1	1kncA	1	1	1kq5A	1	1	1cz3A	1	1
1dfx_	1	1	1g10A	12	69	1a7vA	1	1	1adr_	4	1
1ijxA	1	1	1jkeA	2	1	1mkp_	1	1	1cnoA	1	108
1a7d_	1	1	1an7A	1	1	1dujA	1	49	1ew6A	1	1
1fpzA	2	27	1mnmA	1	20	1acx_	3	3	1h8pA	8	1
1regX	1	1	1i3cA	1	2	1qqzA	1	292	1fvqA	1	2
1j77A	1	2	1otgA	1	1	2ezm_	3	1	1doaB	1	1
1fl7B	1	1	1elkA	1	1	1df7A	1	1	1jfmA	1	1
1jruA	1	59	1dv8A	1	2	1vpu_	1	66	1eh2_	4	19
1eq1A	1	16	1fhoA	1	2	1k3bC	4	25	1b78A	1	1
1gd5A	3	4	101m_	1	1	1ghj_	1	1	1kr7A	1	10
3caoA	4	98	1do6A	1	1	1ndoB	1	1	1bd7A	1	1
1bfs_	1	1	1d9nA	1	15	1f1mA	1	1	1gl4B	1	1
1i1jA	1	1	1cauA	1	2	1a3z_	5	1	1bs4A	1	1
1qmtA	1	1	1h8mA	1	11	1aj5A	1	1	1g1xC	1	1
1tbi_	1	1	1vmpA	3	12	1kxlA	1	1	1ljaA	1	33
1qlcA	1	10	1a1mB	2	3	1c4zD	1	1	1hli_	1	2
1pviA	1	1	1i6wA	1	1	1k3bB	1	1	1occH	2	1
1bjfA	1	1	1itpA	4	1	1b2pA	1	1	1b9wA	1	1
1lgbC	1	1	1qduB	1	1	1f4oA	1	3	1kqaA	1	1
2u1a_	8	270	1bvoA	3	2	1iqzA	1	3	1pauA	1	147
1hmjA	1	—	153l_	1	1	1bal_	1	53	1dqgA	8	1
1aa1S	3	1	1bam_	1	1	1a45_	1	1	1id1A	1	1
1eptB	1	19	1c01A	1	2	1bqz_	1	377	1cmoA	1	145
1dz7A	5	21	1msbA	2	1	1ffkU	1	—	1kxa_	1	1
1bbi_	9	1	1tsrA	1	1	1g9pA	23	217	1nukA	2	1
1a4yB	1	1	1d1rA	6	283	1dg4A	4	6	1abtA	1	2
1jbjA	1	9	1bq0_	1	19						
For	150	148	For	150	148						
AVE	1.872	19.94	FIRSTS	113	92						

Table X: Rankings of the native conformations using the high resolution (HR) force field on the LKF test set. Performance of the LKF and TE13 force field on this test set has been published in Loose et al.[37]

ID	HR	ID	HR	ID	HR	ID	HR
1a7v	2	1a90	1	1a9w	1	1aa0	1
1ab0	1	1ab2	1	1ab5	1	1ab6	1
1abo	1	1adn	7	1adr	2	1ae2	3
1ae3	1	1afi	1	1afj	1	1aj3	93
1an2	14	1ap4	1	1ar2	1	1auc	1
1aud	1	1aum	1	1avs	1	1aw3	2
1awd	1	1awp	1	1axq	1	1axx	6
1azq	1	1b0t	1	1b1a	1	1b20	1
1b21	1	1b27	1	1b2p	1	1b2s	1
1b2z	1	1b3i	1	1b3s	1	1b3t	1
1b4a	1	1b4c	3	1b5a	1	1b5b	1
1b5m	1	1b67	4	1b6c	1	1b7v	1
1b86	1	1b8c	1	1b8m	1	1b8r	1
1b9a	2	1b9l	1	1bai	2	1baj	8
1bbb	1	1bbz	3	1bd6	1	1bdj	1
1be2	1	1beo	1	1bf4	3	1bfe	1
1bfj	1	1bfm	13	1bfs	1	1bfx	1
1bhd	1	1bov	2	1box	1	1bhh	1
1bij	1	1bja	1	1bjx	1	1bk2	1
1bkf	1	1bl4	1	1bl8	1	1blj	1
1blk	1	1blv	1	1bm4	79	1bni	1
1bnj	1	1bnr	1	1bnz	23	1bo9	11
1bpt	99	1bq8	1	1bre	1	1brf	1
1brs	1	1btb	1	1btg	1	1bu1	1
1bu4	1	1buj	3	1buw	1	1buz	1
1bv4	1	1bv8	2	1bwe	1	1bwo	1
1bwu	1	1bwy	1	1bym	3	1byo	1
1byp	1	1bzd	1				
For	110						
AVE	4.32	FIRSTS	86				

Table XI: Testing force fields on 150 proteins of the high resolution decoy set. TE13 force field was only tested on 148 cases.

FF-Name	Average Rank	No of Firsts	Average rmsd
HR	1.87	113 (75.33%)	0.451
LKF	39.45	17 (11.33%)	1.721
TE13	19.94	92 (62.16%)	0.813
HL	44.93	70 (46.67%)	1.092

Table XII: Testing force fields on 150 proteins of the medium resolution set (LKF test set).

FF-Name	Average Rank	No of Firsts	Average rmsd
HR	4.32	86/110 (78.2 %)	1.903
LKF	5.84	93/151 (61.6 %)	3.510
TE13	17.36	43/131 (32.8 %)	not available
HL	92.88	3/150 (2.0 %)	8.436

Appendix 1 : Parameter values for high resolution C^{α}-C^{α} based distance dependent force field

Table A.I: Interaction Energies ($\theta_{IC,ID}$) for C^{α}-C^{α} distance dependent force field for Bin ID-1 (3-4 Å)

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	-5.82	-14.00	-1.00	-9.61	1.33	-1.94	-3.30	2.00	-3.66	2.00	2.00	-3.20	4.58	-1.35	1.83	-2.58	-1.42	-2.83	-4.34	2.00
CYS	-14.00	-5.20	5.71	5.87	-2.15	3.56	4.42	7.61	-0.33	6.01	4.88	-0.32	2.00	3.79	-1.44	-2.08	1.60	-1.21	-14.00	0.67
ASP	-1.00	5.71	9.20	9.55	10.00	-5.53	-7.10	10.00	-8.28	6.93	-7.73	-6.02	-3.03	-2.50	-6.81	-6.90	5.96	0.36	10.77	0.12
GLU	-9.61	5.87	9.55	11.23	-0.12	-5.30	-4.00	4.78	-10.35	-2.97	9.56	-4.33	1.21	-7.38	-7.77	-5.31	-3.35	4.15	3.52	-0.74
PHE	1.33	-2.15	10.00	-0.12	-2.36	3.49	0.54	-2.77	-1.18	-2.91	0.36	7.93	5.67	-7.86	-2.31	3.85	0.17	0.76	5.41	2.41
GLY	-1.94	3.56	-5.53	-5.30	3.49	-4.22	-5.70	1.61	-7.51	6.55	3.04	-6.40	-5.97	-6.03	-3.99	-5.30	1.30	3.76	5.04	0.98
HIS	-3.30	4.42	-7.10	-4.00	0.54	-5.70	-9.00	7.95	-7.91	5.52	3.44	-4.00	-4.00	-4.01	-9.00	-8.00	6.14	-9.21	8.00	2.00
ILE	2.00	7.61	10.00	4.78	-2.77	1.61	7.95	2.11	1.07	2.08	4.55	8.37	8.00	4.89	2.17	7.00	-0.18	3.66	3.56	2.11
LYS	-3.66	-0.33	-8.28	-10.35	-1.18	-7.51	-7.91	1.07	8.00	8.56	6.55	-7.85	-1.76	-4.80	8.88	-4.45	5.28	9.19	5.78	1.74
LEU	2.00	6.01	6.93	-2.97	-2.91	6.55	5.52	2.08	8.56	2.11	-0.60	1.75	9.02	6.25	8.00	5.57	4.85	2.11	2.11	2.11
MET	2.00	4.88	-7.73	9.56	0.36	3.04	3.44	4.55	6.55	-0.60	1.17	6.95	2.00	2.42	1.91	-10.82	8.00	0.36	4.00	1.17
ASN	-3.20	-0.32	-6.02	-4.33	7.93	-6.40	-4.00	8.37	-7.85	1.75	6.95	-8.48	-8.56	-4.00	-5.68	-4.00	4.67	0.40	-0.43	1.74
PRO	4.58	2.00	-3.03	1.21	5.67	-5.97	-4.00	8.00	-1.76	9.02	2.00	-8.56	-5.00	-2.92	-4.65	-5.01	7.95	2.00	4.93	6.12
GLN	-1.35	3.79	-2.50	-7.38	-7.86	-6.03	-4.01	4.89	-4.80	6.25	2.42	-4.00	-2.92	-2.21	-4.16	-4.00	2.10	4.74	-10.04	6.52
ARG	1.83	-1.44	-6.81	-7.77	-2.31	-3.99	-9.00	2.17	8.88	8.00	1.91	-5.68	-4.65	-4.16	13.61	-4.88	-5.00	5.85	-1.08	-0.12
SER	-2.58	-2.08	-6.90	-5.31	3.85	-5.30	-8.00	7.00	-4.45	5.57	-10.82	-4.00	-5.01	-4.00	-4.88	-4.68	-1.10	-1.69	-3.50	-5.11
THR	-1.42	1.60	5.96	-3.35	0.17	1.30	6.14	-0.18	5.28	4.85	8.00	4.67	7.95	2.10	-5.00	-1.10	5.74	0.31	-1.42	-0.53
VAL	-2.83	-1.21	0.36	4.15	0.76	3.76	-9.21	3.66	9.19	2.11	0.36	0.40	2.00	4.74	5.85	-1.69	0.31	2.11	4.88	2.11
TRP	-4.34	-14.00	10.77	3.52	5.41	5.04	8.00	3.56	5.78	2.11	4.00	-0.43	4.93	-10.04	-1.08	-3.50	-1.42	4.88	6.11	5.03
TYR	2.00	0.67	0.12	-0.74	2.41	0.98	2.00	2.11	1.74	2.11	1.17	1.74	6.12	6.52	-0.12	-5.11	-0.53	2.11	5.03	6.11

Table A.II: Interaction Energies ($\theta_{IC,ID}$) for C^{α}-C^{α} distance dependent force field for Bin ID-2 (4-5 Å)

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	-6.00	-6.00	-3.73	-4.66	-6.00	-2.53	-5.00	-6.00	-3.82	-6.00	-6.00	-2.82	-3.42	-3.38	-5.00	-3.84	-6.00	-6.00	-4.00	-6.00
CYS	-6.00	-9.20	-2.29	-2.13	-6.15	-0.75	-3.58	-0.39	-2.72	-1.48	-3.12	-0.12	-6.00	-4.21	-3.13	-3.33	-6.40	-5.21	-6.00	-3.33
ASP	-3.73	-2.29	5.20	5.55	2.00	-1.53	-3.10	2.00	-4.28	3.75	-1.05	-2.02	0.97	1.50	-2.81	-2.90	-2.04	3.62	2.77	2.00
GLU	-4.66	-2.13	5.55	7.23	2.96	-1.30	0.00	1.24	-6.35	1.12	1.56	-0.33	5.21	0.62	-3.77	-1.30	-2.31	-0.40	0.45	1.19
PHE	-6.00	-6.15	2.00	2.96	-6.36	-1.69	-0.86	-6.77	-0.07	-6.98	-7.64	-0.07	-2.33	-1.44	-1.40	-0.07	-3.83	-5.90	-1.06	-5.31
GLY	-2.53	-0.75	-1.53	-1.30	-1.69	-0.22	-1.70	1.75	-3.51	0.80	2.57	-2.40	-1.97	-2.03	0.01	-1.30	-2.49	0.37	-0.25	-1.07
HIS	-5.00	-3.58	-3.10	0.00	-0.86	-1.70	-5.00	0.00	-3.91	-1.09	-4.56	0.00	0.00	0.00	-5.00	0.00	-1.86	-4.31	0.00	-6.00
ILE	-6.00	-0.39	2.00	1.24	-6.77	1.75	0.00	-1.89	1.10	-1.92	-2.83	0.37	0.00	0.26	0.00	0.70	-4.18	-1.89	-0.44	-1.89
LYS	-3.82	-2.72	-4.28	-6.35	-0.07	-3.51	-3.91	1.10	4.00	1.12	0.06	-3.85	2.24	-2.06	4.88	-0.45	-0.94	1.19	-0.85	-1.95
LEU	-6.00	-1.48	3.75	1.12	-6.98	0.80	-1.09	-1.92	1.12	-1.89	-4.60	1.12	1.02	-0.25	0.00	0.21	-3.15	-1.89	-1.89	-1.89
MET	-6.00	-3.12	-1.05	1.56	-7.64	2.57	-4.56	-2.83	0.06	-4.60	-2.83	-1.05	-6.00	-5.58	0.06	-2.82	0.00	-3.64	0.00	-2.83
ASN	-2.82	-0.12	-2.02	-0.33	-0.07	-2.40	0.00	0.37	-3.85	1.12	-1.05	-4.48	-4.56	0.00	-1.68	0.00	-3.33	-0.89	0.00	-1.26
PRO	-3.42	-6.00	0.97	5.21	-2.33	-1.97	0.00	0.00	2.24	1.02	-6.00	-4.56	-1.00	1.08	-0.65	-1.01	-0.05	-6.00	0.75	-1.88
GLN	-3.38	-4.21	1.50	0.62	-1.44	-2.03	0.00	0.26	-2.06	-0.25	-5.58	0.00	1.08	1.79	-0.16	0.00	-3.55	-0.78	-3.30	-1.48
ARG	-5.00	-3.13	-2.81	-3.77	-1.40	0.01	-5.00	0.00	4.88	0.00	0.06	-1.68	-0.65	-0.16	5.61	-0.88	-2.86	-0.56	-6.00	-2.58
SER	-3.84	-3.33	-2.90	-1.30	-0.07	-1.30	0.00	0.70	-0.45	0.21	-2.82	0.00	-1.01	0.00	-0.88	-0.68	-2.06	-4.87	3.45	-1.41
THR	-6.00	-6.40	-2.04	-2.31	-3.83	-2.49	-1.86	-4.18	-0.94	-3.15	0.00	-3.33	-0.05	-3.55	-2.86	-2.06	0.00	-3.69	-6.72	-4.53
VAL	-6.00	-5.21	3.62	-0.40	-5.90	0.37	-4.31	-1.89	1.19	-1.89	-3.64	-0.89	-6.00	-0.78	-0.56	-4.87	-3.69	-1.89	-1.89	-1.89
TRP	-4.00	-6.00	2.77	0.45	-1.06	-0.25	0.00	-0.44	-0.85	-1.89	0.00	0.00	0.75	-3.30	-6.00	3.45	-6.72	-1.89	-1.89	-1.89
TYR	-6.00	-3.33	2.00	1.19	-5.31	-1.07	-6.00	-1.89	-1.95	-1.89	-2.83	-1.26	-1.88	-1.48	-2.58	-1.41	-4.53	-1.89	-1.89	-1.89

Table A.III: Interaction Energies ($\theta_{IC,ID}$) for C^{α}-C^{α} distance dependent force field for Bin ID-3 (5-5.5 Å)

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	-4.00	-4.00	-2.73	-3.66	-4.00	-1.53	-4.00	-4.00	-2.82	-4.00	-4.00	-1.82	-2.42	-2.38	-4.00	-2.84	-4.00	-4.00	-2.00	-4.00
CYS	-4.00	-13.20	-4.00	-4.00	-8.15	-0.67	-4.00	-2.39	-3.87	-3.48	-1.71	-4.00	-4.00	-4.00	-1.16	-4.00	-8.40	-7.21	-4.00	-5.33
ASP	-2.73	-4.00	4.20	3.33	0.00	-4.00	-3.29	0.00	-7.49	0.00	-3.37	-4.00	-3.03	-2.50	-6.08	-4.00	-4.00	0.00	0.77	0.00
GLU	-3.66	-4.00	3.33	3.35	1.04	-1.89	-3.48	-1.04	-7.22	-1.87	1.37	-2.07	1.79	-3.38	-5.61	-4.00	-3.12	-1.89	-2.78	-2.48
PHE	-4.00	-8.15	0.00	1.04	-7.36	-3.74	-3.78	-8.77	-2.95	-8.98	-9.64	-4.00	-4.00	-2.95	-3.51	-3.29	-5.83	-7.90	-2.06	-6.31
GLY	-1.53	-0.67	-4.00	-1.89	-3.74	-3.04	-4.00	-0.25	-2.19	-1.48	-1.05	-2.82	-0.97	-4.00	-1.88	-2.51	-4.00	-1.53	-4.00	-4.00
HIS	-4.00	-4.00	-3.29	-3.48	-3.78	-4.00	-4.00	-2.00	-2.91	-4.00	-3.88	-3.74	-4.00	-1.67	-4.00	-4.00	-4.00	-4.00	-4.00	-3.37
ILE	-4.00	-2.39	0.00	-1.04	-8.77	-0.25	-2.00	-5.89	-0.90	-5.92	-4.83	-1.63	-2.00	-2.00	-2.00	-2.00	-6.18	-5.89	-2.44	-4.71
LYS	-2.82	-3.87	-7.49	-7.22	-2.95	-2.19	-2.91	-0.90	3.00	-2.21	-1.22	-4.00	-1.66	-3.48	3.00	-4.00	-4.00	-2.81	-3.76	-4.00
LEU	-4.00	-3.48	0.00	-1.87	-8.98	-1.48	-4.00	-5.92	-2.21	-5.89	-8.57	-2.21	0.64	-2.25	-4.00	-3.31	-5.15	-5.89	-4.42	-4.18
MET	-4.00	-1.71	-3.37	1.37	-9.64	-1.05	-3.88	-4.83	-1.22	-8.57	-6.17	-3.60	-4.00	-4.00	-3.45	-3.68	-2.00	-5.70	-2.00	-4.83
ASN	-1.82	-4.00	-4.00	-2.07	-4.00	-2.82	-3.74	-1.63	-4.00	-2.21	-3.60	-3.68	-2.68	-4.00	-4.00	-4.00	-4.00	-3.22	-3.33	-1.84
PRO	-2.42	-4.00	-3.03	1.79	-4.00	-0.97	-4.00	-2.00	-1.66	0.64	-4.00	-2.68	3.00	-2.92	-2.09	-2.66	-1.73	-4.00	-0.63	-4.00
GLN	-2.38	-4.00	-2.50	-3.38	-2.95	-4.00	-1.67	-2.00	-3.48	-2.25	-4.00	-4.00	-2.92	-2.21	-1.62	-4.00	-4.00	-4.00	-3.63	-4.00
ARG	-4.00	-1.16	-6.08	-5.61	-3.51	-1.88	-4.00	-2.00	3.00	-4.00	-3.45	-4.00	-2.09	-1.62	3.00	-4.00	-4.00	-3.50	-3.51	-3.51
SER	-2.84	-4.00	-4.00	-4.00	-3.29	-2.51	-4.00	-2.00	-4.00	-3.31	-3.68	-4.00	-2.66	-4.00	-4.00	-2.48	-4.00	-2.81	-0.46	-3.55
THR	-4.00	-8.40	-4.00	-3.12	-5.83	-4.00	-4.00	-6.18	-4.00	-5.15	-2.00	-4.00	-1.73	-4.00	-4.00	-4.00	-2.00	-5.69	-8.72	-6.53
VAL	-4.00	-7.21	0.00	-1.89	-7.90	-1.53	-4.00	-5.89	-2.81	-5.89	-5.70	-3.22	-4.00	-4.00	-3.50	-2.81	-5.69	-5.89	-3.89	-3.99
TRP	-2.00	-4.00	0.77	-2.78	-2.06	-4.00	-4.00	-2.44	-3.76	-4.42	-2.00	-3.33	-0.63	-3.63	-3.51	-0.46	-8.72	-3.89	-1.89	-3.10
TYR	-4.00	-5.33	0.00	-2.48	-6.31	-4.00	-3.37	-4.71	-4.00	-4.18	-4.83	-1.84	-4.00	-4.00	-3.51	-3.55	-6.53	-3.99	-3.10	-3.43

Table A.IV: Interaction Energies ($\theta_{IC,ID}$) for C^{α}-C^{α} distance dependent force field for Bin ID-4 (5.5-6 Å)

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	-2.00	-2.00	-1.73	-2.66	-2.00	-0.53	-2.57	-2.00	-1.82	-2.00	-2.00	-0.82	-1.42	-1.38	-3.00	-1.84	-2.00	-2.00	0.00	-2.00
CYS	-2.00	-17.20	-3.03	-3.49	-10.15	1.78	-4.00	-4.39	-4.00	-5.48	-1.03	-3.24	-2.54	-4.00	-3.27	-1.73	-10.40	-9.21	-4.00	-7.33
ASP	-1.73	-3.03	3.20	2.33	-2.00	-2.35	-3.18	-2.00	-7.10	-2.00	-1.20	-3.50	-4.00	-2.80	-3.81	-3.24	-3.98	-2.00	-1.23	-2.00
GLU	-2.66	-3.49	2.33	2.35	0.61	-1.51	-3.50	-3.04	-6.38	-2.71	-0.47	-4.00	0.26	-4.00	-5.96	-2.57	-3.55	-2.15	-2.10	-4.00
PHE	-2.00	-10.15	-2.00	0.61	-8.36	-1.03	-4.00	-10.77	-3.51	-10.98	-11.64	-4.00	-3.94	-4.00	-3.51	-4.00	-7.83	-9.90	-3.06	-7.31
GLY	-0.53	1.78	-2.35	-1.51	-1.03	-1.61	-1.95	-2.25	-2.48	-2.02	-1.02	-3.02	-0.02	-1.20	-2.76	-1.55	-2.62	-0.54	-1.73	-2.65
HIS	-2.57	-4.00	-3.18	-3.50	-4.00	-1.95	-3.00	-4.00	-1.91	-3.32	-4.00	-4.00	-4.00	-1.80	-3.00	-1.54	-2.15	-3.77	-3.80	-4.00
ILE	-2.00	-4.39	-2.00	-3.04	-10.77	-2.25	-4.00	-9.89	-2.90	-9.92	-6.83	-3.63	-4.00	-4.00	-4.00	-4.00	-8.18	-9.89	-4.44	-6.71
LYS	-1.82	-4.00	-7.10	-6.38	-3.51	-2.48	-1.91	-2.90	2.00	-1.77	-0.37	-3.93	-2.20	-3.37	2.00	-2.63	-3.67	-1.90	-4.00	-4.00
LEU	-2.00	-5.48	-2.00	-2.71	-10.98	-2.02	-3.32	-9.92	-1.77	-9.89	-10.57	-4.00	-3.24	-4.00	-4.00	-3.77	-7.15	-9.89	-6.42	-6.18
MET	-2.00	-1.03	-1.20	-0.47	-11.64	-1.02	-4.00	-6.83	-0.37	-10.57	-8.17	-3.27	-1.70	-3.60	-1.75	-3.40	-4.00	-7.70	-5.64	-6.83
ASN	-0.82	-3.24	-3.50	-4.00	-4.00	-3.02	-4.00	-3.63	-3.93	-4.00	-3.27	-3.94	-3.73	-4.00	-3.68	-4.00	-3.24	-2.29	-2.72	-4.00
PRO	-1.42	-2.54	-4.00	0.26	-3.94	-0.02	-4.00	-4.00	-2.20	-3.24	-1.70	-3.73	0.15	-1.16	-4.00	-2.74	-2.44	-3.39	-0.83	-4.00
GLN	-1.38	-4.00	-2.80	-4.00	-4.00	-1.20	-1.80	-4.00	-3.37	-4.00	-3.60	-4.00	-1.16	-4.00	-1.39	-2.11	-2.96	-2.96	-3.76	-3.89
ARG	-3.00	-3.27	-3.81	-5.96	-3.51	-2.76	-3.00	-4.00	2.00	-4.00	-1.75	-3.68	-4.00	-1.39	2.00	-2.95	-4.00	-3.34	-4.00	-3.72
SER	-1.84	-1.73	-3.24	-2.57	-4.00	-1.55	-1.54	-4.00	-2.63	-3.77	-3.40	-4.00	-2.74	-2.11	-2.95	-3.33	-0.88	-3.00	-2.58	-1.27
THR	-2.00	-10.40	-3.98	-3.55	-7.83	-2.62	-2.15	-8.18	-3.67	-7.15	-4.00	-3.24	-2.44	-2.96	-4.00	-0.88	-4.00	-7.73	-10.72	-8.53
VAL	-2.00	-9.21	-2.00	-2.15	-9.90	-0.54	-3.77	-9.89	-1.90	-9.89	-7.70	-2.29	-3.39	-2.96	-3.34	-3.00	-7.73	-9.89	-5.89	-5.99
TRP	0.00	-4.00	-1.23	-2.10	-3.06	-1.73	-3.80	-4.44	-4.00	-6.42	-5.64	-2.72	-0.83	-3.76	-4.00	-2.58	-10.72	-5.89	-3.96	-2.82
TYR	-2.00	-7.33	-2.00	-4.00	-7.31	-2.65	-4.00	-6.71	-4.00	-6.18	-6.83	-4.00	-4.00	-3.89	-3.72	-1.27	-8.53	-5.99	-2.82	-5.25

Table A.V: Interaction Energies ($\theta_{IC,ID}$) for C^{α}-C^{α} distance dependent force field for Bin ID-5 (6-6.5 Å)

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	-3.33	-0.19	-1.83	-2.76	-2.28	-0.37	-2.51	-2.52	-1.55	-3.81	-2.67	-1.13	-0.82	-2.16	-3.45	-0.00	-0.76	-2.58	-2.62	-2.34
CYS	-0.19	-15.20	-2.08	-2.86	-8.32	2.08	-4.00	-5.34	-1.91	-4.52	-2.99	-4.00	-0.39	-2.50	-4.00	-1.41	-9.21	-7.21	0.00	-5.33
ASP	-1.83	-2.08	2.20	1.33	-1.74	0.54	-1.70	-0.71	-5.98	0.46	-1.30	-1.66	-0.94	-1.70	-3.48	-2.48	-2.70	0.15	-1.48	-0.44
GLU	-2.76	-2.86	1.33	1.35	0.73	0.53	-0.66	-1.50	-6.31	-1.72	0.14	-1.98	1.56	-3.18	-6.65	-0.99	-1.74	-1.36	-3.48	-2.23
PHE	-2.28	-8.32	-1.74	0.73	-8.36	-1.10	-3.20	-8.77	-1.98	-8.98	-9.64	-1.98	-0.71	-3.23	-2.76	-1.98	-6.10	-9.66	-3.06	-7.31
GLY	-0.37	2.08	0.54	0.53	-1.10	-0.47	-2.47	1.29	-1.02	-0.05	-0.59	0.49	0.08	-2.77	-0.30	0.21	-0.49	0.94	-2.90	-1.64
HIS	-2.51	-4.00	-1.70	-0.66	-3.20	-2.47	-3.00	-2.95	-1.93	-2.41	-2.82	-4.00	-2.50	-0.77	-1.82	0.71	-2.81	-3.89	-2.88	-1.80
ILE	-2.52	-5.34	-0.71	-1.50	-8.77	1.29	-2.95	-8.61	-2.10	-9.44	-8.44	-2.10	-3.64	-3.02	-4.00	-2.16	-6.20	-7.89	-6.81	-6.04
LYS	-1.55	-1.91	-5.98	-6.31	-1.98	-1.02	-1.93	-2.10	1.00	-1.72	-1.30	-2.79	-2.01	-2.42	1.00	-2.56	-2.96	-0.58	-1.58	-2.39
LEU	-3.81	-4.52	0.46	-1.72	-8.98	-0.05	-2.41	-9.44	-1.72	-10.65	-8.57	-3.03	-2.12	-3.88	-2.88	-1.72	-5.88	-8.65	-7.16	-5.10
MET	-2.67	-2.99	-1.30	0.14	-9.64	-0.59	-2.82	-8.44	-1.30	-8.57	-8.67	-1.30	-2.55	-1.30	-2.07	-2.62	-2.10	-8.44	-3.64	-4.83
ASN	-1.13	-4.00	-1.66	-1.98	-1.98	0.49	-4.00	-2.10	-2.79	-3.03	-1.30	-2.50	-1.08	-3.28	-1.90	-2.50	-2.36	-1.36	-1.33	-4.00
PRO	-0.82	-0.39	-0.94	1.56	-0.71	0.08	-2.50	-3.64	-2.01	-2.12	-2.55	-1.08	-1.23	-2.73	-1.94	-2.10	0.30	-3.17	-2.15	-2.76
GLN	-2.16	-2.50	-1.70	-3.18	-3.23	-2.77	-0.77	-3.02	-2.42	-3.88	-1.30	-3.28	-2.73	-2.52	-2.95	-0.30	-1.78	-1.62	-1.78	-4.00
ARG	-3.45	-4.00	-3.48	-6.65	-2.76	-0.30	-1.82	-4.00	1.00	-2.88	-2.07	-1.90	-1.94	-2.95	1.00	-3.37	-2.85	-4.00	-3.66	-2.76
SER	-0.00	-1.41	-2.48	-0.99	-1.98	0.21	0.71	-2.16	-2.56	-1.72	-2.62	-2.50	-2.10	-0.30	-3.37	-1.22	-1.35	-1.95	1.04	-0.90
THR	-0.76	-9.21	-2.70	-1.74	-6.10	-0.49	-2.81	-6.20	-2.96	-5.88	-2.10	-2.36	0.30	-1.78	-2.85	-1.35	-3.24	-5.73	-8.72	-6.53
VAL	-2.58	-7.21	0.15	-1.36	-9.66	0.94	-3.89	-7.89	-0.58	-8.65	-8.44	-1.36	-3.17	-1.62	-4.00	-1.95	-5.73	-7.89	-3.89	-4.61
TRP	-2.62	0.00	-1.48	-3.48	-3.06	-2.90	-2.88	-6.81	-1.58	-7.16	-3.64	-1.33	-2.15	-1.78	-3.66	1.04	-8.72	-3.89	-4.84	-2.82
TYR	-2.34	-5.33	-0.44	-2.23	-7.31	-1.64	-1.80	-6.04	-2.39	-5.10	-4.83	-4.00	-2.76	-4.00	-2.76	-0.90	-6.53	-4.61	-2.82	-5.25

Table A.VI: Interaction Energies ($\theta_{IC,ID}$) for C^{α}-C^{α} distance dependent force field for Bin ID-6 (6.5-7 Å)

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	-3.25	-1.15	0.49	-1.82	-0.23	2.05	-0.75	-2.30	-0.57	-2.40	1.33	0.82	-0.88	1.00	-2.05	2.14	0.73	-1.52	-1.80	-1.04
CYS	-1.15	-11.20	1.92	-2.27	-8.28	3.79	-0.17	-3.80	-0.85	-3.98	-0.77	-3.11	-1.78	-1.02	-0.56	-1.94	-5.91	-5.24	-4.00	-3.78
ASP	0.49	1.92	1.20	0.33	1.24	0.67	-2.07	1.51	-2.08	1.69	1.56	0.81	-0.81	1.12	-1.96	0.31	-2.26	1.70	2.52	1.93
GLU	-1.82	-2.27	0.33	0.35	1.67	0.05	-1.05	-0.52	-4.17	-0.61	0.09	-0.60	1.88	0.23	-4.67	-0.06	-1.13	1.20	-0.33	-2.59
PHE	-0.23	-8.28	1.24	1.67	-8.17	0.74	-3.37	-6.17	0.71	-6.15	-6.15	-0.40	1.97	-0.33	-1.48	0.71	-3.35	-6.15	-1.48	-6.15
GLY	2.05	3.79	0.67	0.05	0.74	1.34	-0.36	0.80	-1.07	0.72	0.26	0.70	1.33	0.78	0.46	0.90	-0.03	1.67	-1.07	-0.26
HIS	-0.75	-0.17	-2.07	-1.05	-3.37	-0.36	1.00	-0.52	1.44	-0.68	-0.07	0.00	0.46	-0.76	-0.57	0.15	0.30	0.11	0.65	-0.68
ILE	-2.30	-3.80	1.51	-0.52	-6.17	0.80	-0.52	-7.06	-0.52	-6.53	-4.44	-1.37	-0.85	-3.47	-2.27	-0.52	-3.83	-5.25	-4.44	-4.28
LYS	-0.57	-0.85	-2.08	-4.17	0.71	-1.07	1.44	-0.52	0.00	-0.61	1.09	0.79	-0.80	-1.83	0.00	-1.46	-1.87	0.33	1.93	-1.33
LEU	-2.40	-3.98	1.69	-0.61	-6.15	0.72	-0.68	-6.53	-0.61	-6.68	-4.62	-1.25	-2.04	-0.88	-1.22	-0.61	-4.36	-6.33	-4.62	-4.62
MET	1.33	-0.77	1.56	0.09	-6.15	0.26	-0.07	-4.44	1.09	-4.62	-4.67	0.09	1.45	-2.37	1.09	-1.50	0.07	-4.44	-4.27	-2.41
ASN	0.82	-3.11	0.81	-0.60	-0.40	0.70	0.00	-1.37	0.79	-1.25	0.09	-2.88	-1.04	-0.68	0.16	0.22	-2.23	0.33	-1.47	-2.18
PRO	-0.88	-1.78	-0.81	1.88	1.97	1.33	0.46	-0.85	-0.80	-2.04	1.45	-1.04	1.64	-0.55	-1.10	-2.50	-0.39	-1.50	0.24	-2.19
GLN	1.00	-1.02	1.12	0.23	-0.33	0.78	-0.76	-3.47	-1.83	-0.88	-2.37	-0.68	-0.55	-1.48	-1.53	0.66	0.30	0.33	0.21	-1.34
ARG	-2.05	-0.56	-1.96	-4.67	-1.48	0.46	-0.57	-2.27	0.00	-1.22	1.09	0.16	-1.10	-1.53	0.00	-1.72	-1.55	-0.47	-1.48	-1.48
SER	2.14	-1.94	0.31	-0.06	0.71	0.90	0.15	-0.52	-1.46	-0.61	-1.50	0.22	-2.50	0.66	-1.72	0.86	-0.85	-0.50	0.51	1.62
THR	0.73	-5.91	-2.26	-1.13	-3.35	-0.03	0.30	-3.83	-1.87	-4.36	0.07	-2.23	-0.39	0.30	-1.55	-0.85	-1.65	-3.27	-7.19	-5.59
VAL	-1.52	-5.24	1.70	1.20	-6.15	1.67	0.11	-5.25	0.33	-6.33	-4.44	0.33	-1.50	0.33	-0.47	-0.50	-3.27	-5.25	0.11	-4.16
TRP	-1.80	-4.00	2.52	-0.33	-1.48	-1.07	0.65	-4.44	1.93	-4.62	-4.27	-1.47	0.24	0.21	-1.48	0.51	-7.19	0.11	-2.38	-0.93
TYR	-1.04	-3.78	1.93	-2.59	-6.15	-0.26	-0.68	-4.28	-1.33	-4.62	-2.41	-2.18	-2.19	-1.34	-1.48	1.62	-5.59	-4.16	-0.93	-4.44

Table A.VII: Interaction Energies ($\theta_{IC,ID}$) for C^{α}-C^{α} distance dependent force field for Bin ID-7 (7-8 Å)

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	-2.95	-0.95	0.64	-0.19	-1.96	1.37	1.02	-0.18	0.46	-2.23	-1.48	1.50	-0.35	0.63	-0.36	1.58	-0.40	-0.51	-1.71	-2.07
CYS	-0.95	-7.20	1.12	-2.26	-4.28	2.51	-0.10	-2.98	0.35	-2.49	-4.00	-0.76	-1.20	1.03	-2.65	0.62	-3.99	-2.30	-1.90	-3.62
ASP	0.64	1.12	1.20	0.33	1.28	0.70	0.24	1.35	-1.81	1.48	2.15	1.32	0.18	1.47	-1.98	-0.69	-1.13	3.50	1.45	-0.28
GLU	-0.19	-2.26	0.33	0.00	0.67	1.48	0.29	0.89	-3.04	-1.15	1.18	0.45	0.65	0.83	-2.12	0.89	-2.63	0.94	-2.25	-1.37
PHE	-1.96	-4.28	1.28	0.67	-5.12	1.31	-0.68	-3.66	1.09	-5.43	-2.40	-0.16	0.13	0.12	-0.09	1.57	-3.56	-4.86	-1.45	-3.74
GLY	1.37	2.51	0.70	1.48	1.31	1.11	-0.18	3.02	0.53	1.73	0.23	0.94	1.13	-0.64	1.71	0.75	0.34	2.79	1.61	0.85
HIS	1.02	-0.10	0.24	0.29	-0.68	-0.18	0.40	1.08	0.35	-1.80	-3.89	0.37	-1.55	1.92	-0.23	-0.24	0.26	0.24	-2.45	-0.20
ILE	-0.18	-2.98	1.35	0.89	-3.66	3.02	1.08	-4.73	0.23	-3.70	-4.74	0.52	-0.68	-0.73	0.08	1.01	-1.73	-4.14	-1.82	-2.47
LYS	0.46	0.35	-1.81	-3.04	1.09	0.53	0.35	0.23	0.00	0.65	1.11	1.15	-1.19	-0.43	0.00	-0.49	-1.54	1.59	1.20	0.21
LEU	-2.23	-2.49	1.48	-1.15	-5.43	1.73	-1.80	-3.70	0.65	-4.11	-2.94	2.48	-1.39	-0.90	0.05	0.75	-2.95	-3.85	-2.06	-1.49
MET	-1.48	-4.00	2.15	1.18	-2.40	0.23	-3.89	-4.74	1.11	-2.94	-4.76	0.81	-1.86	1.63	2.80	-0.62	-0.06	-1.84	-0.27	1.09
ASN	1.50	-0.76	1.32	0.45	-0.16	0.94	0.37	0.52	1.15	2.48	0.81	-0.17	0.59	0.09	0.91	0.26	1.16	1.85	1.44	0.04
PRO	-0.35	-1.20	0.18	0.65	0.13	1.13	-1.55	-0.68	-1.19	-1.39	-1.86	0.59	0.84	-0.03	-1.35	-1.19	-0.37	-1.23	0.04	-2.11
GLN	0.63	1.03	1.47	0.83	0.12	-0.64	1.92	-0.73	-0.43	-0.90	1.63	0.09	-0.03	1.92	-1.67	0.74	0.60	0.38	2.85	-1.76
ARG	-0.36	-2.65	-1.98	-2.12	-0.09	1.71	-0.23	0.08	0.00	0.05	2.80	0.91	-1.35	-1.67	0.00	-0.30	1.39	0.66	0.90	0.91
SER	1.58	0.62	-0.69	0.89	1.57	0.75	-0.24	1.01	-0.49	0.75	-0.62	0.26	-1.19	0.74	-0.30	0.79	1.44	0.17	3.70	2.09
THR	-0.40	-3.99	-1.13	-2.63	-3.56	0.34	0.26	-1.73	-1.54	-2.95	-0.06	1.16	-0.37	0.60	1.39	1.44	-2.72	-2.20	-7.89	-1.91
VAL	-0.51	-2.30	3.50	0.94	-4.86	2.79	0.24	-4.14	1.59	-3.85	-1.84	1.85	-1.23	0.38	0.66	0.17	-2.20	-5.66	-2.80	-3.12
TRP	-1.71	-1.90	1.45	-2.25	-1.45	1.61	-2.45	-1.82	1.20	-2.06	-0.27	1.44	0.04	2.85	0.90	3.70	-7.89	-2.80	-2.30	-2.82
TYR	-2.07	-3.62	-0.28	-1.37	-3.74	0.85	-0.20	-2.47	0.21	-1.49	1.09	0.04	-2.11	-1.76	0.91	2.09	-1.91	-3.12	-2.82	-3.77

Table A.VIII: Interaction Energies ($\theta_{IC,ID}$) for C^{α}-C^{α} distance dependent force field for Bin ID-8 (8-9 Å)

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	-0.16	0.18	0.01	-0.11	-2.28	1.04	-0.12	-0.88	-0.12	-0.62	-0.35	0.84	0.23	-0.14	0.32	-0.53	-0.94	-0.47	-0.04	-1.55
CYS	0.18	-4.00	0.71	-0.80	-1.10	0.96	-2.84	-0.63	-1.07	0.22	-1.18	-0.17	0.21	-0.41	-0.92	1.02	-2.27	0.21	-1.06	0.27
ASP	0.01	0.71	1.03	0.00	-0.03	0.74	-0.19	1.15	-1.13	1.25	0.20	-0.07	-0.55	1.21	-1.07	0.87	0.24	1.90	-0.30	0.34
GLU	-0.11	-0.80	0.00	0.00	-0.01	0.21	1.12	-0.93	-1.62	-0.49	2.62	0.59	-0.50	0.85	-0.70	0.76	-0.98	-0.24	-1.08	-1.87
PHE	-2.28	-1.10	-0.03	-0.01	-4.00	0.49	-0.48	-1.13	0.35	-2.84	0.47	1.04	-0.80	-0.39	-0.14	0.21	-2.24	-3.07	0.84	-2.04
GLY	1.04	0.96	0.74	0.21	0.49	1.74	-0.40	1.15	-0.21	0.98	1.82	1.00	1.02	0.02	0.15	0.33	-0.67	1.54	1.76	0.63
HIS	-0.12	-2.84	-0.19	1.12	-0.48	-0.40	-0.97	0.71	0.69	0.22	-2.75	-0.28	-0.07	-0.53	-0.85	0.85	0.68	0.50	1.55	-0.02
ILE	-0.88	-0.63	1.15	-0.93	-1.13	1.15	0.71	-1.46	0.40	-1.44	-1.66	1.22	-1.02	0.05	0.07	0.84	-0.48	-1.05	-1.76	-2.04
LYS	-0.12	-1.07	-1.13	-1.62	0.35	-0.21	0.69	0.40	0.00	0.42	0.26	-0.01	-2.23	-0.05	0.00	-0.34	-0.19	0.29	1.05	0.42
LEU	-0.62	0.22	1.25	-0.49	-2.84	0.98	0.22	-1.44	0.42	-1.83	-0.11	2.16	-0.94	0.15	-0.83	0.24	-1.67	-1.29	-0.54	0.30
MET	-0.35	-1.18	0.20	2.62	0.47	1.82	-2.75	-1.66	0.26	-0.11	-4.00	0.98	-0.23	0.89	1.49	-0.74	-0.49	-1.34	-1.98	0.33
ASN	0.84	-0.17	-0.07	0.59	1.04	1.00	-0.28	1.22	-0.01	2.16	0.98	1.02	0.02	-1.33	0.20	0.53	-1.22	1.35	-1.66	0.38
PRO	0.23	0.21	-0.55	-0.50	-0.80	1.02	-0.07	-1.02	-2.23	-0.94	-0.23	0.02	0.97	0.64	-0.52	-1.17	0.66	-0.80	-2.03	-0.53
GLN	-0.14	-0.41	1.21	0.85	-0.39	0.02	-0.53	0.05	-0.05	0.15	0.89	-1.33	0.64	0.91	0.08	1.07	1.26	0.12	0.12	-0.50
ARG	0.32	-0.92	-1.07	-0.70	-0.14	0.15	-0.85	0.07	0.00	-0.83	1.49	0.20	-0.52	0.08	0.00	0.17	-0.13	0.05	-0.47	1.58
SER	-0.53	1.02	0.87	0.76	0.21	0.33	0.85	0.84	-0.34	0.24	-0.74	0.53	-1.17	1.07	0.17	1.63	0.86	-0.54	1.81	0.89
THR	-0.94	-2.27	0.24	-0.98	-2.24	-0.67	0.68	-0.48	-0.19	-1.67	-0.49	-1.22	0.66	1.26	-0.13	0.86	-1.10	-1.00	-3.89	-0.75
VAL	-0.47	0.21	1.90	-0.24	-3.07	1.54	0.50	-1.05	0.29	-1.29	-1.34	1.35	-0.80	0.12	0.05	-0.54	-1.00	-1.66	-0.65	-0.41
TRP	-0.04	-1.06	-0.30	-1.08	0.84	1.76	1.55	-1.76	1.05	-0.54	-1.98	-1.66	-2.03	0.12	-0.47	1.81	-3.89	-0.65	-0.36	0.36
TYR	-1.55	0.27	0.34	-1.87	-2.04	0.63	-0.02	-2.04	0.42	0.30	0.33	0.38	-0.53	-0.50	1.58	0.89	-0.75	-0.41	0.36	-2.86