JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Binding Affinity Prediction for Protein−Ligand Complexes Based on β Contacts and B Factor

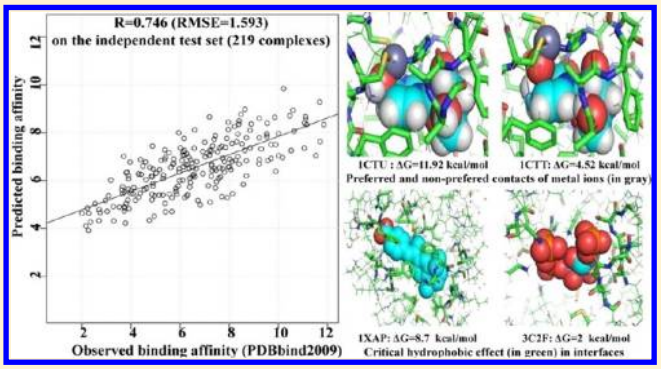
Qian Liu,[†] Chee Keong Kwoh,[‡] and Jinyan Li*,[†]

[†]Advanced Analytics Institute and Center for Health Technologies, University of Technology, Sydney, Sydney, New South Wales, NSW 2007 Australia

[‡]Bioinformatics Research Center, School of Computer Engineering, Nanyang Technological University, 639798 Singapore

Supporting Information

ABSTRACT: Accurate determination of protein−ligand binding affinity is a fundamental problem in biochemistry useful for many applications including drug design and protein−ligand docking. A number of scoring functions have been proposed for the prediction of protein−ligand binding affinity. However, accurate prediction is still a challenging problem because poor performance is often seen in the evaluation under the leave-one-cluster-out cross-validation (LCOCV). We introduce a new scoring function named B2BScore to improve the prediction performance. B2BScore integrates two physicochemical properties for protein−ligand binding affinity prediction. One is the property of β contacts. A β contact between two atoms requires no other atoms to interrupt the atomic contact and assumes that the two atoms should have enough direct contact area. The other is the property of B factor to capture the atomic mobility in the dynamic protein−ligand binding process. Tested on the PDBBind2009 data set, B2BScore shows superior prediction performance to existing methods on independent test data as well as under the LCOCV evaluation framework. In particular, B2BScore achieves a significant LCOCV improvement across 26 protein clusters—a big increase of the averaged Pearson's correlation coefficients from 0.418 to 0.518 and a significant decrease of standard deviation of the coefficients from 0.352 to 0.196. We also identified several important and intuitive contact descriptors of protein−ligand binding through the random forest learning in B2BScore. Some of these descriptors are closely related to contacts between carbon atoms without covalent-bond oxygen/nitrogen, preferred contacts of metal ions, interfacial backbone atoms from proteins, or π rings. Some others are negative descriptors relating to those contacts with nitrogen atoms without covalent-bond hydrogens or nonpreferred contacts of metal ions. These descriptors can be directly used to guide protein−ligand docking.



INTRODUCTION

Protein−ligand binding is a critical interaction in biological systems. A ligand, usually a small molecule, can bind to a biomolecule such as a protein to serve for a special biological function. Ligand binding to a protein may alter the conformational state of the protein to change the functional state of the protein. For example, an agonist (a ligand) binding to a protein adapts the function of the protein to trigger a physiological response. Binding affinity, the strength of the binding measured in numeric value, is a key descriptor for protein−ligand binding interfaces. In biochemistry, binding affinity is calculated according to dissociation constant and the relative concentration of ligands at the protein binding site. Binding affinity calculation and prediction has many applications such as structure-based drug design. It is also a compulsory and difficult step in the protein−ligand docking process after the identification of the binding mode of ligands to target receptors.

Numerous computational methods have been proposed for protein−ligand binding affinity prediction. These methods can be categorized into three groups:[1−4] those using force-field based scoring functions, those with knowledge based scoring functions, and those with empirical based scoring functions. Force-field based scoring functions make use of established mathematical terms and parameters derived from experimental outcomes or high-level quantum mechanical calculations. The most often used terms include the van der Waals interactions computed with a Lennard-Jones potential and the electrostatic interactions with the Coulomb's law. Force-field based scoring functions had been thoroughly investigated by AUTODOCK,[5,6] GOLD,[7,8] DOCK,[9] and CHARMm.[10] Knowledge-based scoring functions take the advantage of a large set of protein−ligand quaternary structures to derive pairwise potential knowledge, such as the widely used distance-dependent atomic pairwise potentials. The knowledge is then converted into the preference scores of atomic pairs to biological binding over the reference state. The sum of these scores in a protein−ligand complex is transformed into the

Received: July 30, 2013
Published: November 5, 2013

pseudoenergy of this complex typically through an inverse Boltzmann law. Scoring functions, such as PMF,[11−14] DrugScore,[15,16] ASP,[17] and ITScore,[18,19] fall into this group. Empirical scoring functions employ interaction terms directly from the predicted complex. The relationship between these terms and the binding affinity is obtained by machine-learning algorithms or regression on a training data of protein−ligand quaternary structures with known binding affinity. The increasing number of protein−ligand complexes and their experimental binding affinity make a number of empirical scoring functions available,[2,20−26] including SCORE,[27] X-CSCORE,[28] LigScore,[29] Glide,[30] RFscore,[4] and CScore.[31]

However, the performance of binding affinity prediction is poor when the test complexes have a low protein sequence similarity with the proteins in the training data.[2,26] We propose a new empirical based scoring function, named B2Bscore, to improve this performance. The main idea of B2Bscore is based on two physicochemical properties of protein−ligand interactions: B factor and β contacts, which have not been used in the affinity prediction before. B factor measures the mobility and flexibility of dynamic atoms in proteins, which is essential to determine the proteins' behavior and functions. β contacts are a small fraction of distance-based contacts.[32] A β contact between two atoms requires that there is no other atom between them. Because of this restriction, a number of unimportant, interrupted distance-cutoff contacts are excluded from β contacts. In B2Bscore, β contacts are integrated by a vector representing for the knowledge from the contacting atoms and their pairs in a protein−ligand binding interface, and the values in the vector are calculated using normalized B factor (lower B factor, larger value, or vice versa). Then B2Bscore uses a random forest algorithm to learn the relation between the binding affinity and the contacting knowledge in protein−ligand binding interfaces.

We evaluate the performance of our B2Bscore method under the leave-cluster-out cross-validation (LCOCV) framework.[26] LCOCV requires that none of the proteins in the test data set has a high sequence similarity to any protein in the training set. Prediction performance under such a requirement is very useful for drug design against new target proteins,[33] because a number of affinity-unknown complexes of low protein sequence similarity to those in known databases may actually have a strong binding affinity. We also evaluate our method on independent data sets as traditionally followed in this research field. An independent test data set is a subset of protein−ligand complexes of known binding affinity. The division of training and test data sets is carried out by random selection or a careful design[34] such that the protein sequence similarity between the test and training data set is not necessary to be low. Thus, it is possible that a protein in the test data set is highly similar to a protein in the training data set.[4]

Our evaluation was conducted on the 2009 version of the PDBbind data set (PDBbind2009 for short)[34] to compare with existing scoring functions and to show the advantages of our B2Bscore method in binding affinity prediction. We also report some constructive or potentially destructive contact descriptors extracted from the random forest learning model for characterizing protein−ligand binding interfaces. Our method and results will be useful in future applications such as protein−ligand docking.

RESULTS AND DISCUSSION

Our results are presented mainly in two parts. The first part reports our B2Bscore method's prediction performance when tested on the core set (CS, the independent data set) of PDBbind2009 as well as its performance when tested under the leave-cluster-out cross-validation (LCOCV) requirement on the refined set (RS) of PDBbind2009. In the second part, we report novel atomic contacts learned by random forest from the protein−ligand binding interface data.

Binding Affinity Prediction for Protein−Ligand Complexes. *Results on the Independent Data Set.* Following the traditional evaluation approach,[2,4,31] B2Bscore was tested on the independent data set CS of PDBbind2009 when trained on $RS - CS$ (the protein−ligand complexes in RS but not in CS). The result is shown in Figure 1 and Table 1. On CS, B2Bscore

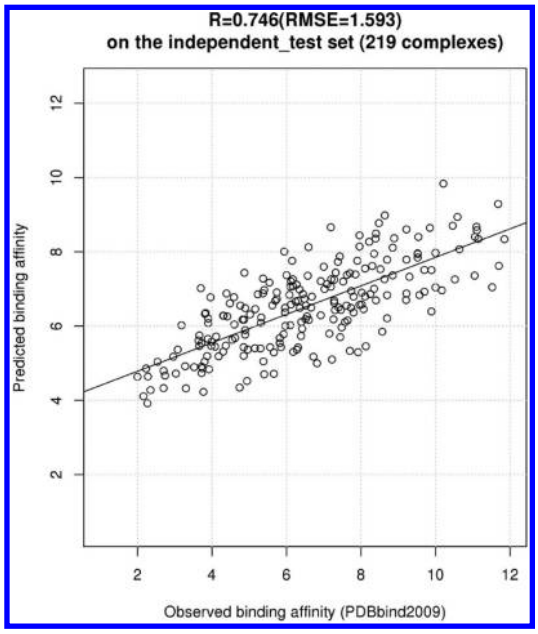


Figure 1. The prediction performance of B2Bscore on the core set of PDBbind2009.

achieves a high Pearson correlation coefficient $R_{PCC} = 0.746$ and a low root-mean-square error (RMSE) 1.593. This performance is better than the performance of all of the literature scoring functions, including RFscore[4] ($R_{PCC} = 0.728$), HSScore ($R_{PCC} = 0.604$), HMScore ($R_{PCC} = 0.599$), and HPScore ($R_{PCC} = 0.598$) in XScore;[28] AutoDock ($R_{PCC} = 0.499$),[5,6] AutoDock Vina ($R_{PCC} = 0.549$),[35] ChemScore ($R_{PCC} = 0.482$), ASP ($R_{PCC} = 0.444$), and GoldScore ($R_{PCC} = 0.326$) in GOLD;[7,8] and the scoring functions in ref 2.

Results under the Leave-One-Cluster-out Cross-Validation (LCOCV). To avoid the effect of high sequence similarity in RS, B2Bscore is also evaluated under LCOCV, for example tested on every protein family (in Table 2) after trained on the other complexes. The performance of B2Bscore under LCOCV is shown in Table 2 in comparison with the performance[26] achieved by RFscore.[4] The super performance of B2Bscore can be demonstrated over the following four aspects. First, B2Bscore has a larger average $R_{PCC} = 0.518$ over the 26 protein families and a smaller standard deviation ($\delta = 0.196$). The R_{PCC} is improved by 24% comparing with RFscore, and the δ is only 55.7% of δ in ref 26. Second, B2Bscore also has a larger average weighted $R_{PCC} = 0.540$ vs 0.456 of RFScore.

dx.doi.org/10.1021/ci400450h | J. Chem. Inf. Model. 2013, 53, 3076−3085

Table 1. Prediction Performance on the Independent Data Set of PDBbind2009

scoring function[a]	R_{PCC}
B2Bscore	0.746
RF-Score	0.728
X-Score	
HSScore	0.604
HMScore	0.599
HPScore	0.598
Autodock Vina	0.549
Autodock	0.499
GOLD	
ChemScore	0.482
ASP	0.444
GoldScore	0.326
Kramer *et al.*[b]	
ddPLATp+MOE	0.693
ddPLATp	0.671
ddPLEp	0.640
MOE	0.583

[a]The performance of existing methods is borrowed from ref 31 because the training data and testing data here are the same as those in ref 31. [b]The performance of a standard cross-validation on a subset of PDBbind2009 with 1387 complexes after the removal of those protein−ligand complexes with larger weight, those more than 20 donors/acceptors, those more than one P atom, or those with reported errors by RDkit.

Third, B2Bscore only has one negative R_{PCC} and another R_{PCC} less than 0.2, while RFScore has four negative R_{PCC} and another four R_{PCC} less than 0.2. The results imply a strong robustness of B2Bscore across the 26 protein families. Fourthly, the improvement of B2Bscore over RFscore increases from the last third protein family (0.025 R_{PCC} improvement) to the last family (0.085 R_{PCC} improvement). In particular, the biggest improvement is for the most diverse protein families 'singletons'. This result agrees with the improvement of overall performance of B2Bscore under LCOCV. All of these comparative results support that B2Bscore is able to predict binding affinity of protein−ligand complexes which have low sequence similarity to those in training data.

Important Descriptors in Protein−Ligand Binding Interfaces. Each descriptor in protein−ligand binding interfaces is also investigated according to its importance score in random forest. A larger score suggests that a contact-based descriptor should contribute to protein−ligand binding remarkably. Evaluated on *RS*, the importance of several top-ranked descriptors is shown in Figure 2, and their meanings are presented in Table 3. Several kinds of important contact knowledge are discussed below. Most of this knowledge is consistent with the review in ref 36.

Top First Contributor to Protein−Ligand Binding Affinity: Contact Descriptor of Hydrophobic Effect. As shown in Figure 2, the most important descriptor is the cross-interface β contacts among carbon/sulfur/phosphorus (CSP for short)

Table 2. LCOCV Performance

			our method		RFscore[a]	
cluster name	cluster id	number in cluster	RMSE	R_{PCC}	RMSE	R_{PCC}
HIV protease	A	188	1.688	0.381	1.910	0.110
trypsin	B	74	1.031	0.679	1.040	0.730
carbonic anhydrase	C	53	1.629	0.642	1.680	0.560
thrombin	D	57	2.169	0.506	2.030	0.370
PTP1B	E	32	1.278	0.523	1.020	0.630
factor Xa	F	32	2.098	0.335	1.760	0.190
urokinase	G	29	1.107	0.685	0.950	0.780
transporters	H	29	1.150	0.495	1.170	−0.120
PKA	I	17	1.316	0.604	1.260	0.540
beta-glucosidase	J	17	1.582	0.479	1.130	0.590
antibodies	K	16	1.550	0.555	1.570	0.580
casein kinase II	L	16	0.955	0.695	1.100	0.440
ribonuclease	M	15	1.186	0.277	1.200	0.180
thermolysin	N	14	1.067	0.698	1.090	0.680
CDK2 kinase	O	13	1.211	0.499	1.110	0.640
glutamate receptor 2	P	13	0.940	−0.015	1.160	−0.200
P38kinase	Q	13	0.870	0.599	0.590	0.790
beta-secretase 1	R	12	1.486	0.958	1.510	0.930
tRNA-guanine transglycosylase	S	12	1.211	0.118	1.080	0.120
endothiapepsin	T	11	0.961	0.699	1.340	0.600
alpha-mannosidase 2	U	10	1.654	0.479	1.880	−0.170
carboxypeptidase A	V	10	2.030	0.404	1.710	0.780
penicillopepsin	W	10	1.670	0.471	2.220	−0.420
complexes 4−9	X	387	1.616	0.585	1.630	0.560
complexes 2−3	Y	340	1.551	0.581	1.610	0.530
singletons	Z	321	1.622	0.525	1.750	0.440
overall	average			0.518		0.418
	standard deviation (δ)			0.196		0.352
	weighted average			0.540		0.456

[a]The performance from ref 26 is achieved using RFscore with distance-cutoff 12 Å which is recommended in ref 4 and used in Table 1.

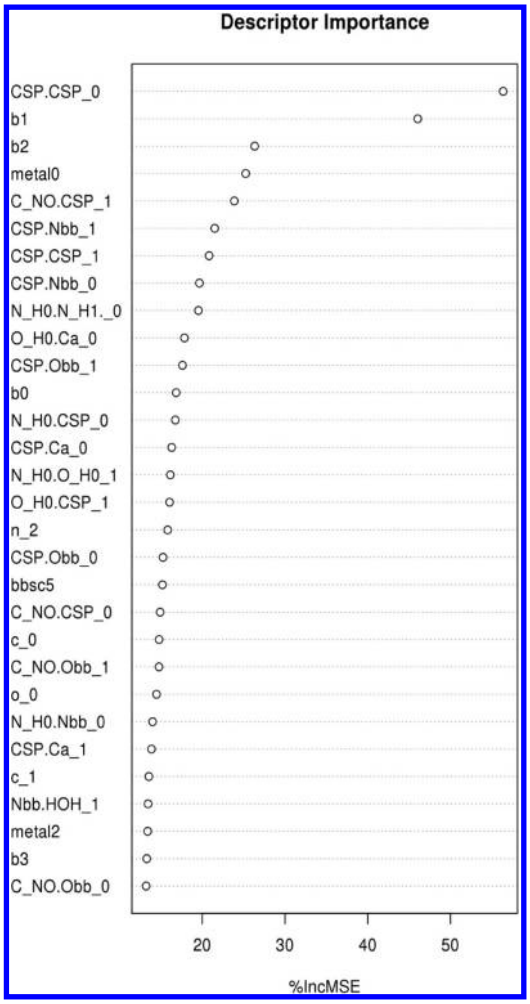


Figure 2. The descriptor's importance in random forest. '%IncMSE" indicates the increase of the mean standard error after the permutation of the descriptors. The meanings of top 15 important descriptors are in Table 3.

which have no covalent-bond with oxygen/nitrogen (CSP.CSP_0 for short). The permutation of this descriptor increases MSE by 56.4%. For the whole RS, its R_{PCC} is also as high as 0.521 against the binding affinity. Furthermore, the within-binding-site β contacts among CSP also have a very large MSE (20.8%) increase and a high R_{PCC} = 0.432. Overall for the top 20 important descriptors in random forest as shown in Figure 2, 11 of them are CSP-involved contacts. This clearly suggests that a large number of CSP atoms and their contacts in protein surfaces is a good signal of a likely protein−ligand binding site, while a large number of CSP contacts in protein−ligand docking is a strong indicator of correct protein−ligand binding modes.

Two protein−ligand complexes are shown in Figure 3 to illustrate the hydrophobic effect: 1XAP has the largest CSP.CSP_0 in Figure 3a and 3C2F has zero CSP.CSP_0 in Figure 3b. Interestingly, the binding affinity of 1XAP is as high as 8.7 kcal/mol, while that of 3C2F is only 2 kcal/mol. This clearly reflects the energetic contribution of hydrophobic effect to protein−ligand binding.

The Protection from Backbone Atoms of Proteins in Protein−Ligand Binding. The backbone atoms in proteins with low B factor (the top third descriptor b2 in Figure 2) contribute a great deal to protein−ligand binding affinity with R_{PCC} =0.491. Under the permutation test in random forest, its MSE can increase by 26.3%. From the top 30 important descriptors in Figure 2, 12 descriptors contain β contacts with backbone atoms. The protection and contacts of backbone atoms with ligands are illustrated here by a protein−ligand complex (2E7L) of high binding affinity 7.54 kcal/mol (Figure 4). The reason why backbone atoms can play a great role here should be closely related to the lock-and-key model of protein−ligand binding by which a protein is considered as a lock while a ligand as a key. According to this model, backbone atoms, generally with a lower B factor than side-chain atoms, should behave in a better way to be a part of a lock than a binding site with much motion where both the protein binding sites and ligands are actually very flexible.

Table 3. Meanings of Top Important Descriptors in Figure 2

descriptor[a]	ID[b]	R_{PCC}	meaning
CSP.CSP_0	9	0.521	cross-interface β contacts among carbon/sulfur/phosphorus (CSP) which have no covalent-bond with oxygen/nitrogen
b1		0.496	the sum of B factor of all nonpeptide-bond atoms in ligand which have β contacts with proteins
b2		0.491	the sum of B factor of backbone atoms in proteins which are level-2 nearby atoms of interfacial atoms
metal0		0.350[c]	the preferred average β contacts of first metal group in SI Table 3 in the Supporting Information
C_NO.CSP_1	8	0.185	β contacts between CSP and carbon/sulfur/phosphorus (CNO) which have covalent-bond with oxygen/nitrogen within a protein
CSP.Nbb_1	24	0.348	within-binding-site β contacts between CSP atoms and backbone nitrogen atoms
CSP.CSP_1	9	0.432	within-binding-site β contacts between CSP atoms
CSP.Nbb_0	24	0.342	cross-interface β contacts between CSP atoms and backbone nitrogen atoms
N_H0.N_H1+_0	1	−0.013	cross-interface β contacts between nitrogen/oxygen atoms without and with covalent-bond hydrogen atoms
O_H0.Ca_0	29	0.272	cross-interface β contacts between CA carbon and oxygen atoms without covalent-bond hydrogen atoms
CSP.Obb_1	39	0.472	within-binding-site β contacts between CSP and backbone oxygen atoms
b0		−0.118[d]	the sum of B factor of backbone atoms in peptide ligands
N_H0.CSP_0	6	0.168	cross-interface β contacts between CSP atoms and nitrogen atoms without covalent-bond hydrogens
CSP.Ca_0	31	0.270	cross-interface β contacts between CSP atoms and CA atoms
N_H0.O_H0_1	2	−0.033[e]	within-binding-site β contacts between nitrogen and oxygen atoms both of which have no covalent-bond hydrogens

[a]Descriptors are in a descending rank according to the importance in random forest in Figure 2, and '.' splits two atomic types. [b]The group id of atomic type pairs in SI Table 2 and 3 in the Supporting Information. [c]R_{PCC} is calculated on those protein−ligand complexes with nonzero contacts. This is because many interfaces have zero values, resulting in R_{PCC} = 0.134 on all protein−ligand complexes. [d]R_{PCC} is calculated on those protein−ligand complexes with nonzero contacts. This is because many interfaces have zero values, resulting in R_{PCC} = 0.055. [e]R_{PCC} is calculated on those protein−ligand complexes with nonzero contacts. This is because many interfaces have zero values, resulting in R_{PCC} = 0.085.

dx.doi.org/10.1021/ci400450h | J. Chem. Inf. Model. 2013, 53, 3076−3085

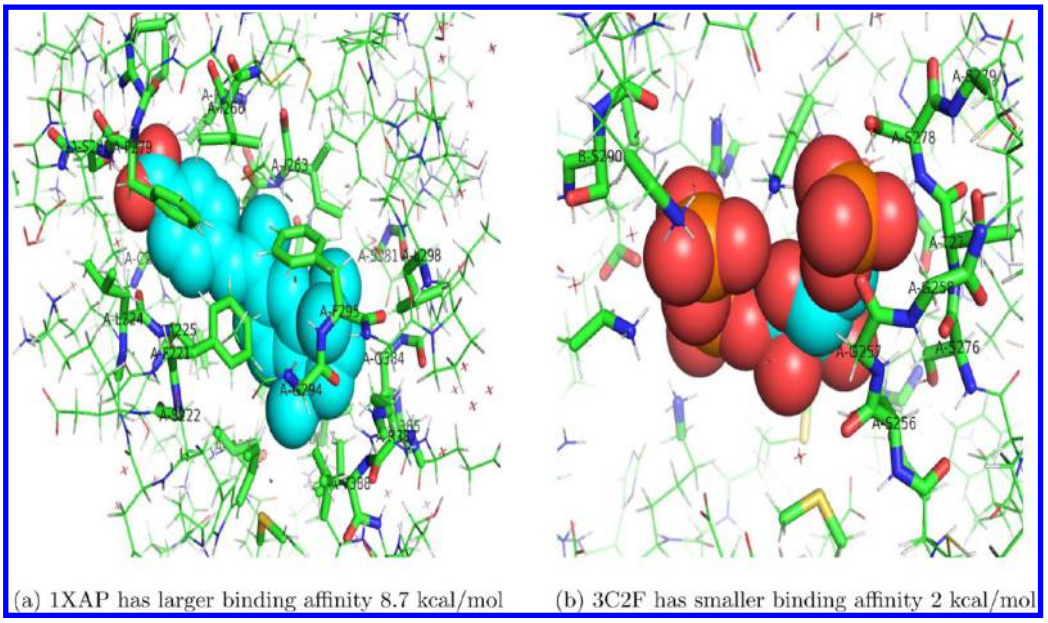


Figure 3. Two protein−ligand complexes showing contacts between CSP atoms. Nitrogen: in blue; Oxygen: in red; Carbon: in cyan for ligands and in green for proteins. Ligands are in sphere and atoms within 6 Å of ligands are in stick. X-YZZZ indicates a residue Y in position ZZZ of chain X. The 2D plot in Ligplot+ is shown in SI Figure 1 in the Supporting Information.

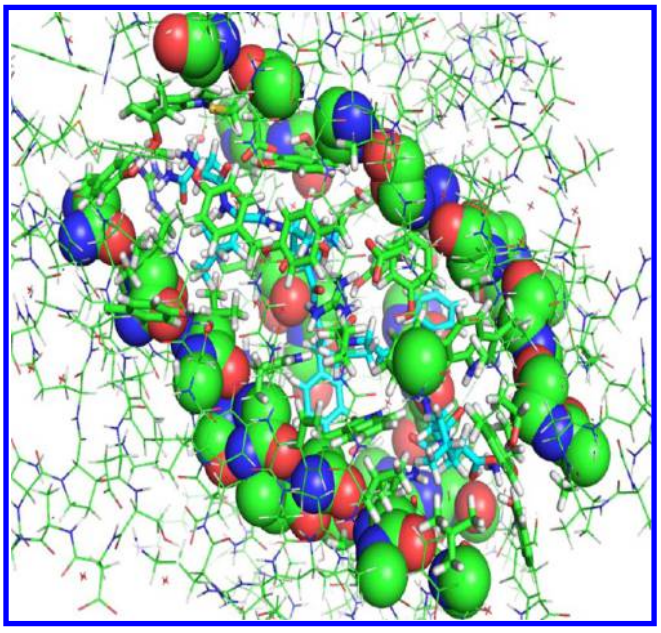


Figure 4. A protein−ligand complexes (2E7L with a binding affinity 7.54 kcal/mol) has backbone-atom rich interfaces. The color of nitrogen, oxygen, and carbon is the same as that in Figure 3. Ligands and atoms within 6 Å of ligands are in stick, while backbone atoms are in sphere. The 2D plot in Ligplot+ is shown in SI Figure 2 in the Supporting Information.

In contrast, the backbone atoms from peptide ligands (b0 in Figure 2) make possibly negative contribution to protein−ligand binding affinity: its permutation increases MSE by 16.8%, and their $R_{PCC} = -0.118$ (R_{PCC} is calculated on the 100 protein-peptide complexes in RS, and a peptide here should have at least three peptide bonds.). Thus, caution should be taken with protein-peptide binding.

Positive Contribution by the Preferred Contacts of Interfacial Metal Atoms to Protein−Ligand Binding. The contribution by the metal-involved β contacts is ranked as the top fourth as shown in Figure 2. This descriptor has $R_{PCC} = 0.35$ in those protein−ligand complexes which have metal ions in their binding interfaces. Its permutation increases MSE by 25.2%. The importance of metal-involved contacts can be seen as examples in Figure 5, where the complex 1CTU in Figure 5a and the complex 1CTT in Figure 5b have the same protein and almost the same ligands except that an oxygen atom in the 1CTU ligand is replaced by a water molecule in the 1CTT ligand. This small difference breaks the strong bridge of the metal contacts between the protein and the ligand, and so the binding affinity decreases from 11.92 to 4.52 kcal/mol. This demonstrates critical importance of metal ions in protein−ligand binding interfaces. Thus, the contact with metal ions at protein surfaces is a useful indicator of protein−ligand binding.

The Constructive Role of π Rings in Protein−Ligand Binding Interfaces. π rings can play an important role in protein−ligand binding and can contribute to protein−ligand binding remarkably. Its importance can be seen from two extreme cases of π ring-involved complexes as shown in Figure 6, where 1Z71 (Figure 6b) contains many contacts between the π rings and the halogen atoms, and 2CCC (Figure 6a) has a binding interface of predominant π rings. Both of these two protein−ligand complexes have a high binding affinity: the binding affinity of 1Z71 is 9.18 kcal/mol, and the binding affinity of 2CCC is 7.76 kcal/mol. Moreover, the contacts between the halogen atoms and the π centers have $R_{PCC} = 0.451$, ranked at the top fifth, when R_{PCC} is calculated on those protein−ligand complexes with nonzero values (e.g., a ligand has halogen atoms here). We would like to note that the existence and importance of the contacts between halogen atoms and π centers has also been highlighted by ref 37.

Potentially Destructive Contacts in Protein−Ligand Binding Interfaces. Potentially destructive contacts are also investigated based on the importance score of each descriptor in random forest and R_{PCC} between the binding affinity and those descriptors which have nonzero values.

An interesting finding is that the permutation of three descriptors for the contacts between two nitrogen atoms

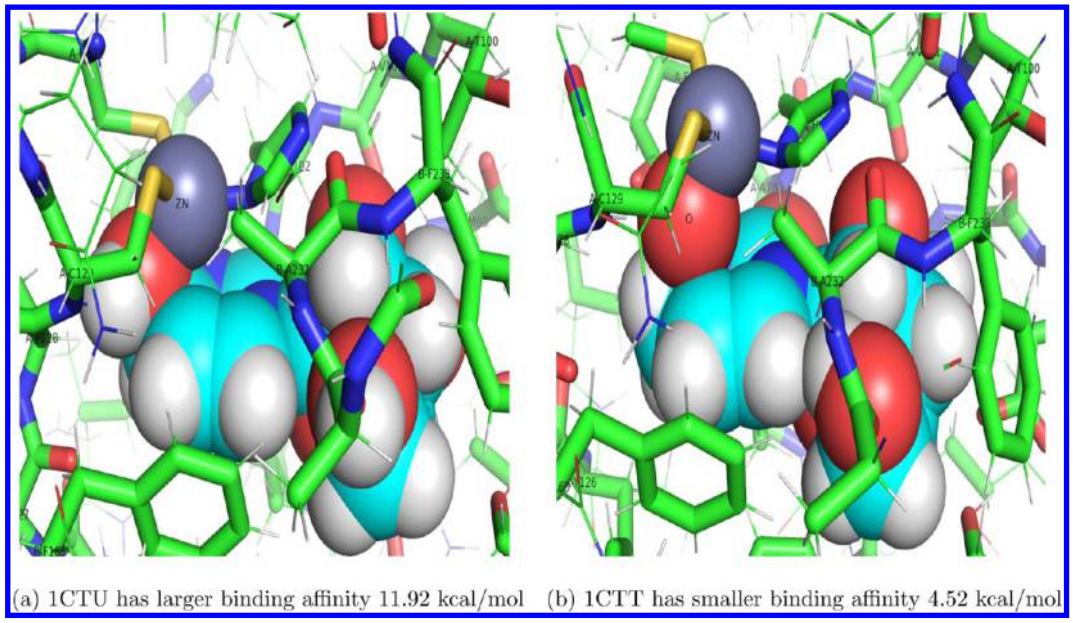


Figure 5. Two protein−ligand complexes showing metal-involved contacts. The notations have the same meaning as those in Figure 3: X-YZZZ, stick, sphere, and the color of nitrogen, oxygen, and carbon. ZN in gray, sulfur atoms in lemon, and hydrogens in white. In (a), the oxygen atom which has very close contact with ZN belongs to the ligand, while in (b), the oxygen is replaced by a water molecule with the label 'O'. The 2D plot in Ligplot+ is shown in SI Figure 3 in the Supporting Information.

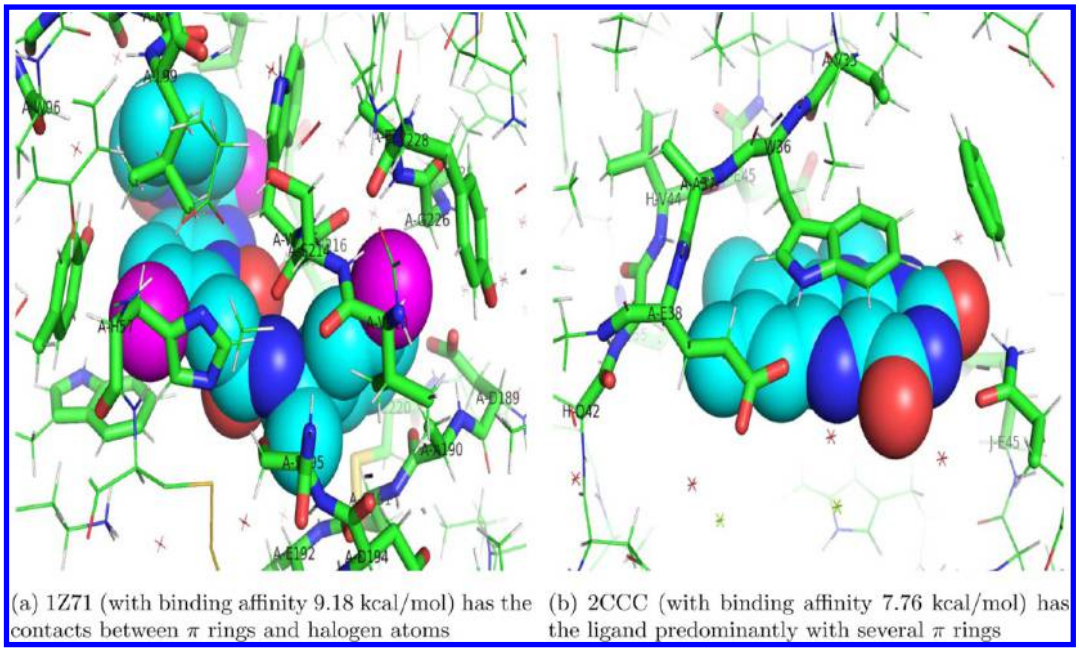


Figure 6. Two protein−ligand complexes containing π-ring-involved contacts. Halogen atoms are in magenta. The following presentations have the same meaning as those in Figure 3: X-YZZZ, stick, sphere, and the color of nitrogen, oxygen, and carbon. The 2D plot in Ligplot+ is shown in SI Figure 4 in the Supporting Information.

(N_H0 for short) without covalent-bond hydrogens can decrease OOB errors. Two of the three descriptors also have R_{PCC} less than −0.2. Meanwhile, several contact descriptors of N_H0-involved within-binding-site contacts also show negative R_{PCC} with regard to binding affinity. This demonstrates that a large number of nitrogen atoms without covalent-bond hydrogens in protein surfaces or in protein−ligand decoys (predicted) should give rise to a signal of a ligand binding site much less favorable which sometimes may not be detectable.

Another interesting finding is that the nonpreferred metal contacts of both metal groups have a negative R_{PCC} about −0.2. The permutation of the nonpreferred metal contacts of the first metal group has a positive importance score in random forest, confirming a negative correlation with regard to binding affinity. Thus, those metal ion contacts with nonpreferred atoms should also weaken instead of strengthen protein−ligand binding.

■ CONCLUSION

In this work, a new scoring function B2Bscore has been proposed to improve the performance of protein−ligand binding affinity prediction. B2Bscore integrates B factor into

dx.doi.org/10.1021/ci400450h | J. Chem. Inf. Model. 2013, 53, 3076−3085

β contacts in a vector representation of protein−ligand binding interfaces. Evaluated on an independent data set or under LCOCV, B2Bscore demonstrates a superior performance to existing scoring functions especially the big improvement under LCOCV. The random forest learning process in B2Bscore also extracts several important contact descriptors in protein−ligand binding interfaces. These descriptors are intuitive and thus can be used in other protein−ligand applications such as protein−ligand docking. Meanwhile, we realized that B factor is mainly available from crystallography structures (those structures are actually predominant in PDB), and it is not in the structures obtained from NMR. Thus in the future, we will overcome this issue from the following two aspects. One is to use or design more efficient methods to predict B factor, and the other is to take the advantage of the dynamic models in NMR to extract useful concepts to reflect the dynamic property of proteins, as B factor does. These two approaches are also very useful for those structures obtained from homology modeling or other sources.

MATERIALS AND METHODS

We first describe the data sets of protein−ligand complexes which are used by this work. Then, we present a method for the normalization of B factor. We also describe what is a β contact and how to produce β contacts from protein−ligand binding structures. After that, we introduce how to use a vector representation derived from β contacts to describe protein−ligand binding interfaces with the integration of B factor. Finally, we provide how to use random forest to learn the relationships between the interface vectors and the binding affinity of protein−ligand complexes, followed by evaluation measures.

Data Sets. Two kinds of data sets, widely used to evaluate a scoring function of protein−ligand binding affinity prediction by existing methods,[2−4,26,31] are closely related to the refined set (RS) of protein−ligand complexes in PDBbind.[34] On the first evaluation data set, the core set (CS) of RS is constructed according to the clusters in RS using a cutoff of 90%.[34] CS was used as the independent test data set, while $RS - CS$ was used in the training process. Here, $RS - CS$ means protein−ligand complexes in RS but not in CS. In the PDBbind version 2009, there are 1741 protein−ligand complexes and 73 clusters in RS and 219 protein−ligand complexes in CS.

In RS, many proteins have more than 90% sequence similarity to others including those in CS. Thus, the work in ref 26 further grouped RS into 26 clusters (refer to Table 2) where proteins within a group may have high sequence similarity but proteins between the groups must have low similarity. These clusters are specially noted as a clustered set, which is the second evaluation data set in our work.

B Factor. B factor quantitates the uncertainty/mobility for each atom of dynamic protein 3D structures, that is, the displacement of the atomic positions from its mean position. B factor is also known as temperature factor or Debye−Waller factor—used in condensed matter physics to describe the attenuation of X-ray scattering or coherent neutron scattering caused by thermal motion. In crystallography, B factor describes the degree to which the electron density is spread out. In protein structures, B factor can be taken to indicate the relative vibrational motion or the disorder in protein crystal. It can be calculated using $B^i = 8\pi^2 U_i^2$, where U_i^2 is the mean square displacement of atom i. B factor increases as U_i^2 increases. A low B factor implies that the atom is in well-ordered parts of the structure, while a large B factor generally suggests a very flexible atom. Protein flexibility is heavily related to protein functions such as catalysis and allostery.[38] Deeply buried atoms in the core of the protein are usually hardly moving at all with low B factor,[39] while interfacial residues in protein binding complexes have lower B-factors in comparison to the rest of the tertiary structural surface.[40] In different PDB structures, the distribution of B factors varies greatly. Thus, a normalized B factor as calculated by eq 1 is used in this work

$$B^i_{norm} = \frac{\overline{B} - B^i}{\delta_B} \times \frac{1}{1.645}$$

$$\ddot{B}^i_{norm} = min[max(B^i_{norm} + 1, 0), 2] \quad (1)$$

where B^i is the B factor of atom i, $\overline{B}$ and δ_B are the mean and the standard deviation of the B factor of all atoms in the PDB biological complexes, and B^i_{norm} is the normalized B factor of atom i. 1.645 is used to indicate that the probability of a value outside $[-1.645,1.645]$ is 0.05 for each of the two tails under a standard normal distribution. *min* or *max* means the minimum or maximum of two values. By eq 1, an atomic B factor, which has more than $1.645 \times \delta$ difference from $\overline{B}$, is considered to have $1.645 \times \delta$ difference. $\ddot{B}^i_{norm}$ is used in this work with the range $[0,2]$.

β Contacts. β contact is a newly proposed definition of atomic contacts for precisely modeling well-organized protein 3D structures.[32] Its detail can be found in ref 32. For easy reference, we provide a brief summarization below about what is β contact and how to produce β contacts from a protein−ligand binding structure.

What Is β Contact. In a given 3D structure of a protein−ligand complex p, a β contact c of two atoms i and j requires (i) the spatial distance between i and j is less than a threshold T_d plus the sum of their van der Waals radii as defined in ref 41, (ii) i and j share a Voronoi facet in p's Voronoi diagram, and (iii) the contact cannot break p's β-skeleton. Here, the β-skeleton[42] of a discrete set p is an undirected graph in computational geometry. In this graph, two points i and j have an edge if any angle ikj (denoted as $\angle ikj$) is sharper than the angle threshold $\angle\beta$ defined by β, $\forall k \in p, k \neq i,j$. The relation of β and $\angle\beta$ is as follows: $\angle\beta = arcsin(1/\beta)$ if $\beta \geq 1$ and $\angle\beta = \pi - arcsin\beta$ if $\beta \leq 1$. β or $\angle\beta$ actually defines a forbidden region fr of the contact between i and j. For example, when $\angle\beta = 90$, that is $\beta = 1$, fr is a sphere with the midpoint of i and j as the center and c's length as the diameter, which is similar to the van der Waals radii of atoms. An atom k is farther from the center, $\angle ikj$ is sharper. Thus, β contacts assume that two atoms should have enough direct contact area to form an important interaction. We note that in the β contact definition, different $\angle\beta$ thresholds define various forbidden regions fr, and fr under a smaller $\angle\beta$ such as $\angle\beta = 75°$ is larger than that under a larger $\angle\beta$ such as $\angle\beta = 90°$; thus, $\angle\beta = 75°$ is stricter than $\angle\beta = 90°$.

How To Construct β Contacts Graphs for 3D Protein−Ligand Binding Structures. A given protein−ligand binding structure p can be represented by a β atomic contact graph $b(p)$, if all of the heavy atoms are represented by points, and the β contacts are represented by edges. To produce $b(p)$ for p, Qhull is first used to obtain the Delaunay triangulation[43] for all points. After that, a distance threshold T_d is used to remove those atomic contacts whose distances are too large. T_d is set to 2.8 Å (the diameter of a water molecule). This threshold is an insensitive factor to β contacts when it is large enough. Third, each atomic contact is checked to guarantee that it satisfies β

skeletons. In this work, the threshold $\angle\beta$ is set according to eq 2

$$\angle\beta = min\{max[75 + int(10*(d_{ij} - 3.5)), 75], 90\} \quad (2)$$

where d_{**} is the spatial distance between two atoms, *int* is a function to convert a float into an integer, while *min* or *max* is a function to return a smaller or larger number of two numbers, and thus restricts $\angle\beta$ to being in [75,90]. In this work, there are three exceptions for the relaxed setting of $\angle\beta$: (i) if d_{ik} or d_{jk} is less than 1.8 Å, $\angle\beta = \angle\beta + 10$, where k is so close to i or j that k should be covalent-bond atoms of i or j, and the covalent bonds with shorter length and the noncovalent bond between i and j make $\angle ikj$ easily larger than 75°; (ii) if d_{ik} or d_{jk} is less than 2.6 Å (for example, disulfide bonds), $\angle\beta = \angle\beta + 5$, which has a similar situation to the first one; and (iii) if i or j represents the center of a planar ring π such as aromatic rings (denoted as a pseudoatom π-center) and k is an atom on π, $\angle\beta = 95$, where i and k, or j and k, is in the same plane.

Vector Representation of Protein−Ligand Binding Interfaces. Given the β atomic contact graph $b(p)$ of a protein−ligand binding complex, a vector representation can be used to describe the protein−ligand binding interface. This vector representation includes several types of subvectors for atomic types (two types) and their atomic pairs (four types). The vector representation is also integrated by normalized B factor.

Some Preliminary Definitions. In protein−ligand binding interfaces, we use β contacts both within protein binding sites and across interfaces, called within-binding-site contacts and across-interface contacts for short. With regard to within-binding-site contacts, we define nearby atoms of an atom i. Suppose $i - j - k - l - m$, where − indicates a covalent bond. From i, the covalent-bond step is 0 to i, is 1 to j, is 2 to k, is 3 to l, and is 4 to m, respectively. We call i,j and k level-2 nearby atoms of i which have no more than two covalent-bond steps from i. Similarly, we call i,j,k and l level-3 nearby atoms of i. In this work, atomic contacts between level-3 nearby atoms are not used in the vector representation.

Atomic Pair Information in Protein−Ligand Interfaces. In $b(p)$, all heavy atoms are grouped into fifteen types as shown in SI Table 1 in the Supporting Information. In particular, the center of a planar ring (π-center for short), such as aromatic rings, is considered as a pseudoatom. The value of each atomic type pair (Ti,Tj) is calculated using eq 3

$$v(Ti, Tj) = \sum_{m\in P\wedge t_m=Ti} \sum_{n\in C_m\wedge t_n=Tj} \frac{(\ddot{B}^m_{norm} + \ddot{B}^n_{norm})}{d^2_{(m,n)}} + \sum_{l\in P\wedge t_l=Tj} \sum_{k\in C_l\wedge t_k=Ti} \frac{(\ddot{B}^l_{norm} + \ddot{B}^k_{norm})}{d^2_{(l,k)}} \quad (3)$$

where P is a set of protein atoms which have at least one β contact with the ligand, $C_{m/l}$ is a set of ligand atoms which have cross-interface β contacts with atom m/l, and d_{**} is the spatial distance between two atoms. T_i or t_i indicates the atomic type of atom i.

Those atomic type pairs are further manually clustered into 41 types (refer to SI Table 2 and 3 in the Supporting Information) to avoid a sparse vector when representing protein−ligand binding interfaces. Thus, an atomic pair vector has 41 descriptors for cross-interface β contacts, and the value of each descriptor is the sum of its belonging atomic type pairs.

Another atomic pair vector is also constructed for those within-binding-site β contacts of level-2 nearby atoms of interfacial atoms. This vector has the same 41 descriptors as the cross-interface contacts and can be calculated using eq 3 with the difference that P is a set of level-2 nearby atoms of interfacial atoms, and C is a set of atoms which have within-binding-site β contacts with atoms in P.

We note that a water molecule in PDB is considered as a part of a binding interface if (i) it has at least 3 potential hydrogen-bonds contacts, or (ii) it has 2 potential hydrogen-bond contacts and also has at least 2 other contacts with spatial distances less than 4 Å. A potential hydrogen-bond contact should have a spatial distance less than 3.2 Å between a hydrogen donor (such as a nitrogen atom) and a hydrogen acceptor (such as an oxygen atom). Under this requirement, these water molecules, especially called bound water molecules, are heavily involved in protein−ligand binding such that they play an integral part; but the contacts between any two water molecules are not considered.

π-Center Involved β Contacts. A vector with three descriptors is especially used to summarize π-center involved β contacts across interfaces: one descriptor represents β contacts between two π-centers, the second denotes β contacts between a π-center and the carbon, sulfur, phosphate, halogen atoms which have no covalent bonds to nitrogen or oxygen atoms, and the third contains β contacts between a π-center and the other atoms. The values of the descriptors are calculated in a similar way to eq 3. A similar vector is used for π-center involved β contacts of level-2 nearby atoms of interfacial atoms. Here, the two descriptors of the contacts between the two π-centers are not explicitly used since they already considered in the vectors in the previous subsection.

Backbone-Atom Involved β Contacts. β contacts involving backbone atoms are also summarized by two vectors each with three descriptors: one descriptor represents β contacts between backbone atoms, the second means β contacts between backbone atoms and other atoms, while the third one is about β contacts between other atoms. One vector is for β contacts with T_d = 0.7 Å, while the other is for β contacts with T_d = 2.8 Å. The values of the descriptors in the two vectors are calculated in a similar way to eq 3.

Metal-Ion Involved β Contacts. Metal atoms are abundant and play a vital role in protein−ligand binding interfaces. Their β contacts are also used in this work. Metal atoms are first categorized into two groups: one group contains ZN, CO, CU, FE, and NI, and the other includes CA, MG, MN, NA, SR, and K. Then for each group, the contact atoms can be preferred by metal ions or not (please refer to SI Table 4 in the Supporting Information for details). After that, a vector of four descriptors is used to summarize the β contacts of interfacial metal ions: the first two descriptors are for preferred contacts and other contact behaviors of the first metal group, and the latter two descriptors for preferred contacts and other contact behaviors of the second metal group. Similarly, another vector of four descriptors is used to represent only across-interface contact behaviors of metal ions. In these two vectors, the values are calculated in a similar way to eq 3, but the values are averaged by the number of metal ions in protein−ligand binding interfaces. Here, a metal ion is considered to contact with other atoms if their spatial distance is not greater than 2.7 Å.

Atomic Information of Contact Atoms in Ligands. Besides those vectors of atomic type pairs above, a vector of atomic type information is used to summarize all interacting ligand

dx.doi.org/10.1021/ci400450h | J. Chem. Inf. Model. 2013, 53, 3076−3085

atoms in β contacts. According to each ligand atomic type in SI Table 5 in the Supporting Information, eq 4 is used to calculate its value

$$v(Ti) = \sum_{m \in L \wedge t_m = Ti} \sum_{n \in C} \frac{(\ddot{B}^i_{norm} + \ddot{B}^n_{norm})}{d^2_{(m,n)} \times N_{Ti}} \quad (4)$$

where L is a set of ligand atoms, C is a set of atoms in the protein which have β contacts with the ligand atom m, while N_{Ti} is the number of ligand atoms with atomic type T_i. Here, three vectors each with all ligand atomic types are used to summarize the contact information with the spatial distance $\leq$3.5 Å, $\leq vdw_1 + vdw_2 + 0.7$, or $\leq vdw_1 + vdw_2 + 2.8$ respectively where vdw_1 and vdw_2 are the van der Waals radii of two contact atoms.

Atomic Information of Interfacial Backbone Atoms. The summarization of backbone atoms in across-interface β contacts is also useful. A vector of two descriptors is used to summarize β interfacial backbone atoms in proteins: one descriptor is the sum of $\ddot{B}^j_{norm}$ of interfacial backbone atoms and the other is of other interfacial atoms. Similarly, the other vector of two descriptors is calculated for backbone atoms and the other atoms in ligands (especially in peptide ligands).

Our Proposed Method and Evaluation Measures. *B2Bscore: A Random Forest Learning Process of Binding Affinity.* Merging the above vectors all together, every protein−ligand binding interface can be then denoted by a vector of 131 descriptors plus the binding affinity measurement. Then, random forest[44] is used to learn the relation between all of the vectors and the binding affinity for a data set of protein−ligand complexes, similar to RFscore.[4] The details of how to use random forest for the binding affinity prediction are provided in the Supporting Information.

The randomForest package also ranks the importance of descriptors using permutation test. Generally, the descriptors with larger importance scores are ranked as more influential in the learning process than descriptors with smaller scores. This score in random forest is thus used to evaluate the importance to protein−ligand binding for all descriptors derived from β contacts.

Evaluation Measures. Finally, the correlation of predicted and experimentally measured binding affinity is quantified using Pearson's correlation coefficient (R_{PCC}) and root-mean-square error (RMSE). Their definitions are provided in the Supporting Information.

■ ASSOCIATED CONTENT

Supporting Information

Description of PDBbind and of random forest, the definition of evaluation measures, the atomic types and the groups of their pairs. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jinyan.li@uts.edu.au.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.* **2004**, *47*, 3032−3047.

(2) Kramer, C.; Gedeck, P. Global free energy scoring functions based on distance-dependent atom-type pair descriptors. *J. Chem. Inf. Model.* **2011**, *51*, 707−720.

(3) Cheng, T.; Liu, Z.; Wang, R. A knowledge-guided strategy for improving the accuracy of scoring functions in binding affinity prediction. *BMC Bioinf.* **2010**, *11*, 193.

(4) Ballester, P. J.; Mitchell, J. B. O. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26*, 1169−1175.

(5) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed automated docking of flexible ligands to proteins: Parallel applications of AutoDock 2.4. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 293−304.

(6) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.

(7) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.

(8) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(9) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.

(10) Momany, F. A.; Rone, R. Validation of the general purpose QUANTA 3.2/CHARMm force field. *J. Comput. Chem.* **1992**, *13*, 888−900.

(11) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.

(12) Muegge, I. In *Virtual Screening: An Alternative or Complement to High Throughput Screening?*; Klebe, G., Ed.; Springer: Netherlands, 2002; Vol. *20*, pp 99−114.

(13) Muegge, I. Effect of ligand volume correction on PMF scoring. *J. Comput. Chem.* **2001**, *22*, 418−425.

(14) Muegge, I. PMF scoring revisited. *J. Med. Chem.* **2006**, *49*, 5895−5902 PMID: 17004705.

(15) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.

(16) Velec, H. F. G.; Gohlke, H.; Klebe, G. DrugScore^{CSD} knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.* **2005**, *48*, 6296−6303.

(17) Mooij, W. T. M.; Verdonk, M. L. General and targeted statistical potentials for protein-ligand interactions. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 272−287.

(18) Huang, S.-Y.; Zou, X. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. *J. Comput. Chem.* **2006**, *27*, 1866−1875.

(19) Huang, S.-Y.; Zou, X. An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. *J. Comput. Chem.* **2006**, *27*, 1876−1882.

(20) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. R.; Fogel, L. J.; Freer, S. T. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: Conformationally flexible docking by evolutionary programming. *Chem. Biol. (Oxford, U. K.)* **1995**, *2*, 317−324.

(21) Jain, A. N. Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 427−440.

(22) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.

(23) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible docking using Tabu search and an empirical

estimate of binding affinity. *Proteins: Struct., Funct., Bioinf.* **1998**, *33*, 367−382.
(24) Bohm, H.-J. Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309−323, DOI: 10.1023/A:1007999920146.
(25) Bohm, H.-J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256, DOI: 10.1007/BF00126743.
(26) Kramer, C.; Gedeck, P. Leave-cluster-out cross-validation is appropriate for scoring functions derived from diverse protein data sets. *J. Chem. Inf. Model.* **2010**, *50*, 1961−1969.
(27) Wang, R.; Liu, L.; Lai, L.; Tang, Y. SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex. *J. Mol. Model.* **1998**, *4*, 379−394.
(28) Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11−26.
(29) Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C.; Waldman, M. LigScore: a novel scoring function for predicting binding affinities. *J. Mol. Graphics Modell.* **2005**, *23*, 395−407.
(30) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **2006**, *49*, 6177−6196.
(31) Ouyang, X.; Handoko, S.; Kwoh, C. Cscore: a simple yet effective scoring function for protein-ligand binding affinity prediction using modified cmac learning architecture. *J. Bioinf. Comput. Biol.* **2011**, *9* Suppl 1.
(32) Liu, Q.; Kwoh, C.-K.; Hoi, S. C. H. Beta atomic contacts: Identifying critical specific contacts in protein binding interfaces. *PLoS One* **2013**, *8*, e59737.
(33) Overington, J.; Al-Lazikani, B.; Hopkins, A. How many drug targets are there? *Nat. Rev. Drug Discovery* **2006**, *5*, 993−996.
(34) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.* **2009**, *49*, 1079−1093.
(35) Trott, O.; Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.
(36) Bissantz, C.; Kuhn, B.; Stahl, M. A medicinal chemist's guide to molecular interactions. *J. Med. Chem.* **2010**, *53*, 5061−5084.
(37) Robertazzi, A.; Krull, F.; Knapp, E.-W.; Gamez, P. Recent advances in anion-π interactions. *CrystEngComm* **2011**, *13*, 3293−3300.
(38) Yuan, Z.; Bailey, T. L.; Teasdale, R. D. Prediction of protein B-factor profiles. *Proteins: Struct., Funct., Bioinf.* **2005**, *58*, 905−912.
(39) Parthasarathy, S.; Murthy, M. R. Analysis of temperature factor distribution in high-resolution protein structures. *Protein Sci.* **1997**, *6*, 2561−2567.
(40) Yuan, Z.; Zhao, J.; Wang, Z.-X. Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Eng.* **2003**, *16*, 109−114.
(41) Hubbard, S. J.; Thornton, J. M. *'NACCESS', computer program*; 1993.
(42) Kirkpatrick, D. G.; Radke, J. D. A framework for computational morphology. *Computational Geometry, Machine Intelligence and Pattern Recognition* **1985**, *2*, 217−248.
(43) Barber, C. B.; Dobkin, D. P.; Huhdanpaa, H. The Quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.* **1996**, *22*, 469−483.
(44) Breiman, L. Random forests. *Machine Learn.* **2001**, *45*, 5−32.