Random Forest Refinement of the KECSA2 Knowledge-based Scoring Function for Protein Decoy Detection

Jun Pei,[§] Zheng Zheng,^{§†} and Kenneth M. Merz Jr.^{§†}*

[§]Department of Chemistry, Michigan State University, 578 S. Shaw Lane, East Lansing,
Michigan 48824, United States

[†]Institute for Cyber Enabled Research, Michigan State University, 567 Wilson Road, East
Lansing, Michigan 48824, United States

Abstract

Knowledge-based potentials generally perform better than physics-based scoring functions in detecting the native structure from a collection of decoy protein structures. Through the use of a reference state, the pure interactions between atom/residue pairs can be obtained through the removal of contributions from ideal-gas state potentials. However, it is a challenge for conventional knowledge-based potentials to assign different importance factors to different atom/residue pairs. In this work, via the use of the ‘comparison’ concept, Random Forest (RF) models were successfully generated using unbalanced data sets that assign different importance factors to atom pair potentials to enhance their ability to identify native proteins from decoy proteins. Individual and combined data sets consisting of twelve decoy sets were used to test the performance of the RF models. We find that RF models increase the recognition of native structures without affecting their ability to identify the best decoy structures. We also created models using scrambled atom types, which create physically unrealistic probability functions, in order to test the ability of the RF algorithm to create useful models based on inputted scrambled probability functions. From this test we find that we are unable to create models that are of similar quality relative to the unscrambled probability functions. Next we created uniform probability functions where the peak positions are the same as the original, but each interaction has the same peak height. Using these uniform potentials we were able to recover models as good as the ones using the full potentials suggesting all that is important in these models are the experimental peak positions.

Introduction

According to the “thermodynamic hypothesis”.¹ the native protein in its preferred milieu should adopt a structure that has the lowest Gibbs free energy. Hence, accurate energy functions are needed to solve the protein folding, the protein structure prediction, and protein design problems. Extant scoring functions can be classified into three broad categories, physics-based, knowledge-based, and machine learning based scoring functions. Physics-based scoring functions typically employ a classical force field that represents the protein at the atomic level through the use of energy terms that represent bond, angle, torsion, van der Waals, and electrostatics interactions using relatively simple terms.³⁻⁹ Alternatively, knowledge-based scoring functions extract radial distributions of atom/residue pairs from a protein structure database and use different statistical analysis to gain “pure” interactions between atom/residue pairs.^{2, 10-40} Machine learning based scoring functions utilize different machine learning/deep learning algorithms and a large variety of information from protein structures.⁴⁹⁻⁵⁵ For example, SVMQA combines different statistical potentials, secondary structure information, and surface area as the descriptors for protein structure prediction.⁵³

Currently, knowledge-based scoring functions have been more successful than physics-based potentials in protein structure prediction.² Knowledge-based scoring function can be classified into coarse-grained residue^{17, 21-31} or atomic level^{2, 32-40} potentials due to different descriptions of the interactions present. The main challenge of building up a knowledge-based potential is how to set up an appropriate reference state. In 1985, Miyazawa and Jernigan introduced the ‘Random-mixing approximation’,²⁶ which states that, in the reference state, amino acid and solvent molecules would be uniformly distributed throughout the volume due to the absence of interactions. Many different kinds of reference states have

been constructed since then. For instance, the freely-jointed chain (FJC) model^{41,42} was applied to construct a random-walk reference state.³⁹ Knowledge-based potentials are very sensitive to the completeness of the structural database used to describe the potential and given the current status of available structural information the contribution or presence of each atom/residue pair combination is not equal for every interaction. Thus, different importance factors need to be applied to each pair wise interactions in a knowledge-based potential to reflect this deficiency. In principle, the reference state removes the contribution from an ideal-gas state potential, but the process of accurately assigning different importance factors to each atom/residue pair remains challenging.

In order to obtain an accurate potential, and estimate of the relative importance of each atom pair is needed. In this work, we focus on using the Random forest (RF) machine learning (ML) model to refine a knowledge-based potential by designating different importance factors to each atom pair in a given potential. Here, the knowledge-based and empirical combined potential, KECSA2, with 167 atom types, 2001 torsion potentials, and 14028 non-bonded atom pairs is used in the RF refinement. Our goal is to construct a ML model that can accurately differentiate the native structure from decoys using a potential refined by RF optimization. Firstly, a ‘comparison’ concept is used to change an unbalanced database to a balanced one for the construction of the RF refinement models. Secondly, the RF refinement models are applied both on individual and combined decoy sets from twelve commonly used decoy sets. As a result of this optimization all of the RF refined models recognize native structure more accurately than conventional potentials. The performance improvement realized by our RF refinement protocol can be applied to the refinement of potentials other than KECSA2. Finally, the importance of KECSA2 was examined through the use of scrambled and uniform probability functions. The comparison result suggests that

only the peak positions are important in the construction of RF models, and the RF models can be used to optimize the peak heights for different atom pairs.

Method

Building up the probability list for the proteins

If all independent pair wise probabilities with different magnitudes in a n -body system are known, the probability of the whole n -particle system can be obtained as:

$$p_n = \prod_{i,j=1, i \neq j}^n c_{ij} * p_{ij}, \quad (1)$$

where p_n is the probability of the n -particle system, c_{ij} is the scaling factor, which can be evaluated using the random forest model, of pair wise probability p_{ij} , i and j represent two different particles. Using a knowledge-based potential with pair wise independent interactions, the independent pair wise probabilities for the bond, angle, torsion, and non-bonding terms can be obtained. If the protein structure is treated as a n -particle system, the probability is:

$$p_{protein} = (\prod_{bond} c_{ij} * p_{ij})(\prod_{angle} c_{kl} * p_{kl})(\prod_{torsion} c_{mn} * p_{mn})(\prod_{nonbond} c_{pq} * p_{pq}) \quad (2)$$

$p_{protein}$ is the protein structure probability, $c_{\alpha\beta}$ and $p_{\alpha\beta}$ represent the scaling factor and the probability of atom pair α and β , the subscripts ij , kl , mn , and pq correspond to bond, angle,

torsion, and non-bonded atom pairs, respectively. In this work, we make two further assumptions: (i) $\prod_{bond} p_{bond}$ and $\prod_{angle} p_{angle}$ are similar for native and all decoys and, hence, are treated as a constant C ; (ii) the probabilities for the torsion and nonbond atom pairs are independent since a reference state is used to remove contributions from the ideal-gas state. With these assumptions, the probability of a n -atom protein can be written as:

$$p_{protein} = C(\prod_{torsion} c_{mn} * p_{mn})(\prod_{nonbond} c_{pq} * p_{pq}) \quad (3)$$

Taking the logarithm on both sides of equation 3 we get:

$$\log(p_{protein}) = \log(C) + \sum_{torsion} x_{mn} * \log(p_{mn}) + \sum_{nonbond} x_{pq} * \log(p_{pq}) \quad (4)$$

where x_{mn} and x_{pq} are the logarithm of c_{mn} and c_{pq} , respectively. A detailed potential database, KECSA2, was utilized to obtain p_{mn} and p_{pq} . Below we use O-MET-CG-MET as an example for what is involved in calculating the pair wise probability of a given protein. From KECSA2, the probability *versus* distance function, shown as a red curve in **Figure 1**, can be found. If the distance between O-MET-CG-MET in the protein is 4.5 Å, we first obtain the corresponding probabilities for the distances from 4 Å to 5 Å with an interval of 0.005 Å. Next, we take the logarithm of the average of the 201 probabilities obtained in the previous step, and use it to represent the probability at distance 4.5 Å.

Figure 1 Probability versus distance plot for atom pair O-MET__CG-MET, shaded region is the averaged region with the length of 1 Å.

Equation 5 represents a general way to obtain the probability of atom pair A-B with distance r_1 , where $KECSA2_{A-B}$ is the potential function of atom pair A-B obtained from the KECSA2 potential data base, r_{ABi} is a distance between $r_1-0.5$ and $r_1+0.5$ with an interval of 0.005 Å.

$$p_{A-B}(r_1) = \log\left[\sum_{r_1-0.5}^{r_1+0.5} KECSA2_{A-B}(r_{ABi})\right] - \log(201) \quad (5)$$

Using equation 5, the probability for each atom pair present in the protein can be obtained; for the same atom pairs, the probabilities were summed yielding the final probability. In this way, the probability list for each protein examined can be generated.

Building up the Machine Learning model

A traditional native structure recognition problem is detecting the native or most native-like structure from a collection of decoys. Many different scoring functions have been described that attempt to solve this problem. At first glance, it is hard to use unbalanced decoy sets as the training data set directly. Yet a balanced data structure can be generated if the decoy set is replaced with a ‘comparison’ data set. Instead of training ML model that focuses on directly finding the native structure from hundreds of decoys, we can create a ML model that can accurately distinguish between native and decoy structures.

In a ‘comparison’ data set for decoy detection native structure should have the highest probability, which means:

$$\begin{aligned} \log(p_{native}) - \log(p_{decoy}) &> 0 \\ \log(p_{decoy}) - \log(p_{native}) &< 0 \end{aligned} \tag{6}$$

Figure 2 shows a detailed workflow for our protocol. If a decoy set consists of one native and m decoys, for each structure, an atom pair wise descriptor (probability) can be built as described above. For each descriptor set, there are 16029 elements in total (KECSA2 has 2001 torsion atom pairs and 14028 nonbonded atom pairs, yielding $16029 = 2001+14028$). The descriptor sets are defined as the ‘Descriptor vector’ in **Figure 2**. Next,

the descriptor vector of the native minus the vector of each decoy are classified as class ‘1’, which means ‘more stable than’ since the native structure is always more stable than the decoys; the descriptor vector of each decoy minus the vector of the native is defined as class ‘0’, which represents ‘less stable than’. The resultant descriptors are described as the ‘final descriptor vector’ in **Figure 2**. In this way, equal members of class ‘0’ and class ‘1’ can be generated, which is an ideal situation for classification. Hence, a RF model can be obtained based on using those two classes. Through the use of this classification system a RF model can be generated where the relative probabilities of two proteins with the same sequence can be compared. A final descriptor vector can be generated using the descriptor vector of the first protein minus the second’s. Then the RF model can be used to predict the class for that final descriptor vector. If the prediction from the RF model is ‘1’, it means the first protein is ‘more stable than’ the second one, and if the prediction is ‘0’, the first protein is ‘less stable than’ the second one.

Figure 2. Protocol used to build up the Random Forest model.

Building up the ranking list for a decoy set

Constructing a RF model that can accurately differentiate native and decoy structures is not enough. For a native recognition blind test, in order to identify the native structure, a ranking of all structures should be generated. Thus, the RF model needs to be used to obtain the ranking list for a decoy set.

Figure 3 gives the protocol used to obtain the ranking of a decoy set with n structures. First, the probability descriptor of each protein structure can be built using the KECSA2 database. Second, a table for each structure was obtained from the probability descriptor of the individual protein structure minus the probability vectors of all the other structures. Then, the RF model is used to predict the class of each column in all tables; in other words, the RF model is used to ‘compare’ two structures. Finally, a row with length $n-1$ can be generated for each structure. The value of each column in the resultant row is either ‘0’ or ‘1’, which represents the comparison result of each structure with all other structures. The sum of the resultant row is defined as a ‘score’, which indicates if the corresponding structure is more stable than the “score” amount of decoys. In this way, the score of each structure can be generated, thereby, creating a ranking list.

Figure 3. Protocol for generating the ranking list for the Random Forest model.

Decoy sets

The decoy sets we used include the multiple decoy sets from the Decoys ‘R’ Us collection (<http://compbio.buffalo.edu/dd/download.shtml>), which include the 4state_reduced, fisa, fisa_casp3, hg_structal, ig_structal, ig_structal_hires, lattice_ssfit, lmds, and lmds_v2 decoy sets. The MOULDER decoy set was downloaded from <https://salilab.org/decoys/>; the I-TASSER decoy set-II was obtained from <https://zhanglab.ccmb.med.umich.edu/decoys/decoy2.html>; and the ROSETTA all-atom decoy set from <https://zenodo.org/record/48780#.WvtCA63MzLF>.

Our RF model was compared to the following potentials designed for decoy detection: KECSA2, GOAP,² DFIRE,⁴⁰ dDFIRE,^{37, 43} and RWplus.³⁹ The programs for these methods were downloaded from the corresponding author’s website.

Protein structure preparation

All protein structures (including both native and decoys) were converted into their biological oligomerization state and prepared with the Protein Preparation Wizard,⁴⁴ which adds missing atoms, optimizes the H-bond network, and performs energy minimization to clean up the structures for subsequent calculations. The decoy sets can be found here https://github.com/JunPei000/protein_folding-decoy-set.

The KECSA2 potential

KECSA is a potential specifically designed for protein and protein/ligand systems that was developed in our group.⁴⁵ A detailed atom type definition was used in this potential; in other words, every atom type represents a specific atom in the twenty naturally occurring amino acids. For instance, 'CA_ALA' corresponds to the alpha carbon in alanine. A detailed description of our atom typing scheme can be found here https://github.com/JunPei000/protein_folding-decoy-set.

Machine learning and validation

The `sklearn.ensemble.RandomForestClassifier` function from Scikit-learn was used to create the proposed classification model.⁴⁶ Specifically an iteration of training and testing involved: (1) the data set was randomly separated into training (80%) and test (20%) sets, (2) a grid search with five-fold cross validation was done on the training set to find the best set of

hyperparameters. The set of hyperparameters with the best performance on cross validation was retained as the final hyperparameters to be used on the test set. (3) The RF model with the best hyperparameters was tested on the test set.

(i) Individual decoy set training

For each decoy set, ten iterations described above were run creating ten sets of hyperparameters. Hyperparameters with the highest frequency are shown in **Table s1**.

(ii) Overall decoy sets training

In order to avoid duplicate protein system that existed both in the training and test set, different decoy sets were combined based on different proteins. For instance, 2cro is a decoy set that is present in both the 4state_reduced and fisa decoy sets; herein, the two 2cro sets were combined together. In total, there were 235 different decoy sets.

In overall decoy set training, the top 100 to top 500 most important features were used instead of using all 16029 descriptor elements, based on an importance feature analysis generated from the combined decoy sets. In order to obtain the overall importance feature analysis, ten importance feature analyses on the decoy subsets were generated. Specifically, for each importance feature analysis on a decoy subset, a grid search was done on 20% of the data from the combined data set, then an importance feature analysis of the best parameter set was generated. This process was repeated ten times in order to cover the whole data set, and the ten importance feature analyses were combined to obtain the overall result.

Based on this analysis, the top 100, 200, 300, 400, and 500 most important features were selected for use in the RF model. Again ten iterations were run to find the best hyperparameter set for the combined data set.

(iii) Potential analysis

(a) Scrambled potentials

Based on the overall decoy set, potential functions from both the top 100 and 500 atom pairs were scrambled to explore the importance of atom typing in the RF model. For each scrambled potential, new descriptor vectors and new final descriptor vectors shown in **Figure 2** were built based using the scrambled probability functions. Then, ten iterations were done for each scrambled data set and the resultant RF models analyzed.

(b) Uniform potential

Based on the overall decoy set, uniform potential using the top 100 and 500 features were generated where only the corresponding distance r_{max} of the lowest potential/highest probability point was employed in the RF analysis.

For torsion interactions, r_{max} values of different torsion atom pairs were obtained from the KECSA2 potential. Using equation 7, uniform probability functions for different torsion atom pairs were constructed based only on different r_{max} values. Thus, for all torsion interactions, the heights of peak in the individual probability functions were identical.

$$p_{AB} = \frac{0.05}{\sqrt{2\pi}} e^{\frac{-(r_{AB}-r_{\max_AB})^2}{0.02}} \quad (7)$$

Where AB represents the torsion atom pair, r_{AB} is the distance between atom A and B, $r_{\max_AB}$ is the corresponding distance of the lowest potential/highest probability point for atom pair AB.

For nonbond interactions, equation 8 was used to generate the uniform probability function of the different atom pairs based only on the $r_{\max}$ values from the KECSA2 nonbond interactions.

$$p_{AB} = e^{\frac{(E1*\left(\frac{3}{r_{AB}}\right)^{12} - E2*\left(\frac{3}{r_{AB}}\right)^6)}{-RT}} \quad (8)$$

Where AB is a nonbond atom pair, r_{AB} represents the distance between atom A and B, $E1$ and $E2$ are parameters for repulsion and attraction interactions, respectively. It is known that the probability function achieves a maximum at $r_{\max_AB}$; hence, the derivative of equation 8 should be equal to 0. By setting the maximum value to a constant, the parameters $E1$ and $E2$ can be obtained using equations 9 and 10.

$$\frac{(E1*3^{12}*\frac{-12}{r_{\max_AB}^{13}} - E2*3^6*\frac{-6}{r_{\max_AB}^7})}{-RT} = 0 \quad (9)$$

$$e^{\frac{(E1*\left(\frac{3}{r_{\max_AB}}\right)^{12} - E2*\left(\frac{3}{r_{\max_AB}}\right)^6)}{-RT}} = Constant \quad (10)$$

In this way, if the constant in equation 10 is the same for all nonbonded interactions, probability functions with the same heights but different peak positions can be constructed. In

the end, descriptor vectors and final descriptor vectors can be generated based on these uniform probability functions.

Results and discussion

Model accuracy (individual decoy set training)

The most important characteristic of the resultant scoring function is its ability to differentiate the native structure from decoys. **Table 1** shows the accuracy for both the RF models and traditional scoring functions. Since ten cycles of independent training and testing were performed for each decoy set, the highest, lowest, and averaged accuracy were used to represent the general performance of RF models on that specific decoy set. In this way, the performance of the RF model can be better interpreted. In general, the RF model shows higher accuracies than all traditional methods for all of the decoy sets. For some decoy sets, like fisa, ig_structal, lmds_v2, and rosetta, RF models significantly improved the averaged accuracy to nearly 1.000, and the lowest accuracy values are still higher than the best accuracies of the other scoring functions. For the other decoy sets, such as 4state_reudeced, fisa_casp3, hg_structal, ig_structal_hires, I-TASSER, lattice_ssfit, lmds, and MOULDER, the averaged accuracies from the RF models are similar to the best accuracies of traditional scoring functions, while the lowest accuracies of the RF models are similar to the accuracies of other methods. Overall, the RF models show better performance.

Table 1. Accuracy values for different models.^a

Decoy sets	RF model			KECSA2	RWplus	DFIRE	dDFIRE	GOAP
	Averaged	Highest	Lowest					

	accuracy	accuracy	accuracy					
4state_reduced	1.000	1.000	0.998	0.997	1.000	1.000	1.000	1.000
fisa	1.000	1.000	0.997	0.751	0.775	0.810	0.761	0.816
fisa_casp3	1.000	1.000	0.999	0.842	1.000	1.000	0.989	1.000
hg_structal	0.971	0.994	0.939	0.828	0.902	0.882	0.881	0.934
ig_structal	1.000	1.000	0.999	0.887	0.540	0.536	0.895	0.955
ig_structal_hires	1.000	1.000	1.000	0.953	0.580	0.567	0.942	1.000
I-TASSER	0.982	0.998	0.966	0.971	0.914	0.856	0.919	0.857
lattice_ssift	0.999	1.000	0.998	1.000	1.000	1.000	1.000	1.000
lmds	0.999	1.000	0.997	0.963	0.722	0.727	0.735	0.798
lmds_v2	0.999	1.000	0.990	0.762	0.861	0.899	0.871	0.906
MOULDER	0.988	1.000	0.969	0.829	0.982	0.985	0.982	0.991
ROSETTA	1.000	1.000	1.000	0.776	0.939	0.770	0.537	0.798

a) RF models were trained on different decoy sets.

Ranking of the native structure

Although the accuracies of RF models are higher than other methods, we still wanted to further validate their performance. Other than accuracy, another important criteria for judging a model/scoring function is whether the model/scoring function identifies the native structure as having the lowest rank. Hence, native structure ranking from the different methods were also compared. **Table 2** shows the rankings of the native structures from several different models. The highest, lowest, and averaged rankings are shown to assess the performance of RF models. For the decoy sets fisa, ig_structal, lmds, lmds_v2, and ROSETTA, the RF models substantially improves native structure ranking over the other models. In the remain decoy sets, the averaged rankings of native structures are similar to the best performance of the other scoring functions. It can be concluded that, in general, the RF

model shows a better performance in ranking the native structure over other methods we tested.

Table 2. Native structure's ranking of different models.^a

Decoy sets	RF model			KECSA2	RWplus	DFIRE	dDFIRE	GOAP
	Averaged	Highest	Lowest					
	native	native	native					
	ranking	ranking	ranking					
4state_reduced	1.70	7.00	1.00	3.29	1.00	1.00	1.00	1.00
fisa	1.50	3.00	1.00	125.50	113.5	95.75	120.25	92.75
fisa_casp3	1.00	1.00	1.00	228.60	1.60	1.60	17.20	1.00
hg_structal	2.43	5.67	1.33	5.93	3.79	4.38	4.41	2.90
ig_structal	1.01	1.08	1.00	8.03	29.7	29.98	7.57	3.79
ig_structal_hires	1.00	1.00	1.00	1.90	8.95	9.20	2.10	1.00
I-TASSER	13.26	39.17	2.83	13.71	38.13	63.25	36.16	62.89
lattice_ssift	1.05	1.50	1.00	1.38	1.00	1.00	1.00	1.00
lmds	1.05	1.50	1.00	138.91	138.90	136.50	132.2	101.10
lmds_v2	1.40	5.00	1.00	29.50	17.70	13.10	16.5	12.30
MOULDER	5.65	11.50	1.00	55.15	6.65	5.75	6.65	3.80
ROSETTA	1.00	1.00	1.00	23.33	7.07	23.84	47.16	21.07

a) RF models were trained on different decoy sets.

RMSD and TMscore of the first selected decoy structure

Although the ability to recognize the native structure as the most stable structure is a crucial characteristic of a good model/potential. For a model/potential to be useful for guiding conformation sampling, it should have a good correlation with structural quality. The RMSD and TM-score were used as two criteria for assessing the quality of each decoy structure. RMSD is the root mean squared deviation of all C α pairs of the decoy to the native structure.

TM-score⁴⁷ gives a large distance a small weight and makes the magnitude of TM-score more sensitive to the topology. **Table 3** and **Table 4** summarize the results of best model selection of different methods.

Table 3. 1st decoy's RMSD for different models.^a

Decoy sets	RF model			KECSA2	RWplus	DFIRE	dDFIRE	GOAP
	Averaged	Highest	Lowest					
	1 st decoy	1 st decoy	1 st decoy					
	RMSD	RMSD	RMSD					
4state_reduced	3.28	6.06	1.34	3.17	2.69	2.61	2.25	1.83
fisa	6.13	9.60	4.68	6.51	5.26	5.77	6.05	4.48
fisa_casp3	11.67	15.76	6.35	12.30	11.80	11.10	9.88	10.40
hg_structal	2.62	4.88	1.39	2.59	2.31	2.45	2.66	2.43
ig_structal	2.21	2.62	1.73	2.02	2.00	2.06	1.86	1.88
ig_structal_hires	2.63	4.10	1.48	2.06	2.14	2.13	2.10	2.08
I-TASSER	1.71	2.21	1.27	1.73	1.73	1.70	1.70	1.65
lattice_ssift	10.37	11.44	9.17	9.55	9.26	9.17	9.21	10.01
lmds	7.91	10.75	4.13	7.72	8.08	8.23	6.69	8.55
lmds_v2	7.60	9.38	4.46	8.01	7.74	7.82	7.67	7.36
MOULDER	9.18	12.83	6.67	10.77	9.74	9.98	10.08	9.96
ROSETTA	7.27	8.75	5.88	8.54	7.65	7.36	7.53	7.53

a) RF models were trained on different decoy sets.

Table 3 shows the 1st decoy's RMSD of RF models and against a range of available scoring functions. The RMSD values of available methods are generally within the range of lowest and highest RMSD values of the RF models for each decoy set. This means the performance of those traditional scoring functions are within the confident range of our RF models. **Table 4** shows the 1st decoy's TM-score for the RF model and against several

models; these results are similar to what we observed for the RMSD analysis. In each decoy set, the 1st decoy's TM-score is within the range of the lowest and the highest TM-score from the RF models. Considering that the independent training and testing process was done ten times, the range of lowest to highest RMSD/TM-score values show the confidence range of RF models for each decoy set. In general, the RMSD/TM-score performance of available models are within the confidence range of the RF models, and the averaged values are similar in RF models and models we tested against. In other words, the performance of the RF models against a range of models, when it comes to selecting the best decoy structure, were similar.

Table 4. 1st decoy's TM-score of different models.^a

Decoy sets	RF model			KECSA2	RWplus	DFIRE	dDFIRE	GOAP
	Averaged	Highest	Lowest					
	1 st decoy	1 st decoy	1 st decoy					
	TM-score	TM-score	TM-score					
4state_reduced	0.620	0.864	0.278	0.617	0.700	0.714	0.725	0.791
fisa	0.398	0.468	0.315	0.411	0.467	0.432	0.389	0.472
fisa_casp3	0.263	0.318	0.233	0.296	0.285	0.286	0.313	0.313
hg_structal	0.871	0.924	0.790	0.888	0.894	0.892	0.869	0.891
ig_structal	0.931	0.941	0.923	0.943	0.945	0.943	0.951	0.950
ig_structal_hires	0.936	0.950	0.914	0.939	0.948	0.947	0.949	0.951
I-TASSER	0.431	0.529	0.377	0.451	0.442	0.451	0.445	0.444
lattice_ssift	0.224	0.291	0.179	0.240	0.270	0.258	0.277	0.249
lmds	0.347	0.430	0.283	0.333	0.344	0.336	0.376	0.342
lmds_v2	0.367	0.484	0.296	0.363	0.442	0.451	0.445	0.444
MOULDER	0.429	0.555	0.211	0.394	0.426	0.418	0.416	0.422
ROSETTA	0.487	0.573	0.410	0.438	0.460	0.466	0.477	0.471

a) RF models were trained on different decoy sets.

Feature importance analysis for the overall decoy set

It is important to understand whether the better performances of RF models are due to overfitting because a large number of descriptors were used. To this end, a feature importance analysis was performed and is shown in **Figure s1**. Based on the analysis, new RF models using only the top 500 features were constructed using the previous procedure. **Table 5** shows the comparison of the accuracies between using only the top 500 and all 16029 features. In general, the highest, lowest, and averaged accuracy values of the RF models using the top 500 features for each decoy set are similar to the corresponding values of the RF models using all features. **Table s2** shows the comparison of the native structure's ranking between RF models with the top 500 and all features. The native rankings of RF models with 500 features are similar to those rankings for the RF models using the full feature set except for decoy set fisa_casp3. For fisa_casp3, the native ranking increases from 1.00 to 36.40 due to the loss of the remaining 15529 features. Among the 15529 features, they might contain important information to differentiate the native from decoys in that particular decoy set. But in general, the performance of RF models with 500 features for native structure ranking is still similar to the RF models with all features. **Table s3** and **Table s4** show the results of best decoy selection for RF models with the top 500 and whole feature set. Similar to comparisons of accuracies and native rankings, for the best decoy selection, the performance of the RF models using the top 500 features is still similar to the RF models with all features. Hence, we conclude that the better performance of RF models with all features is not simply due to overfitting and the current method is robust even in the face of potentially non-essential features.

Table 5. Comparison of accuracies of RF models with different numbers of features.^a

Decoy sets	RF model_with_IMP_500			RF_model_with_all_features		
	Averaged	Highest	Lowest	Averaged	Highest	Lowest
	accuracy	accuracy	accuracy	accuracy	accuracy	accuracy
4state_reduced	1.000	1.000	0.999	1.000	1.000	0.998
fisa	1.000	1.000	1.000	1.000	1.000	0.997
fisa_casp3	0.992	0.999	0.987	1.000	1.000	0.999
hg_structal	0.955	0.977	0.909	0.971	0.994	0.939
ig_structal	0.999	1.000	0.998	1.000	1.000	0.999
ig_structal_hires	1.000	1.000	1.000	1.000	1.000	1.000
I-TASSER	0.978	0.997	0.955	0.982	0.998	0.966
lattice_ssift	1.000	1.000	0.997	0.999	1.000	0.998
lmds	0.994	1.000	0.948	0.999	1.000	0.997
lmds_v2	1.000	1.000	1.000	0.999	1.000	0.990
MOULDER	0.989	1.000	0.974	0.988	1.000	0.969
ROSETTA	0.999	1.000	0.996	1.000	1.000	1.000

a) RF models were trained on different decoy sets.

Comparison of overall RF model with traditional scoring functions

Besides creating RF models for each individual decoy set, combined RF models using all decoy sets were also constructed. This examines the situation where in a study one might generate decoys using one method and then score them with another. There were 291 individual systems across the 12 decoy sets that were combined finally, yielding 235 different protein systems (several proteins overlapped amongst the decoy sets). In these studies, 80% of the combined data set was used as the training data to build the RF models instead of choosing several specific decoy sets (like 4state_reduced, fisa, etc.). This was done to insure that the training and testing data set covered the same feature space and had the same distribution – this is known as an independent and identical distribution (IID).⁴⁸ The feature

space and distribution of decoys from different decoy sets are different because different models were used to generate those structures.

Table 6 shows the result of comparing the overall performance of RF models with a number of available potentials. Due to the large number of descriptors, it is impossible to obtain RF models using the entire 16029 feature set. Based on the importance analysis discussed previously, instead of using all features, top 100, 200, 300, 400, and 500 features were used to build up the overall RF models on the combined decoy sets. First, all RF models with different importance features provide higher averaged accuracy values than other traditional scoring functions. Clearly, the accuracies of the RF models outperform the other conventional methods. Second, the highest rankings of the native structure from RF models are smaller than the rankings of other methods, and all of the averaged rankings of the RF models were ~10 or less, which means the RF models can identify the native structure within the top ten structures. Hence, the RF models outperform other methods on this task. Finally, the RMSD and TM-score values of conventional potentials are within the corresponding confidence range of the RF models, and those values are similar to the averaged RMSDs and TM-scores of the RF models. Both the RMSD and TM-score results suggest that the performance of the RF models is similar to other conventional potentials.

Table 6. Comparison of overall performance of RF models (with different number of features) with traditional potentials on overall data set

		nIMP _100	nIMP _200	nIMP _300	nIMP _400	nIMP _500	KECSA2	RWplus	DFIRE	dDFIRE	GOAP
Accuracy	Averaged	0.963	0.972	0.976	0.977	0.981	0.908	0.916	0.886	0.904	0.917
	Lowest	0.931	0.947	0.933	0.956	0.965	-	-	-	-	-
	Highest	0.987	0.989	0.993	0.993	0.994	-	-	-	-	-

Ranking of native structure	Averaged	10.62	8.01	8.69	6.06	7.95	25.67	23.43	31.35	26.49	23.09
	Lowest	4.86	6.05	3.92	2.36	2.77	-	-	-	-	-
	Highest	21.64	11.17	13.31	15.34	17.59	-	-	-	-	-
RMSD of 1 st selected decoy	Averaged	4.62	5.06	4.78	4.74	4.57	4.84	4.53	4.51	4.44	4.45
	Lowest	3.77	4.18	3.56	4.00	3.52	-	-	-	-	-
	Highest	5.49	5.71	5.43	5.48	5.72	-	-	-	-	-
TM-score of 1 st selected decoy	Averaged	0.634	0.598	0.604	0.629	0.614	0.610	0.622	0.623	0.625	0.674
	Lowest	0.574	0.561	0.546	0.582	0.536	-	-	-	-	-
	Highest	0.685	0.656	0.679	0.684	0.695	-	-	-	-	-

Importance of potential

Based on the previous discussion, it is clear that the RF models with KECSA2 perform the best in accuracy and ranking both on individual and overall decoy sets. This directly leads to the interesting question: does the potential plays a significant role in RF models? Here, two analyses were done to address the importance of the KECSA2 potential in the performance of the RF model. First, the probability functions of top 100 and 500 features (atom pairs) were scrambled to test if the probability functions played a role in the RF model, for example, after scrambling, the probability function of atom pair O-PRO and N-ALA might be changed to the probability function of CA-GLY and C-THR. If the KECSA2 potential plays a role in RF models, the performance of the scrambled probability functions should be worse than KECSA2. The peak positions in the probability functions represent the most favorable distances between different atom pairs found in the experimental structure database. For example, the peak position of the atom pair O-PRO and N-ALA is 3.04 Å, which means those two atoms are most stable when they form a hydrogen bond at that distance. However, after scrambling, the peak position might change to 4.51 Å (peak position of CA-GLY and C-THR), which no longer represents a hydrogen bond. Thus, the scrambled

probability function suggests that the atom pair O-PRO and N-ALA is most stable when they do not form a hydrogen bond. It is clear that the scrambled probability functions are unphysical. We expect that it would be unlikely that the RF model, as employed herein, could correct these deficiencies so, the performance of the scrambled probability function is expected to be worse than original KECSA2.

Second, the uniform probability functions (or potential functions) were built for the top 100 and 500 atom pairs to test if the KECSA2 probability peak heights (or well-depths) are important in RF models. The uniform probability functions have the same peak positions found in KECSA2, but with same heights. By doing so, the interaction strength ‘bias’ of different atom types from KECSA2 can be eliminated via use of uniform probability functions. If the KECSA2 probability peak heights (or interaction potentials) are significant, the performance of uniform potential should be worse than KECSA2.

The comparison of the result between the original KECSA2 potential, scrambled potential, and uniform potential are shown in **Table 7**. From the comparison between the original KECSA2 and the scrambled potentials, we find the accuracy of the models decreased by ~ 0.15 , which gives a clear signal that the full KECSA2 potential (well depth and energy minimum) plays a role in RF models. The comparison between the uniform and original KECSA2 potential gives an evidence of how important the r_{max} component of the KECSA2 potential is in building an effective model. For the RF models with the top 100 features, the averaged, highest, and lowest accuracies based on the original KECSA2 potential are slightly higher than the corresponding accuracies from the RF models based on uniform potentials. However, if the number of features is increased to 500, the averaged, highest, and lowest accuracies from the RF models based on the original KECSA2 are similar to the uniform

potentials. This provides strong evidence that only peak positions in the probability functions are critical in building up RF models for native protein structure detection. More importantly, the result also implies that RF models can be used to tune the height of peaks in probability functions (or the depth of potential functions) only with the information of peak positions in protein structures.

Table 7. Comparison of RF models based on different potentials

	RF model_nIMP_100			RF model_nIMP_500		
	Averaged	Highest	Lowest	Averaged	Highest	Lowest
	accuracy	accuracy	accuracy	accuracy	accuracy	accuracy
Original	0.963	0.987	0.931	0.981	0.994	0.965
Scrambled	0.822	0.854	0.805	0.827	0.864	0.799
Uniform	0.940	0.968	0.872	0.977	0.990	0.959

Conclusion

In this work, we utilized a ‘comparison’ concept to construct RF models on an unbalanced data set. With these RF models, the knowledge-based potential, KECSA2, was refined via assignment of different importance factors to different atom pairs present in the scoring function. The performance of the resultant RF models were assessed with individual and combined decoy sets and compared with the results from conventional models. We find that the RF models perform better in accuracy and native ranking and have similar performance in the RMSD and TM-score tests. In other words, the RF models improved the effectiveness of finding native structures from a set of decoys, without compromising their ability to find the best decoy structures. This RF model based refinement not only can be

used to improve the performance of KECSA2, but it can also be applied to other atom/residue pair based potentials. More importantly, we find that only peak positions in probability functions play a significant role in constructing the RF models. This result implies that, with peak position information, RF models can be created to construct probability functions (or potential functions) by tuning the height of peaks in those functions based on native and decoy protein structures.

AUTHOR INFORMATION

Corresponding Author

* E-mail: kmerz1@gmail.com

ORCID

Jun Pei: [0000-0002-0204-0896](https://orcid.org/0000-0002-0204-0896)

Zheng Zheng: [0000-0001-5221-3209](https://orcid.org/0000-0001-5221-3209)

Kenneth M. Merz Jr: [0000-0001-9139-5893](https://orcid.org/0000-0001-9139-5893)

Present Address

‡Z.Z.: School of Chemistry, Chemical Engineering and Life Science, Wuhan University of Technology, 122 Luoshi Road, Wuhan 430070, PR China

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank the high-performance computing center (HPCC) at Michigan State University for providing computational resources.

References

- (1) John, B., & Sali, A. (2003). Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Research*, *31*, 3982–3992.
- (2) Zhou, H., & Skolnick, J. (2011). GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. *Biophysical Journal*, *101*, 2043–2052.
- (3) MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., ... Karplus, M. (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *The Journal of Physical Chemistry B*, *102*, 3586–3616.

- (4) Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4, 187–217.
- (5) Weiner, S. J., Kollman, P. A., Nguyen, D. T., & Case, D. A. (1986). An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.*, 7, 230–252.
- (6) Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., ... Woods, R. J. (2005). The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26, 1668–1688.
- (7) Arnautova, Y. A., Jagielska, A., & Scheraga, H. A. (2006). A New Force Field (ECEPP-05) for Peptides, Proteins, and Organic Molecules. *The Journal of Physical Chemistry B*, 110, 5025–5044.
- (8) Ponder, J. W., & Case, D. A. (2003). Force fields for protein simulations. *Advances in Protein Chemistry*, 66, 27–85.
- (9) Jagielska, A., Wroblewska, L., & Skolnick, J. (2008). Protein model refinement using an optimized physics-based all-atom force field. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 8268–8273.
- (10) Skolnick, J. (2006). In quest of an empirical potential for protein structure prediction. *Current Opinion in Structural Biology*, 16, 166–171.
- (11) Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Current Opinion in Structural Biology*, 5, 229–235.
- (12) Jernigan RL, and B. I. (1996). Structure-derived potentials and protein simulations. *Curr Opin Struct Biol*, 6, 195–209.
- (13) Moult, J. (1997). Comparison of database potentials and molecular mechanics force fields. *Current Opinion in Structural Biology*, 7, 194–199.
- (14) Lazaridis, T., & Karplus, M. (2000). Effective energy functions for protein structure prediction. *Current Opinion in Structural Biology*, 10, 139–145.
- (15) Gohlke, H., & Klebe, G. (2001). Statistical potentials and scoring functions applied to protein-ligand binding. *Current Opinion in Structural Biology*, 11, 231–235.
- (16) Russ, W. P., & Ranganathan, R. (2002). Knowledge-based potential functions in protein design. *Current Opinion in Structural Biology*, 12, 447–452.
- (17) Buchete, N. V., Straub, J. E., & Thirumalai, D. (2004). Development of novel statistical potentials for protein fold recognition. *Current Opinion in Structural Biology*, 14, 225–232.
- (18) Poole, A. M., & Ranganathan, R. (2006). Knowledge-based potentials in protein design. *Current Opinion in Structural Biology*, 16, 508–513.
- (19) Zhou, Y., Zhou, H., Zhang, C., & Liu, S. (2006). What is a desirable statistical energy functions for proteins and how can it be obtained? *Cell Biochemistry and Biophysics*, 46, 165–174.
- (20) Ma, J. (2009). Explicit Orientation Dependence in Empirical Potentials and Its Significance to Side-Chain Modeling. *Accounts of Chemical Research*, 42, 1087–1096.
- (21) Gilis, D., Biot, C., Buisine, E., Dehouck, Y., & Rooman, M. (2006). Development of Novel Statistical Potentials Describing Cation- π Interactions in Proteins and Comparison with Semiempirical and Quantum Chemistry Approaches. *Journal of Chemical Information and Modeling*, 46, 884–893.
- (22) Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., ... Sippl, M. J. (1990). Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *Journal of Molecular Biology*, 216, 167–180.

- (23) Hoppe, C., & Schomburg, D. (2005). Prediction of protein thermostability with a direction- and distance-dependent knowledge-based potential. *Protein Science*, 14, 2682–2692.
- (24) Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, 358, 86.
- (25) Koliński, A., & Bujnicki, J. M. (2005). Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins: Structure, Function and Genetics*, 61, 84–90.
- (26) Miyazawa, Sanzo and Jernigan, R. L. (1985). Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation. *Macromolecules*, 18, 534–552.
- (27) DeBolt, S. E., & Skolnick, J. (1996). Evaluation of atomic level mean force potentials via inverse folding and inverse refinement of protein structures: atomic burial position and pairwise non-bonded interactions. *Protein Engineering*, 9, 637–655.
- (28) Zhang, C., Vasmataz, G., Cornette, J. L., & DeLisi, C. (1997). Determination of atomic desolvation energies from the structures of crystallized proteins. *Journal of Molecular Biology*, 267, 707–726.
- (29) Tobi, D., & Elber, R. (2000). Distance-dependent, pair-potential for protein folding: Results from linear optimization. *Proteins*, 41, 40–46.
- (30) Wu, Y., Lu, M., Chen, M., Li, J., & Ma, J. (2007). OPUS-Ca: A knowledge-based potential function requiring only C α positions. *Protein Science*, 16, 1449–1463.
- (31) Zhang, Y., Kolinski, A., & Skolnick, J. (2003). TOUCHSTONE II: A New Approach to Ab Initio Protein Structure Prediction. *Biophysical Journal*, 85, 1145–1164.
- (32) Sippl, M. J. (1990). Calculation of Conformational Ensembles from Potential of Mean Force. *Journal of Molecular Biology*. 213, 859-883.
- (33) Lu, H., & Skolnick, J. (2001). A distance dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, 44, 223–232.
- (34) Lu, M., Dousis, A. D., & Ma, J. (2008). OPUS-PSP: An Orientation-dependent Statistical All-atom Potential Derived from Side-chain Packing. *Journal of Molecular Biology*, 376, 288–301.
- (35) Samudrala, R., & Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Molecular Biology*, 275, 895–916.
- (36) Shen, M.-Y., & Sali, A. (2006). Statistical potential for assessment and prediction of protein structures., 15, 2507–2524.
- (37) Yang, Y., & Zhou, Y. (2008). Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins: Structure, Function and Genetics*, 72, 793–803.
- (38) Skolnick, J., Kolinski, A., & Ortiz, A. (2000). Derivation of protein-specific pair potentials based on weak sequence fragment similarity. *Proteins: Structure, Function and Genetics*, 38, 3–16.
- (39) Zhang, J., & Zhang, Y. (2010). A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLoS ONE*, 5, e15386.
- (40) Zhou, H., & Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science*, 11, 2714–2726.
- (41) M.DoI, & S.E.Edwards. (1986). The Theory of Polymer Dynamics. Clarendon Press Oxford.

- (42) Zhang, Y., Zhou, H., & Ou-Yang, Z.-C. (2001). Stretching Single-Stranded DNA: Interplay of Electrostatic, Base-Pairing, and Base-Pair Stacking Interactions. *Biophysical Journal*, 81, 1133–1143.
- (43) Yang, Y., and Zhou, Y. (2008). Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. *Protein Science*, 17, 1212–1219.
- (44) Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R., & Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, 27, 221–234.
- (45) Zheng, Z., & Merz, K. M. (2013). Development of the Knowledge-Based and Empirical Combined Scoring Algorithm (KECSA) To Score Protein–Ligand Interactions. *Journal of Chemical Information and Modeling*, 53, 1073–1083.
- (46) Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- (47) Zhang, Y., & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function and Genetics*, 57, 702–710.
- (48) Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- (49) Benkert, P., Tosatto, S. C. E., & Schomburg, D. (2008). QMEAN: A comprehensive scoring function for model quality assessment. *Proteins: Structure, Function, and Bioinformatics*, 71, 261–277.
- (50) Cao, R., Bhattacharya, D., Hou, J., & Cheng, J. (2016). DeepQA: Improving the estimation of single protein model quality with deep belief networks. *BMC Bioinformatics*, 17, 1–9.
- (51) Uziela, K., Shu, N., Wallner, B., & Elofsson, A. (2016). ProQ3: Improved model quality assessments using Rosetta energy terms. *Scientific Reports*, 6, 33509
- (52) Uziela, K., Hurtado, D. M., Shu, N., Wallner, B., & Elofsson, A. (2017). ProQ3D: Improved model quality assessments using deep learning. *Bioinformatics*, 33, 1578–1580.
- (53) Manavalan, B., & Lee, J. (2017). SVMQA: support-vector-machine-based protein single-model quality assessment. *Bioinformatics*, 33, 2496–2503.
- (54) Olechnovič, K., & Venclovas, Č. (2017). VoromQA: Assessment of protein structure quality using interatomic contact areas. *Proteins: Structure, Function and Bioinformatics*, 85, 1131–1145.
- (55) Hurtado, D. M., Uziela, K., & Elofsson, A. (2018). Deep transfer learning in the assessment of the quality of protein models. *arXiv preprint arXiv: 1804.06281*