JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

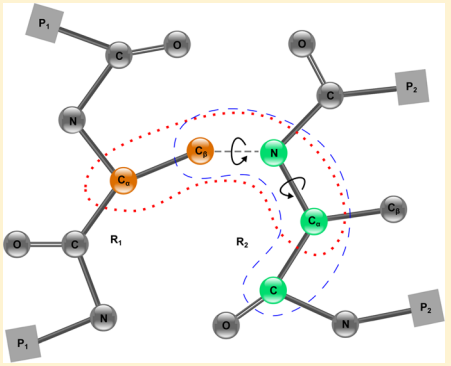
Probabilistic Models for Capturing More Physicochemical Properties on Protein−Protein Interface

Fei Guo,[†] Shuai Cheng Li,[‡] Pufeng Du,[†,‡] and Lusheng Wang*[,‡]

[†]School of Computer Science and Technology, Tianjin University, 92 Weijin Road, Nankai District, Tianjin, P.R. China

[‡]Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong

ABSTRACT: Protein−protein interactions play a key role in a multitude of biological processes, such as signal transduction, de novo drug design, immune responses, and enzymatic activities. It is of great interest to understand how proteins interact with each other. The general approach is to explore all possible poses and identify near-native ones with the energy function. The key issue here is to design an effective energy function, based on various physicochemical properties. In this paper, we first identify two new features, the coupled dihedral angles on the interfaces and the geometrical information on $\pi-\pi$ interactions. We study these two features through statistical methods: a mixture of bivariate von Mises distributions is used to model the correlation of the coupled dihedral angles, while a mixture of bivariate normal distributions is used to model the orientation of the aromatic rings on $\pi-\pi$ interactions. Using 6438 complexes, we parametrize the joint distribution of each new feature. Then, we propose a novel method to construct the energy function for protein−protein interface prediction, which includes the new features as well as the existing energy items such as dDFIRE energy, side-chain energy, atom contact energy, and amino acid energy. Experiments show that our method outperforms the state-of-the-art methods, ZRANK and ClusPro. We use the CAPRI evaluation criteria, I_{rmsd} value, and F_{nat} value. On Benchmark v4.0, our method has an average I_{rmsd} value of 3.39 Å and F_{nat} value of 62%, which improves upon the average I_{rmsd} value of 3.89 Å and F_{nat} value of 49% for ZRANK, and the average I_{rmsd} value of 3.99 Å and F_{nat} value of 46% for ClusPro. On the CAPRI targets, our method has an average I_{rmsd} value of 3.56 Å and F_{nat} value of 42%, which improves upon the average I_{rmsd} value of 4.27 Å and F_{nat} value of 39% for ZRANK, the average I_{rmsd} value of 5.15 Å and F_{nat} value of 30% for ClusPro.



INTRODUCTION

Protein−protein interactions play a key role in a multitude of biological processes, such as signal transduction, de novo drug design, and enzymatic activities. It is of great interest to understand how proteins interact with each other, and the knowledge can help us understand energetics of complexes. There exists many techniques for protein−protein interface prediction.[1−5] The general approach is to explore all possible poses and use the energy function to identify near-native poses. The problem of exploring "all" possible poses has been well-solved by some methods.[6−8] However, the results of those existing methods are not accurate. The key issue here is to design an effective energy function for identifying near-native poses, based on various physicochemical properties.

Protein−protein interactions depend on many conditions such as sequence, structure, as well as other physical and chemical properties. Hydrogen bonds and salt bridges are known to be essential in identifying binding specificity.[9] Most binding sites are hydrophobic and conserved polar residues at specific locations.[10] Secondary structure composition analysis shows that neither helices nor β-sheets are dominantly populated at interfaces.[11] Several geometrical features such as the weighted atomic packing density, relative surface area burial, and weighted hydrophobicity are the most effective features for predicting interfaces.[12] The interfaces are known to be more conserved in local surface similarities and physical-chemical properties than the rest of proteins.[13,14] ZRANK combines an atom-based potential (IFACE) with five residue-based potentials for ranking solutions.[15,16] ZRANK provides fast and accurate rescoring of models from ZDOCK.[6] ClusPro develops a fast algorithm for filtering docked conformations with good surface complementarity and ranking them based on their clustering properties.[17] RosettaDock constructs the energy function by using van der Waals energies, orientation-dependent hydrogen bonding, implicit Gaussian solvation, side-chain rotamer probabilities, and a low-weighted electrostatics energy.[18] HADDOCK makes use of the biochemical and biophysical interaction data, such as chemical shift perturbation data resulting from NMR titration experiments.[19] Force field energy evaluation, such as CHARMM,[20] Amber[21] and Gromacs,[22] are most often used to identify protein−protein interface.[23−25] There also exist knowledge-based methods[26−28] and probabilistic methods.[29,30]

In this paper, we identify two new features: the coupled dihedral angles on the interfaces and the geometrical

Received: April 16, 2014
Published: May 31, 2014

dx.doi.org/10.1021/ci5002372 | J. Chem. Inf. Model. 2014, 54, 1798−1809

information on $\pi-\pi$ interactions. Inspired by dihedral angle biases in conformations,[31] we perform a systematic study on the dihedral angles formed by the interface residues. We use a mixture of bivariate von Mises distributions to model the correlation of the coupled dihedral angles. The attraction of $\pi-\pi$ interaction depends on the relative orientation of two aromatic rings.[12,32,33] We use a mixture of bivariate normal distributions to capture the orientation of two aromatic rings on the interfaces. Using 6438 complexes, we parametrize the joint distribution of each new feature to denote local structural biases of the interface residues. These two features can be adopted to limit the search space, for discovering near-native poses.

Then, we produce a new method to construct the energy function for protein−protein interface prediction, which includes the new features as well as the existing energy items, such as dDFIRE energy, side-chain energy, atomic contact energy, and amino acid energy. The dDFIRE energy is an all-atom statistical function.[34] The side-chain atoms of interface residues are packed by SCWRL4,[35] and the side-chain energy is extracted. The problem of modeling side-chain is a well-studied.[35−37] The atomic contact energy is produced by an atomic energy measure.[38,39] The amino acid energy is constructed by probabilities of the interface residue pairs.[40] Based on these energy items, we train 79 SVM models, one for each complex in the training set. Finally, given a pair of input proteins, we define a score to select one of the 79 SVM models and use such an SVM model to further select the output poses.

We develop a software package, NewDoBi, to include all the proposed ideas. Experiments show that our method achieves better results than the state-of-the-art methods. Here, we use the CAPRI evaluation criteria, I_{rmsd} value and F_{nat} value. On Benchmark v4.0, our method has an average I_{rmsd} value of 3.39 Å and F_{nat} value of 62%, which improves upon the average I_{rmsd} value of 3.89 Å and F_{nat} value of 49% for ZRANK and the average I_{rmsd} value of 3.99 Å and F_{nat} value of 46% for ClusPro. On the CAPRI targets, our method has an average I_{rmsd} value of 3.56 Å and F_{nat} value of 42%, which improves upon the average I_{rmsd} value of 4.27 Å and F_{nat} value of 39% for ZRANK and the average I_{rmsd} value of 5.15 Å and F_{nat} value of 30% for ClusPro.

METHODS

We introduce two new features, namely, the coupled dihedral angles on the interfaces and the geometrical information on $\pi-\pi$ interactions. We study these two features through a statistical method: a mixture of bivariate von Mises distributions is used to model the correlation of the coupled dihedral angles, while a mixture of bivariate normal distributions is used to model the orientation of two aromatic rings on $\pi-\pi$ interactions. Then, we propose a new method to construct the energy function for protein−protein interface prediction, which includes the new features as well as the existing energy items.

Throughout this paper, a complex may contain several subunits and multiple binding interfaces. Each binding interface in a complex occurs in a pair of subunits. Two residues in a pair of subunits are called interface residues if any two atoms, one from each residue, interact. By interact, we mean the distance between two atoms is <6 Å.

Two Features on Protein−Protein Interface. We now described two new features: the coupled dihedral angles on the interfaces and the geometrical information on $\pi-\pi$ interactions.

Dihedral Angle. The dihedral angle biases on the interfaces are observed.[31] A dihedral angle on the interface is formed by four atoms of two interface residues from different subunits, where there exist two atoms, one from each residue, that are within distance 6 Å. Here, we focus on the following five special types of dihedral angles on the interface: $C_\alpha-C_\beta\cdots N-C_\alpha$, C − O ···N−C $_\alpha$, N−$C_\alpha-C_\beta\cdots$N, $C_\beta\cdots N-C_\alpha-$C, and O···N−$C_\alpha-$C, where − represents a bond in the amino acid, and ··· represents a pseudobond on the interface. An example of type $C_\alpha-C_\beta\cdots N-C_\alpha$ is given in Figure 1.

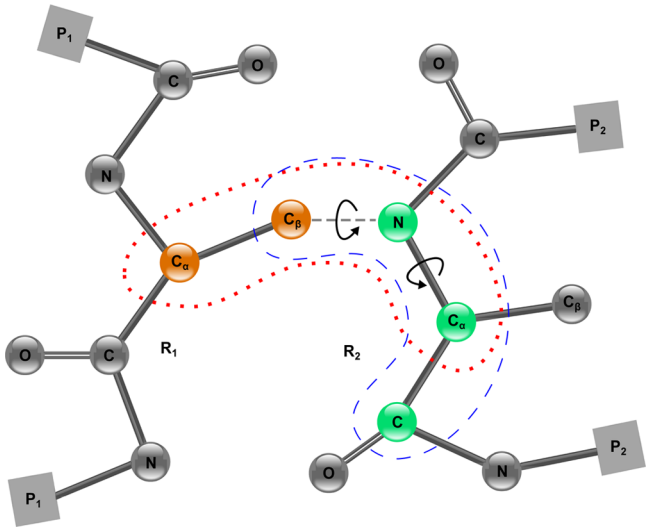


Figure 1. Dihedral angle and coupled dihedral angle pair: The dotted circle contains a dihedral angle of type $C_\alpha-C_\beta\cdots N-C_\alpha$ formed by two interface residues R_1 from and R_2 from P_2. The dashed circle contains a dihedral angle of type $C_\beta\cdots N-C_\alpha-C$. Those two dihedral angles form the coupled dihedral angle pair.

Each coupled dihedral angle pair consists of five atoms of two interface residues from different subunits. These five atoms must form two types of dihedral angles. The above five types of dihedral angles can form three types of coupled dihedral angle pairs: $C_\alpha-C_\beta\cdots N-C_\alpha$ and $N-C_\alpha-C_\beta\cdots$N, $C_\alpha-C_\beta\cdots N-C_\alpha$ and $C_\beta\cdots N-C_\alpha-C$, and C − O···N−$C_\alpha$ and O···N−$C_\alpha-$C. Scientists have studied the model of two correlated dihedral angles comprehensively. One example is the Ramachandran plot of backbone dihedral angles.[41] The Ramachandran biases are often captured by the von Mises models.[42,43] In this paper, we model the correlation of the coupled dihedral angles on the interface, through a mixture of bivariate von Mises distributions.

$\pi-\pi$ Interaction. Among 20 types of amino acid, phenylalanine, tyrosine, and tryptophan contain aromatic six-membered rings, while histidine and tryptophan contain aromatic five-membered rings. Two aromatic rings can form a $\pi-\pi$ interaction, if there exists a pair of atoms, one from each aromatic ring, within a distance of 6 Å. Here, we focus on the following three special types of $\pi-\pi$ interactions: six-membered ring vs six-membered ring, six-membered ring vs five-membered ring, and five-membered ring vs five-membered ring.

Strong attractive interactions between aromatic amino acids have been found to be important in the protein−protein interface.[32] As the aromatic rings incorporate π-systems as well as sigma-framework, the interaction between two aromatic rings depends on their relative position and orientation. When the

dx.doi.org/10.1021/ci5002372 | J. Chem. Inf. Model. 2014, 54, 1798−1809

aromatic rings are stacked in a face-to-face manner, the repulsion between the π-systems dominates. However, when the aromatic rings create a T-shape configuration, the attraction effect between the π-system and the sigma-framework dominates. For other configurations, the effect may involve both attraction and repulsion depending on the orientation of two aromatic rings, as shown in Figure 2.

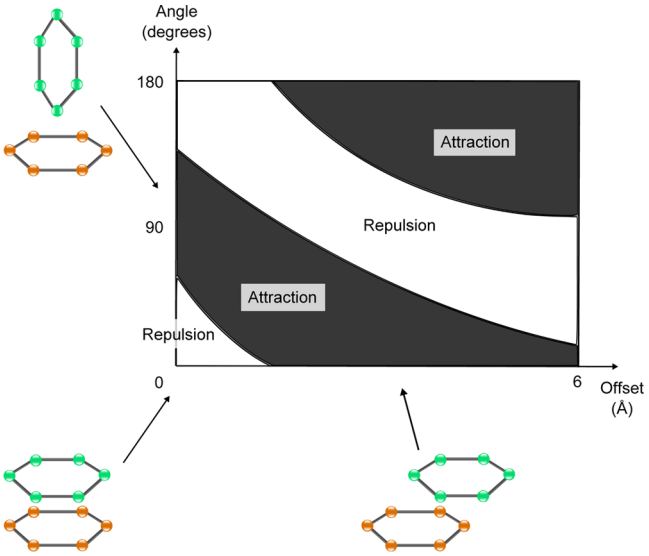


Figure 2. The $\pi-\pi$ interaction between two aromatic rings as a function of orientation.

To represent the relative position and orientation of two aromatic rings, we focus on three geometrical properties of $\pi-\pi$ interactions: the offset between the centers of two aromatic rings, the angle of two normal vectors for the planes of two aromatic rings, and the vertical separation between the centers of two aromatic rings. Scientists have study the arrangement of $\pi-\pi$ interactions in proteins. They found that particular orientations and distances of $\pi-\pi$ interactions contribute specificity to protein folding and stability to native state.[44] In this paper, we model the orientation of two aromatic rings on the interface by a mixture of bivariate normal distributions.

Distribution of Coupled Dihedral Angles. We study each dihedral angle through the cosine model, a von Mises distribution on the circle. The probability density function is defined as

$$f(\tau) = \frac{e^{\kappa\cos(\tau-\mu)}}{2\pi I_0(\kappa)} \tag{1}$$

where $\kappa > 0$ is a measure of concentration, μ is both the mode and circular mean direction, and $I_0(\kappa)$ denotes the modified Bessel function of the first kind and order 0.[45]

Furthermore, we model the correlation of the coupled dihedral angles through a bivariate von Mises distribution. The probability density function is defined as

$$f(\tau_1, \tau_2) = c(\kappa_1, \kappa_2, \kappa_3)e^{\kappa_1\cos(\tau_1-\mu)+\kappa_2\cos(\tau_2-\upsilon)+\kappa_3\cos(\tau_1-\mu-\tau_2+\upsilon)} \tag{2}$$

where μ and υ are the means for τ_1 and τ_2, κ_1 and κ_2 are their concentrations, and κ_3 is related to their correlation. A normalization constant is defined as

$$c(\kappa_1, \kappa_2, \kappa_3)^{-1} = (2\pi)^2 \times \{I_0(\kappa_1)I_0(\kappa_2)I_0(\kappa_3) + 2\sum_{p=1}^{\infty} I_p(\kappa_1)I_p(\kappa_2)I_p(\kappa_3)\} \tag{3}$$

where $I_p(\kappa)$ is the modified Bessel function of the first kind and order P.[45]

A single cosine model is unable to capture the distribution of dihedral angle, since the angular data contain two or more distinct clusters. We utilize a mixture of M cosine models to describe the angular data. The mixture of models are formulated as

$$F_M = \sum_{j=1}^{M} w_j f_j \tag{4}$$

where f_j denotes a density function, and w_j is the weight of model j with $\sum_j w_j = 1$, $1 \le j \le M$.

To estimate the parameters of the mixture model, we adopt a high-quality, nonredundant experimental data set. We select 6438 complexes from Protein Data Bank,[46] each complex consists of two or more protein subunits. These complexes are determined from X-ray data with resolution <2.2 Å. Any two complexes share no more than 30% identity to avoid overlaps in the data set. Using 6438 complexes, we apply the expectation maximization algorithm[47] to estimate the parameters of the mixture model.[48]

We use the negative logarithm of the joint distribution to denote the coupled dihedral angles on the interface. For a given pose, the effective free energy of coupled dihedral angles (τ_1, τ_2) can be calculated as

$$E(\tau_1, \tau_2) = -k_B T \sum_{\tau_1,\tau_2 \in D} \ln \frac{F(\tau_1, \tau_2)}{F(\tau_1) \times F(\tau_2)} \tag{5}$$

where $F(\tau_1,\tau_2)$ is a mixture of bivariate von Mises distributions of dihedral angles τ_1 and τ_2, $F(\tau_1)$ and $F(\tau_2)$ are mixture models of von Mises distributions of two dihedral angles, respectively. D is the set of all coupled dihedral angles on the interface.

Distribution of $\pi-\pi$ Interaction. We describe each geometrical property of $\pi-\pi$ interaction by using the normal distribution. The probability density function is defined as

$$f(x_i) = \frac{1}{\sigma_i\sqrt{2\pi}}\exp\left[-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right] \tag{6}$$

where μ_i is the mean for x_i, and σ_i is its standard deviation.[49]

The orientations of two aromatic rings can be modeled by a bivariate normal distribution, which describes the correlation of offset and angle under a fixed vertical separation. The probability density function is defined as

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left[-\frac{\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}}{2(1-\rho^2)}\right] \tag{7}$$

where μ_1 and μ_2 are the means for x_1 and x_2, σ_1 and σ_2 are their standard deviations, and ρ is the correlation between x_1 and x_2.[50]

One single model is incapable of capturing all the preferred orientations of two aromatic rings. We utilize a mixture of M models to describe the data. Using 6438 selected complexes, we

dx.doi.org/10.1021/ci5002372 | J. Chem. Inf. Model. 2014, 54, 1798−1809

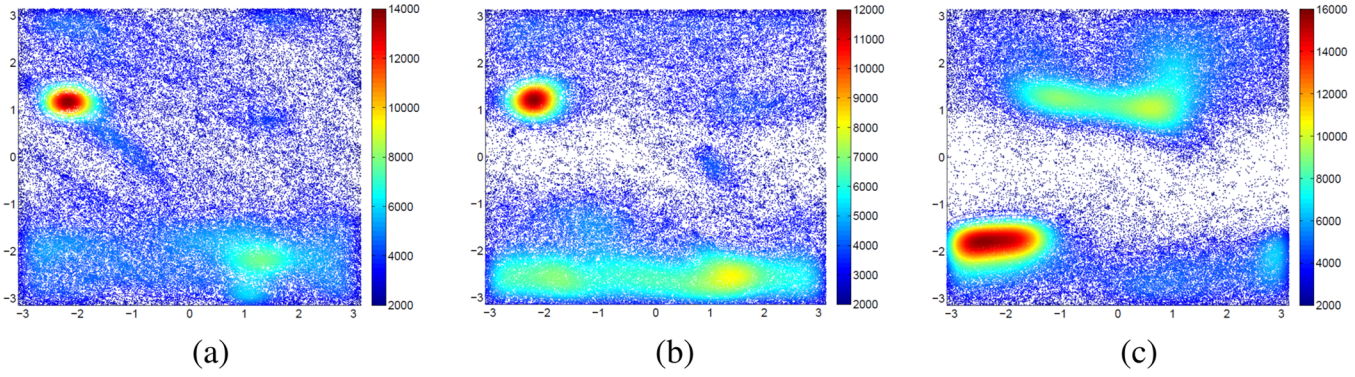


Figure 3. Density plot of the coupled dihedral angle pairs: (a) C_α−C_β···N−C_α and N−C_α−C_β···N; (b) C_α−C_β···N−C_α and C_β···N−C_α−C; (c) C − O···N−C_α and O···N−C_α−C.

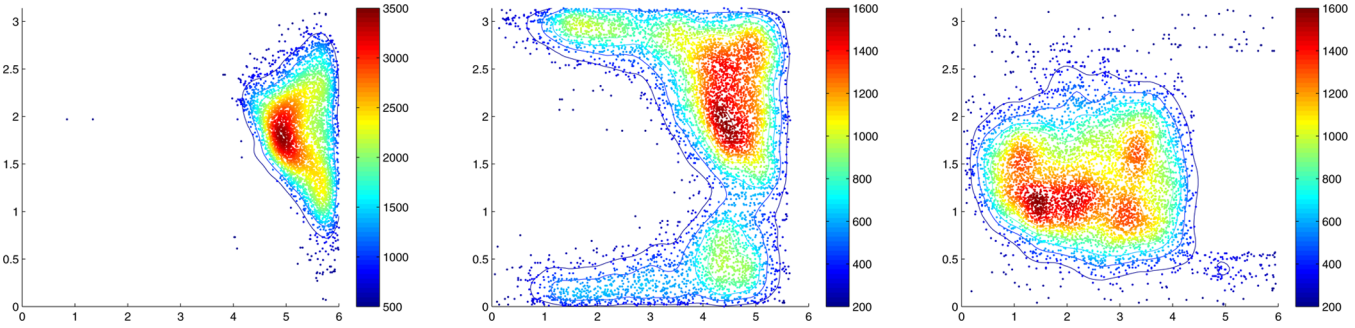


Figure 4. Density plot of offset and angle under different vertical separations for π−π interaction between six-membered rings.

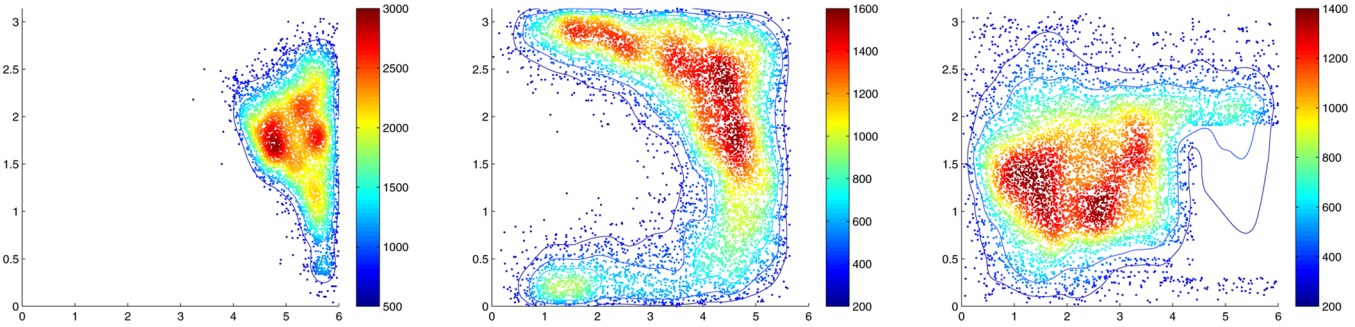


Figure 5. Density plot of offset and angle under different vertical separations for π−π interaction between six-membered ring and five-membered ring.

employ EM algorithm to estimate the parameters of the mixture model.

We use the negative logarithm of the joint distribution to denote the orientation of two aromatic rings on the interface. For a given pose, the effective free energy of two aromatic rings can be calculated as

$$S(x_1, x_2) = -k_B T \sum_{x_1, x_2 \in R} \ln \frac{F(x_1, x_2)}{F(x_1) \times F(x_2)} \quad (8)$$

where $F(x_1,x_2)$ is the bivariate normal distribution of offset and angle under a fixed vertical separation, $F(x_1)$ and $F(x_2)$ are normal distributions of offset and angle, respectively. R is the set of all the orientations of two aromatic rings on the interface.

Energy Function. We now introduce our energy function for protein−protein interface prediction. The following lists all energy items, and how they are computed: (1) The π−π interaction energy is calculated by the distributions of geometrical properties on π−π interactions of known structure complexes. (2) The dihedral angle energy is calculated by statistical analysis of dihedral angle frequencies and correlations on the interfaces. (3) The amino acid energy is constructed by probabilities of interface residue pairs. (4) The atomic contact energy is produced by an atomic energy measure.[38,39] (5) The side-chain atoms of interface residues are packed by SCWRL4,[35] and the side-chain energy is extracted. (6) The dDFIRE energy is an all-atom statistical function,[34] based on the atom distance and three angles involved in dipole−dipole interactions.

We use a linear combination of these energy items, referred to as the initial energy function, to rank the poses. The coefficient of each item is optimized by using the linear combination method.[51] We output the top 100 poses with the lowest energy values. For conformational changed structures, our method calculates a set of possibly changed conformations of interfaces. Finally, we will use the trained SVM models to further select 10 poses for each pair of input proteins.

dx.doi.org/10.1021/ci5002372 | J. Chem. Inf. Model. 2014, 54, 1798−1809

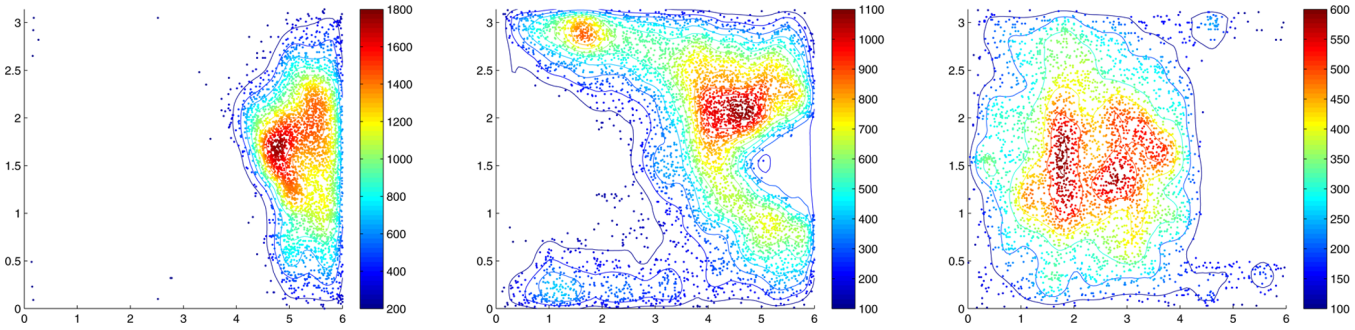


Figure 6. Density plot of offset and angle under different vertical separations for $\pi-\pi$ interaction between five-membered rings.

Table 1. Results by Our Method, ZRANK+FiberDock, and ClusPro on Benchmark v4.0

		NewDoBi			ZRANK+FiberDock			ClusPro		
subset[a]	no. of case	I_{rmsd}[b]	F_{nat}[b]	$F_{non\text{-}nat}$[b]	I_{rmsd}	F_{nat}	$F_{non\text{-}nat}$	I_{rmsd}	F_{nat}	$F_{non\text{-}nat}$
rigid body	123	2.98	68%	36%	3.31	56%	49%	3.33	55%	51%
medium difficult	29	3.46	58%	40%	4.46	39%	59%	4.71	30%	69%
difficult	24	5.42	36%	58%	6.18	28%	67%	6.53	21%	77%
overall	176	3.39	62%	39%	3.89	49%	53%	3.99	46%	58%

[a]Subset is based on the magnitude of conformational change after binding. [b]Measures according to CAPRI assessors. I_{rmsd}: interface backbone rmsd; F_{nat}: fraction of native contacts; $F_{non\text{-}nat}$: fraction of non-native contacts.

SVM Models. We use a training set consisting of 79 complexes from Dockground[52] to produce 79 SVM models, one for each of the 79 complexes, based on the six energy items. For each complex, we obtain the top 100 poses computed by our method based on the initial energy function and use these 100 poses to train an SVM model.[40] The difference here is that we have the 79 SVM models instead of one overall model.

Given a pair of input proteins P_1 and P_2, we use the following method to choose an SVM model M to further select the computed poses. For each complex corresponding to an SVM model, there are two subsequences R_1 and R_2 on the interface. We then use BLAST[53] to find the local alignments between P_i and R_j, for $i = 1, 2$ and $j = 1, 2$. Let $S(P_i, R_j)$ be the obtained alignment similarity score for P_i and R_j based on BLOSUM matrix.[54] The similarity score between an SVM model M and the pair of input proteins (P_1,P_2) is defined as

$$S(M, (P_1, P_2)) = \max\{S(R_1, P_1) + S(R_2, P_2), S(R_1, P_2) + S(R_2, P_1)\} \quad (9)$$

For each pair of input proteins P_1 and P_2, we use the SVM model M with the maximum similarity score $S(M, (P_1,P_2))$ to finally output 10 poses with the lowest energy values.

RESULTS

In this section, we have done three kinds of experiments. First, we present the statistical analysis of two new features on the interface. The statistics are carried out on 6438 complexes in the data set. Then, we examine our method on Benchmark v4.0 and the CAPRI targets. Here, we use the CAPRI evaluation criteria, I_{rmsd} value and F_{nat} value. Experiments show that our method outperforms the state-of-the-art methods, ZRANK and ClusPro. Finally, we assess the effectiveness of each energy item in our energy function.

Coupled Dihedral Angles. We analyze three types of coupled dihedral angle pairs. For each type, we model the angular data with a mixture of bivariate von Mises distributions. Most of these joint distributions cluster at two or three particular centers.

First, we represent an assessment for the dihedral angle pair $C_\alpha-C_\beta\cdots N-C_\alpha$ and $N-C_\alpha-C_\beta\cdots N$ on the interface, as shown in Figure 3a. Two distinct clusters centered at (−123.76°, 60.73°) and (−75.63°, −119.75°) can be obtained. Second, the local bias preferences of the dihedral angle pair $C_\alpha-C_\beta\cdots N-C_\alpha$ and $C_\beta\cdots N-C_\alpha-C$ are shown in Figure 3b. The distribution of these dihedral angles consists of three distinct clusters centered at (−123.76°, −136.94°), (75.63°, −136.94°) and (−123.76°, 69.91°), respectively. Third, we analyze the dihedral angle pair $C-O\cdots N-C_\alpha$ and $O\cdots N-C_\alpha-C$ on the interface, as shown in Figure 3c. The angular data contain three distinct clusters centered at (−140.37°, −119.18°), (−72.19°, 89.38°) and (46.98°, 55.01°), respectively.

$\pi-\pi$ Interaction. We analyze three types of $\pi-\pi$ interactions. For each type, we model the orientation of two aromatic rings by a mixture of bivariate normal distributions. The joint distributions of offset and angle under different vertical separations cluster at different centers.

Six-Membered Ring vs Six-Membered Ring. First, we represent an assessment for $\pi-\pi$ interaction between six-membered rings, as shown in Figure 4. For a vertical separation of 0−2.0 Å, a single cluster centered at (5.28 Å, 105.45°) for offset and angle can be obtained. The distribution of offset and angle under a vertical separation of 2.0−4.0 Å consists of three distinct clusters centered at (3.44 Å, 18.74°), (4.54 Å, 115.89°), and (2.91 Å, 162.99°), respectively. The data under a vertical separation of 4.0−6.0 Å cluster at (2.45 Å, 76.8°) for offset and angle.

Six-Membered Ring vs Five-Membered Ring. Second, the local bias preferences of $\pi-\pi$ interaction between six-membered ring and five-membered ring are shown in Figure 5. For a vertical separation of 0−2.0 Å, a single cluster at (5.22 Å, 100.78°) for offset and angle can be obtained. The distribution of offset and angle under a vertical separation of 2.0−4.0 Å consists of three distinct clusters centered at (2.58 Å, 18.46°), (4.5 Å, 96.39°), and (2.91 Å, 155.53°), respectively.

dx.doi.org/10.1021/ci5002372 | J. Chem. Inf. Model. 2014, 54, 1798−1809

Table 2. Detailed Results in Rigid-Body Group

	NewDoBi			ZRANK+FiberDock			ClusPro		
complex	I_{rmsd}[a]	F_{nat}[a]	category	I_{rmsd}	F_{nat}	category	I_{rmsd}	F_{nat}	category
1ahw	3.86	48%	acceptable	1.45	70%	medium	2.67	63%	acceptable
1bvk	2.55	72%	acceptable	3.73	42%	acceptable	5.42	19%	incorrect
1dqj	1.79	78%	medium	2.43	56%	acceptable	4.59	33%	incorrect
1e6j	1.12	95%	medium	1.73	66%	medium	2.21	76%	acceptable
1jps	9.53	3%	incorrect	1.15	79%	medium	4.57	35%	incorrect
1mlc	1.96	94%	medium	0.87	81%	high	1.90	82%	medium
1vfb	2.53	71%	acceptable	4.1	36%	incorrect	5.05	19%	incorrect
1wej	1.31	85%	medium	1.16	95%	medium	1.56	85%	medium
2fd6	2.44	85%	acceptable	6.67	3%	incorrect	4.08	43%	incorrect
2i25	1.48	81%	medium	1.74	84%	medium	3.02	59%	acceptable
2vis	1.59	90%	medium	1.51	75%	medium	3.06	58%	acceptable
1bj1	9.13	3%	incorrect	2.98	68%	acceptable	4.33	32%	incorrect
1 fsk	3.75	64%	acceptable	0.74	84%	high	1.98	74%	medium
1i9r	1.25	94%	medium	2.28	70%	acceptable	1.41	92%	medium
1iqd	2.76	74%	acceptable	1.84	76%	medium	5.17	21%	incorrect
1k4c	4.15	81%	incorrect	12.18	0%	incorrect	4.92	28%	incorrect
1kxq	1.89	77%	medium	1.16	91%	medium	3.51	52%	acceptable
1nca	1.44	88%	medium	1.09	83%	medium	1.81	76%	medium
1 nsn	5.69	29%	incorrect	1.76	68%	medium	6.89	0%	incorrect
1qfw[b]	1.89	89%	medium	5.29	11%	incorrect	4.83	22%	incorrect
1qfw[b]	2.83	61%	acceptable	4.24	41%	incorrect	3.12	55%	acceptable
2jel	1.83	89%	medium	3.43	40%	acceptable	3.44	56%	acceptable
1avx	7.74	15%	incorrect	3.95	42%	acceptable	6.81	0%	incorrect
1ay7	1.74	86%	medium	0.84	85%	high	1.50	91%	medium
1bvn	3.89	31%	acceptable	1.11	77%	medium	4.30	35%	incorrect
1cgi	2.41	62%	acceptable	2.52	54%	acceptable	4.45	35%	incorrect
1clv	1.85	72%	medium	1.44	76%	medium	1.38	88%	medium
1d6r	2.13	70%	acceptable	5.42	12%	incorrect	3.02	54%	acceptable
1dfj	1.44	82%	medium	1.75	74%	medium	3.10	53%	acceptable
1e6e	3.11	62%	acceptable	1.42	88%	medium	5.50	21%	incorrect
1eaw	1.17	87%	medium	0.97	91%	high	1.57	82%	medium
1ewy	1.17	88%	medium	1.54	87%	medium	1.48	89%	medium
1ezu	2.34	90%	acceptable	1.95	62%	medium	1.80	79%	medium
1f34	1.55	95%	medium	1.87	81%	medium	1.10	91%	medium
1fle	2.58	69%	acceptable	2.67	63%	acceptable	1.78	86%	medium
1gl1	3.69	36%	acceptable	1.47	79%	medium	1.43	83%	medium
1gxd	1.29	76%	medium	2.43	67%	acceptable	6.79	0%	incorrect
1hia	2.15	80%	acceptable	3.87	21%	acceptable	3.58	50%	acceptable
1jtg	3.20	53%	acceptable	1.33	86%	medium	1.45	82%	medium
1mah	1.45	91%	medium	1.87	78%	medium	4.69	26%	incorrect
1n8o	1.96	92%	medium	3.09	36%	acceptable	0.94	92%	high
1oc0	1.47	85%	medium	3.12	56%	acceptable	1.87	78%	medium
1oph	1.21	91%	medium	2.81	66%	acceptable	2.60	67%	acceptable
1oyv[c]	1.24	90%	medium	1.54	74%	medium	3.75	49%	acceptable
1oyv[c]	2.89	57%	acceptable	1.68	86%	medium	1.83	83%	medium
1ppe	0.67	83%	high	1.67	91%	medium	3.25	54%	acceptable
1r0r	1.56	64%	medium	1.98	87%	medium	2.62	63%	acceptable
1tmq	1.45	94%	medium	1.21	66%	medium	2.50	71%	acceptable
1udi	2.38	46%	acceptable	1.9	64%	medium	5.36	13%	incorrect
1yvb	1.82	95%	medium	1.12	77%	medium	1.12	92%	medium
2abz	1.37	96%	medium	3.81	0%	acceptable	3.65	46%	acceptable
2b42	2.05	74%	acceptable	1.07	89%	medium	0.85	95%	high
2j0t	1.46	88%	medium	2.26	73%	acceptable	1.89	75%	medium
2mta	1.41	83%	medium	1.59	63%	medium	4.10	35%	incorrect
2o8v	1.29	87%	medium	2.42	71%	acceptable	2.12	74%	acceptable
2oul	1.47	80%	medium	1.88	78%	medium	0.81	98%	high
2pcc	5.68	11%	incorrect	3.45	50%	acceptable	5.65	15%	incorrect
2sic	3.18	53%	acceptable	0.91	85%	high	7.60	0%	incorrect
2sni	2.15	85%	acceptable	2.11	67%	acceptable	3.15	58%	acceptable

dx.doi.org/10.1021/ci5002372 | J. Chem. Inf. Model. 2014, 54, 1798−1809

Table 2. continued

	NewDoBi			ZRANK+FiberDock			ClusPro		
complex	I_{rmsd}[a]	F_{nat}[a]	category	I_{rmsd}	F_{nat}	category	I_{rmsd}	F_{nat}	category
2uuy	1.25	94%	medium	1.47	78%	medium	2.64	62%	acceptable
3sgq	1.38	84%	medium	2.6	61%	acceptable	1.61	88%	medium
4cpa	2.49	75%	acceptable	1.22	94%	medium	1.77	82%	medium
7cei	1.64	82%	medium	0.95	99%	high	2.78	61%	acceptable
1a2k	3.39	53%	acceptable	3.81	46%	acceptable	4.34	32%	incorrect
1ak4	3.69	49%	acceptable	2.76	69%	acceptable	3.57	53%	acceptable
1akj	6.53	0%	incorrect	4.89	27%	incorrect	8.10	0%	incorrect
1azs	2.06	72%	acceptable	1.18	87%	medium	0.90	98%	high
1b6c	1.29	95%	medium	2.27	69%	acceptable	1.71	81%	medium
1buh	10.55	0%	incorrect	3.52	33%	acceptable	4.78	30%	incorrect
1e96	1.55	88%	medium	3.2	58%	acceptable	1.90	79%	medium
1efn	3.91	37%	acceptable	1.82	77%	medium	4.85	26%	incorrect
1f51	2.97	58%	acceptable	1.13	97%	medium	1.64	84%	medium
1fc2	1.35	81%	medium	3.26	42%	acceptable	5.82	6%	incorrect
1fcc	2.38	73%	acceptable	5.97	4%	incorrect	5.34	15%	incorrect
1ffw	2.11	88%	acceptable	2.5	66%	acceptable	2.12	77%	acceptable
1fqj	3.14	61%	acceptable	4.75	31%	incorrect	5.16	16%	incorrect
1gcq	1.44	91%	medium	3.7	46%	acceptable	3.88	45%	acceptable
1ghq	2.04	94%	acceptable	8.27	9%	incorrect	7.17	0%	incorrect
1gla	2.07	74%	acceptable	4.11	40%	incorrect	2.47	66%	acceptable
1gpw	2.97	81%	acceptable	3.42	48%	acceptable	1.52	83%	medium
1h9d	8.82	17%	incorrect	1.64	87%	medium	2.26	73%	acceptable
1hcf	1.96	95%	medium	6.84	11%	incorrect	6.93	0%	incorrect
1he1	1.20	95%	medium	2.3	72%	acceptable	1.40	89%	medium
1i4d	1.39	92%	medium	1.96	77%	medium	1.34	84%	medium
1j2j	3.75	44%	acceptable	2.18	77%	acceptable	1.57	83%	medium
1jwh	9.56	17%	incorrect	1.88	73%	medium	2.71	68%	acceptable
1k74	1.74	85%	medium	2.3	73%	acceptable	1.22	92%	medium
1kac	1.48	90%	medium	2.24	63%	acceptable	3.57	46%	acceptable
1klu	2.26	83%	acceptable	6.01	0%	incorrect	3.89	45%	acceptable
1ktz	2.51	70%	acceptable	6.06	02%	incorrect	3.83	44%	acceptable
1kxp	1.15	86%	medium	9.83	0%	incorrect	1.40	93%	medium
1ml0	1.86	81%	medium	2.23	77%	acceptable	0.84	97%	high
1ofu	7.28	0%	incorrect	1.89	74%	medium	3.67	46%	acceptable
1pvh	2.36	67%	acceptable	5.69	15%	incorrect	3.58	51%	acceptable
1qa9	1.88	82%	medium	4.15	42%	incorrect	5.61	16%	incorrect
1rlb	15.11	0%	incorrect	1.68	80%	medium	6.72	0%	incorrect
1rv6	1.65	93%	medium	6.59	11%	incorrect	4.49	35%	incorrect
1s1q	10.27	0%	incorrect	7.1	0%	incorrect	6.95	0%	incorrect
1sbb	2.15	92%	acceptable	3.92	36%	acceptable	4.92	29%	incorrect
1t6b	5.86	4%	incorrect	8.27	0%	incorrect	5.25	17%	incorrect
1us7	2.92	58%	acceptable	3.58	50%	acceptable	2.60	66%	acceptable
1wdw	3.84	52%	acceptable	1.41	84%	medium	1.49	84%	medium
1xd3	1.67	85%	medium	1.9	79%	medium	1.43	84%	medium
1xu1	2.45	84%	acceptable	11.33	0%	incorrect	8.70	0%	incorrect
1z0k	2.61	67%	acceptable	1.94	74%	medium	1.82	81%	medium
1z5y	1.98	74%	medium	1.78	79%	medium	1.50	89%	medium
1zhh	9.49	0%	incorrect	4.96	26%	incorrect	5.78	16%	incorrect
1zhi	1.39	88%	medium	3.23	65%	acceptable	1.41	83%	medium
2a5t	9.59	9%	incorrect	7.06	0%	incorrect	4.34	38%	incorrect
2a9k	3.02	59%	acceptable	8.72	0%	incorrect	3.90	43%	acceptable
2ajf	2.29	82%	acceptable	1.98	69%	medium	2.34	67%	acceptable
2ayo	1.97	82%	medium	2.02	88%	acceptable	2.71	63%	acceptable
2b4j	2.14	88%	acceptable	3.66	67%	acceptable	3.98	45%	acceptable
2btf	1.87	89%	medium	10.67	0%	incorrect	1.10	94%	medium
2fju	4.26	40%	incorrect	5.81	11%	incorrect	3.31	52%	acceptable
2g77	2.37	66%	acceptable	2.44	73%	acceptable	1.92	82%	medium
2hle	2.04	89%	acceptable	7.85	8%	incorrect	1.84	80%	medium
2hqs	1.89	94%	medium	2.67	78%	acceptable	3.91	41%	acceptable

dx.doi.org/10.1021/ci5002372 | J. Chem. Inf. Model. 2014, 54, 1798−1809

Table 2. continued

	NewDoBi			ZRANK+FiberDock			ClusPro		
complex	I_{rmsd} [a]	F_{nat} [a]	category	I_{rmsd}	F_{nat}	category	I_{rmsd}	F_{nat}	category
2oob	4.65	29%	incorrect	7.94	0%	incorrect	5.04	22%	incorrect
2oor	3.38	54%	acceptable	3.91	45%	acceptable	4.11	37%	incorrect
2vdb	2.49	82%	acceptable	8.91	0%	incorrect	3.32	56%	acceptable
3bp8	3.87	40%	acceptable	8.84	0%	incorrect	5.08	24%	incorrect
3d5s	1.69	83%	medium	1.73	82%	medium	1.37	91%	medium

[a]Measures according to CAPRI assessors. I_{rmsd}: interface backbone rmsd; F_{nat}: fraction of native contacts. [b]The first complex is 1qfw(HL:AB) and the second complex is 1qfw(IM:AB). [c]The first complex is 1oyv(B:I) and the second complex is 1oyv(A:I).

Table 3. Detailed Results in Medium-Difficulty Group

	NewDoBi			ZRANK+FiberDock			ClusPro		
complex	I_{rmsd} [a]	F_{nat} [a]	category	I_{rmsd}	F_{nat}	category	I_{rmsd}	F_{nat}	category
1bgx	3.22	86%	acceptable	7.76	6%	incorrect	8.28	0%	incorrect
1acb	1.67	87%	medium	2.79	65%	acceptable	16.2	0%	incorrect
1ijk	1.98	81%	medium	2.12	79%	acceptable	1.86	58%	medium
1jiw	1.37	89%	medium	3.28	40%	acceptable	7.38	0%	incorrect
1kkl	5.31	18%	incorrect	12.55	0%	incorrect	1.92	64%	medium
1m10	1.68	82%	medium	3.98	37%	acceptable	5.83	12%	incorrect
1nw9	1.29	88%	medium	3.85	24%	acceptable	4.87	22%	incorrect
1gp2	3.68	51%	acceptable	2.71	66%	acceptable	3.39	39%	acceptable
1grn	3.48	85%	acceptable	5.56	15%	incorrect	3.06	49%	acceptable
1he8	4.76	32%	incorrect	2.34	70%	acceptable	2.38	60%	acceptable
1i2m	7.41	6%	incorrect	3.64	33%	acceptable	3.84	39%	acceptable
1ib1	5.39	20%	incorrect	7.6	0%	incorrect	5.89	3%	incorrect
1k5d	2.64	67%	acceptable	4.94	32%	incorrect	2.51	56%	acceptable
1lfd	5.41	17%	incorrect	6.04	16%	incorrect	4.94	23%	incorrect
1mq8	1.59	80%	medium	2.25	78%	acceptable	5.92	9%	incorrect
1n2c	2.89	86%	acceptable	8.29	0%	incorrect	4.72	18%	incorrect
1r6q	1.74	90%	medium	2.72	55%	acceptable	4.80	17%	incorrect
1syx	1.81	76%	medium	5.61	5%	incorrect	5.01	19%	incorrect
1wq1	2.35	66%	acceptable	2.64	70%	acceptable	3.71	39%	acceptable
1xqs	3.45	82%	acceptable	7.53	0%	incorrect	3.06	49%	acceptable
1zm4	2.47	64%	acceptable	3.67	43%	acceptable	2.44	52%	acceptable
2cfh	3.81	43%	acceptable	1.69	85%	medium	1.53	64%	medium
2h7v	2.97	61%	acceptable	2.36	68%	acceptable	2.64	55%	acceptable
2 hk	7.93	14%	incorrect	2.38	60%	acceptable	3.05	50%	acceptable
2j7p	5.68	6%	incorrect	7.65	0%	incorrect	6.89	0%	incorrect
2nz8	4.71	32%	incorrect	1.98	73%	medium	2.87	42%	acceptable
2oza	1.74	86%	medium	3.94	31%	acceptable	8.06	0%	incorrect
2z0e	2.16	91%	acceptable	3.34	38%	acceptable	5.67	6%	incorrect
3cph	5.88	2%	incorrect	4.16	43%	incorrect	3.91	31%	acceptable

[a]Measures according to CAPRI assessors. I_{rmsd}: interface backbone rmsd; F_{nat}: fraction of native contacts.

The data under a vertical separation of 4.0−6.0 Å cluster at a single center (2.47 Å, 84.72°) for offset and angle.

Five-Membered Ring vs Five-Membered Ring. Third, we analyze $\pi-\pi$ interaction between five-membered rings, as shown in Figure 6. For a vertical separation of 0−2.0 Å, the data cluster at a single center (5.22 Å, 92.42°) for offset and angle. The distribution of offset and angle under a vertical separation of 2.0−4.0 Å consists of three distinct clusters centered at (3.88 Å, 31.43°), (2.33 Å, 158.09°), and (4.42 Å, 116.88°), respectively. For a vertical separation of 4.0−6.0 Å, a single cluster at (2.37 Å, 89.49°) for offset and angle can be obtained.

Training Set. According to CAPRI evaluation criteria,[55] three evaluation measures are commonly used in protein−protein interface prediction. A pair of residues on different sides of the interface is considered to be in contact if any of their atoms are within 6 Å. One is the fraction of native contacts F_{nat}, defined as the number of correct residue−residue contacts in the predicted complex divided by the number of contacts in the native complex. The other is the fraction of non-native contacts $F_{non\text{-}nat}$, defined as the number of incorrect residues-residue contacts in the predicted complex divided by the total number of contacts in that predicted complex. The third is the interface rmsd I_{rmsd}, defined as the rmsd between the backbone atoms of interface in the predicted structure and in the native complex, after the interface residues are superimposed. The evaluated predictions are grouped into four categories on the basis of two parameters F_{nat} and I_{rmsd}. These categories are high ($F_{nat} \geq 0.5$ and $I_{rmsd} \leq 1.0$), medium ($F_{nat} \geq 0.3$ and $1.0 \leq I_{rmsd} \leq 2.0$), acceptable ($F_{nat} \geq 0.1$ and $2.0 \leq I_{rmsd} \leq 4.0$), and incorrect ($F_{nat} < 0.1$).

We produce a new method to construct the energy function for protein−protein interface prediction. We consider the 79

dx.doi.org/10.1021/ci5002372 | J. Chem. Inf. Model. 2014, 54, 1798−1809

Table 4. Detailed Results in Difficulty Group

	NewDoBi			ZRANK+FiberDock			ClusPro		
complex	I_{rmsd}[a]	F_{nat}[a]	category	I_{rmsd}	F_{nat}	category	I_{rmsd}	F_{nat}	category
1e4k	5.98	11%	incorrect	3.66	40%	acceptable	7.07	0%	incorrect
2hmi	4.68	40%	incorrect	8.64	0%	incorrect	7.99	0%	incorrect
1f6m	5.14	30%	incorrect	7.23	5%	incorrect	8.24	0%	incorrect
1fq1	6.23	0%	incorrect	7.61	12%	incorrect	8.05	0%	incorrect
1pxv	1.97	91%	medium	3.87	47%	acceptable	3.57	56%	acceptable
1zli	5.72	17%	incorrect	8.69	0%	incorrect	9.25	0%	incorrect
2o3b	3.76	61%	acceptable	6.78	3%	incorrect	6.09	1%	incorrect
1atn	5.86	12%	incorrect	4.27	48%	incorrect	4.74	32%	incorrect
1bkd	8.12	0%	incorrect	6.38	0%	incorrect	7.33	0%	incorrect
1de4	2.94	79%	acceptable	3.77	51%	acceptable	3.17	60%	acceptable
1eer	1.29	85%	medium	3.47	49%	acceptable	4.56	30%	incorrect
1fak	5.98	17%	incorrect	7.37	2%	incorrect	5.62	12%	incorrect
1h1v	8.12	0%	incorrect	14.53	0%	incorrect	15.72	0%	incorrect
1ibr	10.36	0%	incorrect	8.86	0%	incorrect	9.83	0%	incorrect
1ira	9.17	0%	incorrect	9.18	11%	incorrect	11.13	0%	incorrect
1jk9	2.54	88%	acceptable	2.97	77%	acceptable	2.16	83%	acceptable
1jmo	3.36	70%	acceptable	6.93	4%	incorrect	5.35	13%	incorrect
1jzd	2.64	87%	acceptable	3.2	73%	acceptable	5.95	8%	incorrect
1r8s	7.82	11%	incorrect	3.03	83%	acceptable	5.67	9%	incorrect
1y64	6.59	0%	incorrect	8.53	0%	incorrect	9.73	0%	incorrect
2c0l	6.45	0%	incorrect	5.05	32%	incorrect	4.36	31%	incorrect
2i9b	9.73	10%	incorrect	7.69	8%	incorrect	3.50	55%	acceptable
2ido	2.41	85%	acceptable	3.41	56%	acceptable	3.12	69%	acceptable
2ot3	3.29	74%	acceptable	3.25	78%	acceptable	4.40	33%	incorrect

[a]Measures according to CAPRI assessors. I_{rmsd}: interface backbone rmsd; F_{nat}: fraction of native contacts.

complexes from Dockground[52] as the training set. In order to avoid overfitting, we exclude the complexes, which share more than 30% identity with the cases in the testing set. The average value of I_{rmsd} between the predicted structures and the native complexes is 1.52 Å, and the overall F_{nat} and $F_{non\text{-}nat}$ are 85% and 17%, respectively. The solutions of 79 complexes in training set are all correct: 7 high cases, 64 medium cases, and 8 acceptable cases.

Evaluation on Benchmark v4.0. In this study, NewDoBi is used to predict the protein−protein interfaces of 176 complexes in Benchmark v4.0.[56] We compare our results with ZRANK[15,16] and the external tool, FiberDock,[8] specifically designed to handle the conformation change after binding. We also compare our results with ClusPro.[17]

On Benchmark v4.0, NewDoBi produces one high-quality predictions, 65 medium-quality predictions, 68 acceptable predictions and 42 incorrect predictions. ZRANK+FiberDock produces six high-quality predictions, 48 medium-quality predictions, 66 acceptable predictions, and 56 incorrect predictions. ClusPro produces five high-quality predictions, 41 medium-quality predictions, 56 acceptable predictions, and 74 incorrect predictions. The average values of I_{rmsd} predicted by NewDoBi, ZRANK+FiberDock, and ClusPro are 3.3, 3.89, and 3.99 Å, respectively. The overall F_{nat} predicted by these three methods are 62%, 49%, and 46%, respectively. The results are shown in Table 1. The complexes are classified into three categories, according to the magnitude of conformational change after binding. In rigid-body group, the average values of I_{rmsd} predicted by NewDoBi, ZRANK, and ClusPro are 2.98, 3.31, and 3.33 Å, respectively. The overall F_{nat} predicted by these three methods are 68%, 56%, and 55%, respectively. The detailed results are shown in Table 2. In medium-difficulty group, the average values of I_{rmsd} predicted by NewDoBi, ZRANK+FiberDock, and ClusPro are 3.46 Å, 4.46 and 4.71 Å, respectively. The overall F_{nat} predicted by these three methods are 58%, 39%, and 30%, respectively. The detailed results are shown in Table 3. In difficulty group, the average values of I_{rmsd} predicted by NewDoBi, ZRANK+FiberDock, and ClusPro are 5.42, 6.18, and 6.53 Å, respectively. The overall F_{nat} predicted by these three methods are 36%, 28%, and 21%, respectively. The detailed results are shown in Table 4. NewDoBi produces 112 better predictions than ZRANK+FiberDock and 106 better predictions than ClusPro. Experiments show that our method outperforms the state-of-the-art methods, ZRANK and ClusPro.

Evaluation on CAPRI. We evaluate docking results of our method, ZRANK and ClusPro on the CAPRI targets. CAPRI[55] is a community-wide experiment to assess the capacity of protein docking methods to predict protein−protein interactions. NewDoBi produces 3 high-quality predictions, 10 medium-quality predictions, 9 acceptable predictions, and 14 incorrect predictions. ZRANK+FiberDock produces 6 high-quality predictions, 7 medium-quality predictions, 8 acceptable predictions, and 15 incorrect predictions. ClusPro produces 3 high-quality predictions, 8 medium-quality predictions, 11 acceptable predictions, and 14 incorrect predictions. The average values of I_{rmsd} predicted by NewDoBi, ZRANK +FiberDock and ClusPro are 3.56, 4.27, and 5.15 Å, respectively. The overall F_{nat} predicted by these three methods are 42%, 39%, and 30%, respectively. The results are shown in Table 5. NewDoBi produces 19 better predictions than ZRANK+FiberDock, and 24 better predictions than ClusPro. In general, docking results predicted by our method are somewhat closer to the native complexes.

Assessment of Energy Items. To assess the effectiveness of energy items, we analyze the performance with Benchmark

dx.doi.org/10.1021/ci5002372 | J. Chem. Inf. Model. 2014, 54, 1798−1809

Table 5. Results by Our Method, ZRANK+FiberDock, and ClusPro on the CAPRI Targets

	NewDoBi			ZRANK+FiberDock			ClusPro		
target	I_{rmsd}[a]	F_{nat}[a]	category	I_{rmsd}	F_{nat}	category	I_{rmsd}	F_{nat}	category
T01	4.28	12%	incorrect	8.10	7%	incorrect	12.6	0%	incorrect
T02	1.23	77%	medium	0.51	96%	high	19.0	0%	incorrect
T03	8.48	9%	incorrect	1.92	60%	medium	3.61	23%	acceptable
T04	3.98	34%	acceptable	4.56	23%	incorrect	10.5	1%	incorrect
T05	10.1	8%	incorrect	10.7	5%	incorrect	1.95	56%	medium
T06	6.51	5%	incorrect	3.10	28%	acceptable	3.68	17%	acceptable
T07	5.12	4%	incorrect	6.43	3%	incorrect	12.1	0%	incorrect
T08	6.69	8%	incorrect	1.09	47%	medium	6.50	8%	incorrect
T09	2.85	33%	acceptable	9.77	8%	incorrect	24.7	0%	incorrect
T10	3.52	29%	acceptable	5.05	11%	incorrect	6.18	5%	incorrect
T11	2.55	67%	acceptable	2.63	61%	acceptable	3.12	42%	acceptable
T12	1.55	67%	medium	0.65	84%	high	0.78	93%	high
T13	0.63	94%	high	2.38	54%	acceptable	3.98	12%	acceptable
T14	9.80	10%	incorrect	0.95	53%	acceptable	1.89	31%	medium
T15	1.40	54%	medium	0.86	91%	high	1.83	49%	medium
T17	4.96	21%	incorrect	8.12	3%	incorrect	5.70	24%	incorrect
T18	3.08	25%	acceptable	1.86	66%	medium	3.70	21%	acceptable
T19	1.74	59%	medium	10.3	2%	incorrect	2.58	29%	acceptable
T20	8.13	1%	incorrect	6.31	7%	incorrect	3.24	11%	acceptable
T21	1.39	88%	medium	3.23	56%	acceptable	2.78	67%	acceptable
T22	1.81	76%	medium	5.61	5%	incorrect	3.12	42%	acceptable
T23	1.90	61%	medium	1.34	72%	medium	4.80	16%	incorrect
T24	2.01	50%	acceptable	3.13	20%	acceptable	5.65	2%	incorrect
T25	2.13	57%	acceptable	1.51	64%	medium	1.85	65%	medium
T26	0.89	84%	high	0.33	78%	high	1.21	54%	medium
T27	1.95	60%	medium	1.86	49%	medium	3.70	21%	acceptable
T29	2.46	69%	acceptable	3.13	49%	acceptable	3.57	42%	acceptable
T30	7.81	2%	incorrect	4.84	16%	incorrect	5.40	9%	incorrect
T32	2.98	34%	acceptable	9.45	1%	incorrect	0.52	87%	high
T35	3.71	29%	acceptable	8.71	4%	incorrect	6.90	7%	incorrect
T36	3.71	29%	acceptable	3.64	37%	acceptable	6.90	7%	incorrect
T37	1.25	53%	medium	0.93	92%	high	6.89	5%	incorrect
T39	0.87	75%	high	15.6	0%	incorrect	1.60	56%	medium
T40	2.17	56%	acceptable	0.43	86%	high	1.17	62%	medium
T41	1.09	67%	medium	1.45	46%	medium	1.20	41%	medium
T42	3.70	28%	acceptable	4.13	15%	incorrect	0.91	75%	high

[a]Measures according to CAPRI assessors. I_{rmsd}: interface backbone rmsd; F_{nat}: fraction of native contacts.

Table 6. Performance of Different Energy Items on Benchmark v4.0

	I_{rmsd}[a]	F_{nat}[a]	$F_{non\text{-}nat}$[a]
cases without $\pi-\pi$ interaction energy	3.58	51%	50%
cases without dihedral angle energy	3.62	49%	53%
cases without amino acid energy	3.55	53%	50%
cases without atomic contact energy	3.51	57%	43%
cases without side-chain atoms	3.63	46%	65%
cases without dDFIRE energy	3.61	52%	49%

[a]Measures according to CAPRI assessors. I_{rmsd}: interface backbone rmsd; F_{nat}: fraction of native contacts; $F_{non\text{-}nat}$: fraction of non-native contacts.

v4.0. For evaluating the effectiveness of each item, we reoptimize the coefficients in each case with only five of six items. We re-evaluate the configurations of 176 complexes, and the results are shown in Table 6. The overall F_{nat} and $F_{non\text{-}nat}$ for cases without the $\pi-\pi$ interaction energy are 51% and 50%, for cases without the dihedral angle energy are 49% and 53%, for cases without the amino acid energy are 53% and 50%, for cases without the atomic contact energy are 57% and 43%, for cases without the side-chain energy are 46% and 65%, and for cases without the dDFIRE energy are 52% and 49%. The average values of I_{rmsd} for these six cases are less than that for the case using all energy items. In improving interface prediction, the $\pi-\pi$ interaction energy increases overall I_{rmsd} value by 0.19 Å, the dihedral angle energy increases overall I_{rmsd} value by 0.23 Å, the amino acid energy increases overall I_{rmsd} value by 0.16 Å, the atomic contact energy increases overall I_{rmsd} value by 0.12 Å, the side-chain energy increases overall I_{rmsd} value by 0.24 Å, and the dDFIRE energy increases overall I_{rmsd} value by 0.22 Å. When the $\pi-\pi$ interaction energy and the dihedral angle energy excluded, the overall I_{rmsd} values dropped significantly. This confirms our hypothesis that the physicochemical properties are the important factors to consider in interface prediction.

CONCLUSION

We first identify two new features, the coupled dihedral angles on the interfaces and the geometrical information on $\pi-\pi$ interactions. We study these two features through a statistical method: a mixture of bivariate von Mises distributions is used

dx.doi.org/10.1021/ci5002372 | J. Chem. Inf. Model. 2014, 54, 1798−1809

to model the correlations of the coupled dihedral angles, while a mixture of bivariate normal distributions is used to model the orientations of two aromatic rings on $\pi-\pi$ interactions. Using 6438 complexes, we parametrize the joint distribution of each new feature to denote local structural biases of the interface residues. The probabilistic models obtained in this study can be effectively applied in protein−protein interface prediction.

Then, we propose a new method to construct the energy function for the interface prediction, which includes the new features as well as various existing energy items, such as dDFIRE energy, side-chain energy, atomic contact energy, and amino acid energy. On Benchmark v4.0, our method has an average I_{rmsd} value of 3.39 Å and F_{nat} value of 62%, which improves upon the average I_{rmsd} value of 3.89 Å and F_{nat} value of 49% for ZRANK, and the average I_{rmsd} value of 3.99 Å and F_{nat} value of 46% for ClusPro. On the CAPRI targets, our method has an average I_{rmsd} value of 3.56 Å and F_{nat} value of 42%, which improves upon the average I_{rmsd} value of 4.27 Å and F_{nat} value of 39% for ZRANK, the average I_{rmsd} value of 5.15 Å and F_{nat} value of 30% for ClusPro.

The test sets of protein complexes and the prediction results are available for download from https://sites.google.com/site/guofeics/newdobi.

AUTHOR INFORMATION

Corresponding Author

*E-mail: cswangl@cityu.edu.hk.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work is supported by the grant from Research Grants Council of Hong Kong [project no. CityU 124512] and the grant [project no. 9610025] from City University of Hong Kong.

REFERENCES

(1) Alcaro, S.; Gasparrini, F.; Incani, O.; Caglioti, L.; Pierini, M.; Villani, C. Quasi flexible automatic docking processing for studying stereoselective recognition mechanisms. *J. Comput. Chem.* **2007**, *28*, 1119−1128.

(2) Heifetz, A.; Katchalski-Katzir, E.; Eisenstein, M. Electrostatics in protein-protein docking. *Protein Sci.* **2002**, *11*, 571−587.

(3) Andrusier, N.; Nussinov, R.; Wolfson, H. J. FireDock: Fast interaction refinement in molecular docking. *Proteins* **2007**, *69*, 139−159.

(4) Konc, J.; Janežič, D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* **2010**, *26*, 1160−1168.

(5) Fernández-Recio, J.; Totrov, M.; Abagyan, R. Identification of protein-protein interaction sites from docking energy landscapes. *J. Mol. Biol.* **2004**, *335*, 843−865.

(6) Chen, R.; Li, L.; Weng, Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **2003**, *52*, 80−87.

(7) Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.* **2005**, *33*, 363−367.

(8) Mashiach, E.; Nussinov, R.; Wolfson, H. J. FiberDock: flexible induced-fit backbone refinement in molecular docking. *Proteins* **2009**, *78*, 1503−1519.

(9) Xu, D.; Tsai, C. J.; Nussinov, R. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng.* **1997**, *10*, 999−1012.

(10) Ma, B.; Elkayam, T.; Wolfson, H.; Nussinov, R. Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 5772−5777.

(11) Ansari, S.; Helms, V. Statistical analysis of predominantly transient protein-protein interfaces. *J. Comput. Chem.* **2005**, *61*, 344−355.

(12) Cho, K.; Kim, D.; Lee, D. A feature-based approach to modeling protein-protein interaction hot spots. *Nucleic Acids Res.* **2009**, *37*, 2672−2687.

(13) Konc, J.; Janežič, D. Protein-protein binding-sites prediction by protein surface structure conservation. *J. Chem. Inf. Model.* **2007**, *47*, 940−944.

(14) Li, Y.; Liu, Z.; Han, L.; Li, C.; Wang, R. Mining the characteristic interaction patterns on protein-protein binding interfaces. *J. Chem. Inf. Model.* **2013**, *53*, 2437−2447.

(15) Pierce1, B.; Weng, Z. A combination of rescoring and refinement significantly improves protein docking performance. *Proteins* **2008**, *72*, 270−279.

(16) Vreven, T.; Hwang, H.; Weng, Z. Integrating atom-based and residue-based scoring functions for protein-protein docking. *Proteins* **2011**, *20*, 1576−1586.

(17) Comeau, S. R.; Gatchell, D. W.; Vajda, S.; Camacho, C. J. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics* **2004**, *20*, 45−50.

(18) Schueler-Furman, O.; Wang, C.; Baker, D. Progress in protein-protein docking: atomic resolution predictions in the CAPRI experiment using RosettaDock with an improved treatment of side-chain flexibility. *Proteins* **2005**, *60*, 187−194.

(19) Dominguez, C.; Boelens, R.; Bonvin, A. M. J. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **2003**, *125*, 1731−1737.

(20) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.

(21) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general AMBER force field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.

(22) Lindahl, E.; Hess, B.; Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model* **2001**, *7*, 306−317.

(23) Grosdidier, A.; Zoete, V.; Michielin, O. Fast docking using the CHARMM force field with EADock DSS. *J. Comput. Chem.* **2011**, *32*, 2149−2159.

(24) Wu, G.; Robertson, D. H.; Brooks, C. L.; Vieth, M. Detailed analysis of grid-based molecular docking: a case study of CDOCKER-A CHARMm-based MD docking algorithm. *J. Comput. Chem.* **2003**, *24*, 1549−1562.

(25) Carl, N.; Hodošček, M.; Vehar, B.; Konc, J.; Brooks, B. R.; Janežič, D. Correlating protein hot spot surface analysis using ProBiS with simulated free energies of protein-protein interfacial residues. *J. Chem. Inf. Model.* **2012**, *52*, 2541−2549.

(26) Clark, L. A.; van Vlijmen, H. W. A knowledge-based forcefield for protein-protein interface design. *Proteins* **2008**, *70*, 1540−1550.

(27) Liu, S.; Vakser, I. A. DECK: Distance and environment-dependent, coarse-grained, knowledge-based potentials for protein-protein docking. *BMC Bioinformatics* **2011**, *12*, 1−7.

(28) Summa, C. M.; Levitt, M. Near-native structure refinement using in vacuo energy minimization. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 3177−3182.

(29) Glaser, F.; Steinberg, D. M.; Vakser, I. A.; Ben-Tal, N. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins* **2001**, *43*, 89−102.

(30) Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.* **2003**, *331*, 281−299.

(31) Ghose, A. K.; Crippen, G. M. Use of physicochemical parameters in distance geometry and related three-dimensional quantitative structure-activity relationships: a dsemonstration using

escherichia coli dihydrofolate reductase inhibitors. *J. Med. Chem.* **1985**, *28*, 333−346.
(32) Hunter, C.; Sanders, J. The nature of π-π interactions. *J. Am. Chem. Soc.* **1990**, *112*, 5525−5532.
(33) Burley, S. K.; Petsko, G. A. Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science* **1985**, *229*, 23−28.
(34) Yang, Y.; Zhou, Y. Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins* **2008**, *72*, 793−803.
(35) Krivov, G. G.; Shapovalov, M. V.; L, D. R. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* **2009**, *77*, 778−795.
(36) Xu, J.; Berger, B. Fast and accurate algorithms for protein side-chain packing. *J. ACM* **2006**, *53*, 533−557.
(37) Brown, J. B.; Bahadur, D.; Tomita, E.; Akutsu, T. Multiple methods for protein side chain packing using maximum weight cliques. *Genome Inf.* **2006**, *3*, 191−200.
(38) Zhang, C.; Vasmatzis, G.; Cornette, J. L.; DeLisi, C. Determination of atomic desolvation energies from the structures of crystallized protein. *J. Mol. Biol.* **1997**, *267*, 707−726.
(39) Zhang, C. Extracting contact energies from protein structures: a study using a simplified model. *Proteins* **1998**, *31*, 299−308.
(40) Guo, F.; Li, S. C.; Ma, W. J.; Wang, L. *RECOMB* **2013**, *7821*, 58−74.
(41) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **1963**, *7*, 95−99.
(42) Mardia, K.; Taylor, C.; Subramaniam, G. Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics* **2007**, *63*, 505−512.
(43) Li, S. C.; Bu, D.; Xu, J.; Li, M. Fragment-HMM: A new approach to protein structure prediction. *Protein Sci.* **2008**, *17*, 1925−1934.
(44) Misura, K. M.; Morozov, A. V.; Baker, D. Analysis of anisotropic side-chain packing in proteins and application to high-resolution structure prediction. *J. Mol. Biol.* **2004**, *342*, 651−664.
(45) Abramowitz, M. *Handbook of mathematical functions*, National Bureau of Standards; Dover Publications: Mineola, NY, 1965.
(46) Bernstein, F.; Koetzle, T.; Williams, G.; Meyer, E., Jr.; Brice, M.; Rodgers, J.; Kennard, T.; Shimanouchi, O.; Tasumi, M. The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **1977**, *112*, 535−542.
(47) McLachlan, G. F.; Krishnan, T. *The EM algorithm and extensions*, A Wiley-Interscience Publication; John Wiley: New York, 2008.
(48) Mardia, K. V.; Taylor, C. C.; Subramaniam, G. K. Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics* **2007**, *63*, 505−512.
(49) Trumpler, R.; Weaver, H. *Statistical astronomy*; Dover Publications: Mineola, NY, 1962.
(50) Hamedani, G.; Tata, M. On the determination of the bivariate normal distribution from distributions of linear combinations of the variables. *Am. Math. Monthly* **1975**, *82*, 913−915.
(51) Guo, F.; Li, S. C.; Wang, L. P-Binder: a system for the protein-protein binding sites identification. 2012.
(52) Liu, S.; Gao, Y.; Vakser, I. Dockground protein-protein docking decoy set. *Bioinformatics* **2008**, *24*, 2634−2635.
(53) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403−410.
(54) Henikoff, S.; Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 10915−10919.
(55) Janin, J.; Henrick, K.; Moult, J.; Eyck, L. T.; Sternberg, M.; Vajda, S.; Vakser, I.; Wodak, S. CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins* **2003**, *52*, 2−9.
(56) Hwang, H.; Vreven, T.; Janin, J.; Weng, Z. Protein-protein docking benchmark version 4.0. *Proteins* **2010**, *78*, 3111−3114.