JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Structural and Sequence Similarity Makes a Significant Impact on Machine-Learning-Based Scoring Functions for Protein−Ligand Interactions

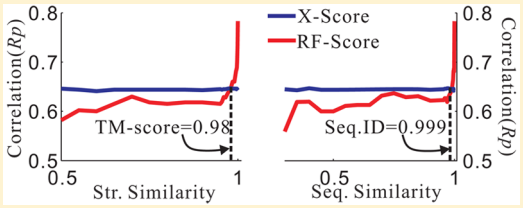
Yang Li[†,‡] and Jianyi Yang*[,‡]

[†]College of Life Sciences, Nankai University, Tianjin 300071, China

[‡]School of Mathematical Sciences, Nankai University, Tianjin 300071, China

Supporting Information

ABSTRACT: The prediction of protein−ligand binding affinity has recently been improved remarkably by machine-learning-based scoring functions. For example, using a set of simple descriptors representing the atomic distance counts, the RF-Score improves the Pearson correlation coefficient to about 0.8 on the core set of the PDBbind 2007 database, which is significantly higher than the performance of any conventional scoring function on the same benchmark. A few studies have been made to discuss the performance of machine-learning-based methods, but the reason for this improvement remains unclear. In this study, by systemically controlling the structural and sequence similarity between the training and test proteins of the PDBbind benchmark, we demonstrate that protein structural and sequence similarity makes a significant impact on machine-learning-based methods. After removal of training proteins that are highly similar to the test proteins identified by structure alignment and sequence alignment, machine-learning-based methods trained on the new training sets do not outperform the conventional scoring functions any more. On the contrary, the performance of conventional functions like X-Score is relatively stable no matter what training data are used to fit the weights of its energy terms.



INTRODUCTION

Scoring is a key component in molecular docking for estimating the binding affinity between a target protein and a small-molecule ligand.[1−3] Utilizing the structures of protein−ligand complexes, the goal of a scoring function is to computationally measure the binding free energy by producing scores that are supposed to be linearly correlated with the experimentally determined binding affinities. This ability is known as the scoring power of a scoring function.[4] Scoring functions find their applications in many fields, such as binding pocket prediction,[5] drug lead optimization,[6] target druggability assessment,[7] and virtual screening.[8]

Conventional scoring functions can be separated into three groups: physics-based,[9] knowledge-based,[1,10,11] and empirical.[12] The empirical scoring functions first derive multiple energy terms (e.g., van der Waals interactions) and then fit them to the experimental binding affinity data through a predetermined additive functional form.[13] Hence, the prediction results can be directly interpreted by decomposing the energy terms and associated with the biochemical characteristics of the binding patterns. A comprehensive discussion and classification of current scoring functions is provided in the review by Liu and Wang.[14]

Despite intensive research in the past two decades, the progress made by conventional scoring functions is slow, with a maximal Pearson correlation coefficient of around 0.6.[4,15] However, remarkable improvement in this area has been reported recently (the correlation was increased to about 0.8[16−20]) with a set of custom-designed descriptors and machine learning algorithms. We classify them as machine-learning-based methods, as they were extensively trained with advanced machine learning algorithms such as random forests[16] and support vector machines.[19]

With the strikingly superior scoring power that significantly outperforms conventional scoring functions, the performance of machine-learning-based methods has been investigated and discussed,[21−23] with the best-performing function, RF-Score, selected as a representative method. Kramer and Gedeck[21] demonstrated that the prediction quality of RF-Score was significantly degraded when the model was trained and validated using leave-cluster-out cross-validation within the family-specific context. However, Ballester, the author of RF-Score, commented that the above conclusion might be inappropriate because no comparison with other scoring functions was performed.[22] Besides the scoring power, Gabel et al.[23] assessed the virtual screening power and docking power of machine-learning-based methods and found that these methods were much worse than conventional scoring functions. They concluded that the higher accuracy on the binding affinity prediction achieved by machine-learning-based methods was suspicious.[23] From these studies, we can see that the reason for the improvement of the scoring power made by machine-learning-based methods remains unclear.

Received: January 25, 2017
Published: March 30, 2017

DOI: 10.1021/acs.jcim.7b00049
J. Chem. Inf. Model. 2017, 57, 1007−1012

In this work, we performed a closer investigation of the scoring power of machine-learning-based methods, using RF-Score as a representative because of its outstanding performance and availability of open source codes. The conventional empirical scoring function X-Score was used as a control. By filtering the training data according to different levels of structural and sequence similarity to the test data, we demonstrate that the scoring power of the machine-learning-based methods is in fact affected by highly similar samples in the training set. On the contrary, the conventional empirical scoring functions do not have such an issue.

METHODS

Data Sets. The PDBbind database,[24] a standard benchmark widely used in this field,[25,26] was adopted to train and test the selected scoring functions. Though updated versions of PDBbind are available, we used the 2007 version to allow direct comparison with scoring functions that have been previously tested on the same data set. In the "refined" set of PDBbind, there are 1300 protein−ligand complexes with high-resolution X-ray crystal structures and experimentally determined binding affinity data. This set was clustered at 90% sequence similarity, resulting in a total of 65 clusters. For each cluster, the three complexes with the highest, median, and lowest binding affinity were selected to form a new set containing 195 diverse complexes named the "core" set, which was used as the test set by most studies. To avoid overlap with the test set, the complexes appearing in the test set were removed from the refined set to construct the training set.[4,16] As a result, we obtained 1105 (=1300 − 195) complexes in the training set. However, such construction of the training set may not be stringent enough, especially for training machine-learning-based methods, which will be discussed subsequently. The refined and core sets were downloaded from the official PDBbind Web site.[27] The original rather than optimized structures from PDBbind were used here.

In this work, the test set (TS) was kept unchanged to make direct comparisons with the results from different settings. The training set was further filtered according to its similarity to the test set. To systematically determine the similarity between training and test sets, we performed all-against-all comparisons between training and test samples. A series of new training sets (NTs) were constructed by gradually removing samples from the original training set (OT) according to specified similarity cutoffs:

$$\mathrm{NT}(c) = \{p_i | p_i \in \mathrm{OT} \text{ and } \forall\, q_j \in \mathrm{TS},\ s(p_i, q_j) \le c\} \quad (1)$$

where p_i and q_j represent the ith and jth samples from the OT and TS, respectively; $s(p_i, q_j)$ is the similarity between p_i and q_j, which will be defined later; and c is the similarity cutoff. We can see that each NT is a subset of OT but that the similarity between the samples in an NT and the TS is less than or equal to the specified cutoff. The pairwise similarities between samples in the training and test sets are available at http://yanglab.nankai.edu.cn/download/SF.

Structural Similarity. The structural similarity between two protein structures is defined as the TM-Score,[28] which is calculated by the structure alignment program TM-align.[29] TM-score is a length-independent scoring function for measuring protein structure similarity:[28]

$$\text{TM-score} = \max\left[\frac{1}{L}\sum_{i=1}^{L_{\mathrm{ali}}}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right] \quad (2)$$

where d_i is the distance between the ith pair of Cα atoms of the two structures, L_{ali} is the number of aligned residue pairs identified by TM-align,[30] L is the length of the test protein, and d_0 is given by $d_0 = 1.24\sqrt[3]{L-15} - 1.8$. TM-score is in the range of 0 to 1, and a higher value of TM-score indicates more similar structures. Because most of the proteins in the PDBbind benchmark contain multiple chains, every chain structure for each protein is extracted and compared. The TM-score for two proteins is defined as the lowest pairwise-chains TM-score. We also tested the definition using the highest value, but the conclusion did not change (see Table S1 in the Supporting Information).

Sequence Identity. Needleman−Wunsch dynamic programming,[31] as implemented in the software NW-align,[32] was used to compute the sequence identity between two proteins. The sequence identity is defined as the number of aligned identical residues divided by the length of the test protein. Similar to the calculation of structural similarity, the comparison is made in an all-chains-against-all-chains way. The sequence identity for the two proteins is then defined as the lowest pairwise-chains sequence identity. We tested the definition using the highest value as well, but the conclusion did not change (see Table S2).

RF-Score. A total of 36 descriptors representing the atomic distance counts from the original RF-Score[16] was used for retraining and retesting. It should be noted that six more descriptors from Autodock Vina[33] are also included in the updated version of RF-Score.[34] We performed similar experiments with the updated version, but the conclusion did not change (data not shown), and thus, we decided to use the 36 descriptors from the original version of RF-Score. The random forest algorithm was used in RF-Score to train models and make predictions. As an ensemble method, random forest fits a number of decision trees and improves the predictive accuracy using the average of each tree output. The configuration for training was the same as for the original RF-Score.[16] The number of trees in the forest was set to 500, the number of features to consider when looking for the best split was set to 5, and the out-of-bag strategy was enabled. The random forest model was implemented using the scikit-learn package.[35]

X-Score. Conventional empirical scoring functions are composed of various energy terms and predict the binding affinity through a linear combination of these terms. As a representative of empirical scoring functions, X-Score contains four energy terms: (1) van der Waals interactions, (2) hydrogen bonding, (3) deformation effects, and (4) hydrophobic effects. In X-Score, the last term can be calculated with three different algorithms: hydrophobic surface (HS), hydrophobic contact (HP), and hydrophobic match (HM). Correspondingly, three scores are obtained by combining the four terms: X-Score::HS, X-Score::HP, and X-Score::HM. It was reported that the best-performing one is X-Score::HM.[4] Thus, X-Score::HM was used to represent X-Score in this study. The calculation of the individual energy terms was done with the X-Score package (version 1.3), downloaded from its official Web site.[36] Default options were used in running the X-Score program. To make a fair comparison with RF-Score, we

DOI: 10.1021/acs.jcim.7b00049
J. Chem. Inf. Model. 2017, 57, 1007−1012

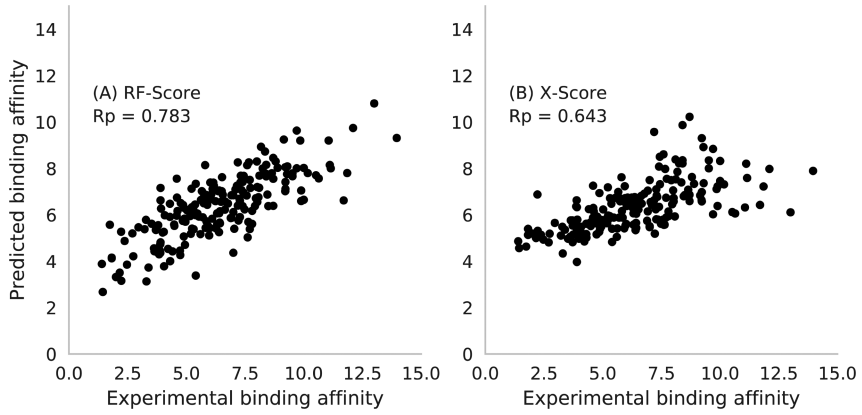


Figure 1. Correlation between the predicted and experimental binding affinities on the 195 test complexes using (A) RF-Score and (B) X-Score.

Table 1. Performance of RF-Score and X-Score Retrained at Different Similarity Levels; The Training Sets Were Constructed Using Equation 1, with TM-score as the Similarity Measure

			RF-Score			X-Score		
cutoff	no. in training set	TC	R_p	R_s	RMSE	R_p	R_s	RMSE
1.000	1105	0.620	0.783	0.769	1.542	0.643	0.707	1.867
0.999	964	0.608	0.731	0.729	1.669	0.644	0.705	1.862
0.998	867	0.606	0.705	0.711	1.719	0.645	0.706	1.859
0.997	810	0.608	0.689	0.698	1.763	0.646	0.706	1.863
0.996	793	0.607	0.684	0.694	1.774	0.645	0.706	1.864
0.995	779	0.605	0.682	0.693	1.776	0.646	0.706	1.864
0.990	710	0.601	0.660	0.660	1.818	0.646	0.707	1.858
0.985	685	0.598	0.660	0.664	1.819	0.646	0.708	1.857
0.980	645	0.597	0.647	0.648	1.836	0.647	0.710	1.850
0.975	605	0.598	0.641	0.646	1.853	0.646	0.709	1.854
0.970	582	0.599	0.635	0.640	1.863	0.646	0.709	1.855
0.965	567	0.600	0.628	0.635	1.872	0.645	0.708	1.859
0.960	563	0.600	0.629	0.633	1.873	0.645	0.708	1.860
0.955	548	0.599	0.619	0.619	1.887	0.644	0.707	1.860
0.950	540	0.597	0.615	0.613	1.893	0.645	0.709	1.856
0.900	510	0.588	0.618	0.615	1.895	0.643	0.708	1.861
0.850	476	0.587	0.618	0.618	1.900	0.644	0.707	1.848
0.800	462	0.587	0.615	0.618	1.906	0.644	0.708	1.845
0.750	443	0.590	0.618	0.629	1.903	0.644	0.708	1.843
0.700	416	0.594	0.630	0.637	1.880	0.644	0.709	1.841
0.650	400	0.596	0.615	0.628	1.899	0.644	0.709	1.841
0.600	371	0.591	0.600	0.617	1.923	0.641	0.707	1.843
0.550	330	0.598	0.602	0.616	1.928	0.644	0.709	1.848
0.500	294	0.594	0.582	0.597	1.960	0.646	0.710	1.842
0.450	222	0.600	0.571	0.599	1.975	0.640	0.704	1.870
0.400	116	0.572	0.556	0.584	1.991	0.624	0.666	1.870

ignored the fixed weights of the energy terms in the X-Score package and refitted them with exactly the same new training sets used in the retraining of RF-Score. The linear regression model was implemented using the scikit-learn package.[35]

Metrics for Performance Evaluation. The scoring power of a scoring function is usually evaluated by the Pearson correlation coefficient (R_p), the Spearman correlation coefficient (R_s), and the root-mean-square error (RMSE) between the predicted and experimental binding affinities. The first two metrics have been extensively used in the literature[4,11,15,18,21] to assess various scoring functions. Thus, the analyses made here are mainly based on the correlation coefficients R_p and R_s.

RESULTS AND DISCUSSION

The Scoring Power of RF-Score Exceeds That of X-Score When Sample Similarity Is Not Considered. When we set the similarity cutoff to 1.0, the training set is exactly the same as that used by previous studies.[4,16] The correlations between the experimental binding affinity and those predicted by RF-Score and X-Score are shown in Figure 1. They are quite similar to those reported in their publications,[4,16] confirming the correctness of our retraining of the two scoring functions. The binding affinity predicted by RF-Score is apparently better than that predicted by X-Score. Ballester et al.[17] provided possible factors that might contribute to this improved result. However, it remains disputable why the machine-learning-

DOI: 10.1021/acs.jcim.7b00049
J. Chem. Inf. Model. 2017, 57, 1007−1012

Table 2. Performance of RF-Score and X-Score Trained at Different Similarity Levels; The Training Sets Were Constructed Using Equation 1, with Sequence Identity as the Similarity Measure

			RF-Score			X-Score		
cutoff	no. in training set	TC	R_p	R_s	RMSE	R_p	R_s	RMSE
1.000	1105	0.620	0.783	0.769	1.542	0.643	0.707	1.867
0.999	786	0.607	0.681	0.698	1.783	0.643	0.708	1.878
0.998	785	0.606	0.687	0.703	1.774	0.643	0.708	1.878
0.997	781	0.605	0.682	0.693	1.782	0.642	0.708	1.878
0.996	746	0.605	0.682	0.694	1.784	0.646	0.708	1.861
0.995	726	0.601	0.665	0.673	1.814	0.645	0.708	1.863
0.990	692	0.600	0.647	0.657	1.836	0.645	0.707	1.862
0.985	678	0.601	0.639	0.649	1.849	0.646	0.709	1.863
0.980	670	0.600	0.634	0.643	1.855	0.646	0.709	1.862
0.975	640	0.594	0.628	0.633	1.864	0.647	0.709	1.857
0.970	638	0.594	0.619	0.624	1.878	0.647	0.709	1.858
0.965	606	0.597	0.627	0.635	1.869	0.646	0.710	1.863
0.960	606	0.597	0.627	0.635	1.869	0.646	0.710	1.863
0.955	584	0.595	0.623	0.626	1.881	0.645	0.708	1.875
0.950	584	0.595	0.623	0.626	1.881	0.645	0.708	1.875
0.900	565	0.598	0.622	0.621	1.886	0.644	0.707	1.863
0.850	558	0.599	0.631	0.638	1.867	0.645	0.709	1.855
0.800	554	0.598	0.629	0.637	1.871	0.645	0.709	1.857
0.750	553	0.598	0.637	0.640	1.859	0.645	0.709	1.858
0.700	552	0.597	0.632	0.640	1.864	0.645	0.708	1.852
0.650	545	0.595	0.613	0.623	1.896	0.645	0.709	1.853
0.600	542	0.595	0.613	0.622	1.896	0.644	0.708	1.856
0.550	535	0.596	0.611	0.618	1.903	0.644	0.709	1.858
0.500	522	0.591	0.600	0.607	1.913	0.644	0.708	1.856
0.450	508	0.593	0.600	0.602	1.918	0.644	0.707	1.857
0.400	451	0.592	0.620	0.627	1.896	0.647	0.709	1.857
0.350	350	0.591	0.619	0.640	1.925	0.643	0.705	1.870
0.300	181	0.587	0.559	0.569	2.007	0.645	0.697	1.838

based methods outperform empirical scoring functions in terms of scoring power.[21−23]

The Scoring Power of RF-Score Is Significantly Affected by Highly Similar Training Samples. As mentioned earlier, the similarity between the training and test proteins was limited under a given cutoff by removing samples from the original training set. Training on the new training data and validating on the original test data should be able to reflect the impact of protein similarity. As shown in Table 1, when the structural similarity (TM-score) between the training and test data decreases from 1 to 0.98, the scoring power of RF-Score is significantly worse (R_p = 0.647, R_s = 0.648, and RMSE = 1.836) than that for training with the whole training set (R_p = 0.783, R_s = 0.769, and RMSE = 1.542). It is interesting to see that at the TM-score cutoff of 0.98, RF-Score is comparable with X-Score (R_p = 0.647 for both functions). This indicates that the outstanding performance of RF-Score over X-Score reported in the literature can be ascribed to those highly similar training samples (TM-score > 0.98). At more stringent TM-score cutoffs, the performance of RF-Score keeps decreasing, though small fluctuations exist. For example, at the TM-score cutoff of 0.5 (at which two structures are in the same topology[37]), its scoring power becomes R_p = 0.582, R_s = 0.597, and RMSE = 1.96. If the TM-score cutoff is further reduced to 0.4, the R_p and R_s of both RF-Score and X-Score decrease. A possible reason for this drop is the small size of the training set (116), which may not be enough to be used for training. To avoid this effect, we focused mainly on the data when the TM-score cutoff was set to 0.5, which is also recommended for constructing new benchmarks in the future.

A similar situation happens when the training data are filtered on the basis of sequence identity (Table 2). Interestingly, a very dramatic decrease (e.g., R_p decreasing from 0.783 to 0.681) was observed when the cutoff was reduced from 1 to 0.999. This suggests that the RF-Score's performance depends heavily on those training samples that are almost identical to the test samples. When the sequence identity goes down to 0.3, RF-Score's performance drops to R_p = 0.559, R_s = 0.569, and RMSE = 2.007.

In addition, the impact of ligand similarity between the test and training ligands was also investigated as follows. At a specified protein similarity cutoff c, the ligand similarity between the test set and a training set NT(c) (constructed from eq 1) is defined as

$$\mathrm{TC}(c) = \frac{1}{195}\sum_{i=1}^{195} \max_{1\le j\le N_c} \{\mathrm{TC}(i, j)\} \quad (3)$$

where TC(i, j) is the fingerprint-based Tanimoto coefficient (TC) between ligand i in the TS and ligand j in NT(c), as calculated by OpenBabel,[38] and N_c is the number of ligands in NT(c).

As shown in Tables 1 and 2, when the protein structural similarity decreases from 1 to 0.4, the ligand similarity goes down from 0.62 to 0.572. Similarly, when the sequence similarity is reduced from 1 to 0.3, the ligand similarity also declines from 0.62 to 0.587. These data are consistent with the

DOI: 10.1021/acs.jcim.7b00049
J. Chem. Inf. Model. 2017, 57, 1007−1012

assumption that similar proteins tend to bind to similar ligands, which was also reported in the work by Brylinski.[19] Therefore, we may conclude that the high performance of RF-Score can be explained by a high similarity of not only proteins but also ligands between the training and testing sets.

The Scoring Power of X-Score Is Stable and Independent of the Sample Similarity. As a control for RF-Score, using exactly the same division method, we refitted X-Score's energy terms by naïve linear regression on the new training sets and validated the new functions on the original test data. As shown in Tables 1 and 2, though the similarity between training and test data gradually increases, the performance of X-Score does not change much (R_p and R_s are always about 0.64 and 0.71, respectively), no matter whether structural or sequence similarity is considered. This indicates that unlike RF-Score, X-Score is independent of sample similarity, which is a necessary feature for real-world applications, such as molecular docking, virtual screening, and drug design. Indeed, the X-Score function has been successfully incorporated into the popular docking software Autodock Vina.[33]

It Is the Sample Similarity Rather Than the Size of the Training Set That Affects the RF-Score. It was reported that the performance of RF-Score improved with an increase in training set size.[16,34] This conclusion may need to be adjusted as well when considering the sample similarity between the training and test data, which can be roughly seen from Tables 1 and 2. For example, when the TM-score cutoff changes from 0.965 to 0.65, the R_p and R_s of RF-Score are relatively stable (around 0.62), though the size of the training set decreases from 567 to 400. When the sequence identity cutoff changes from 0.97 to 0.4, the R_p and R_s of RF-score are also almost stable (around 0.62), though the size of the training set decreases from 638 to 451. To further demonstrate our hypothesis that the outstanding scoring power of the RF-Score is attributable to the increase in the sample similarity rather than the increase in training set size, we trained the RF-Score by continuously raising the structural similarity cutoff to remove those dissimilar training samples. The results are listed in Table 3. With 538 training samples at TM-score > 0.965, similar performance was achieved (R_p = 0.784, R_s = 0.771, and RMSE = 1.561) as with the original 1105 samples (R_p = 0.783, R_s = 0.769, and RMSE = 1.542). Moreover, only 295 samples with TM-score > 0.997 were used, the R_p was still as high as 0.737, which is much better than that (R_p = 0.689) using the remaining 810 samples with TM-score ≤ 0.997 (Table 1). Similar results were obtained when the sequence identity was used to conduct the above experiments (see Table S3).

Table 3. Performance of RF-Score Trained with Similar Samples Only; In the Training Sets, the Samples Are Required To Have Structural Similarity (with the Test Samples) Higher Than the Specified Cutoffs

			RF-Score		
cutoff	no. in training set	TC	R_p	R_s	RMSE
0.400	989	0.626	0.786	0.772	1.531
0.450	883	0.625	0.769	0.747	1.577
0.500	811	0.630	0.771	0.758	1.570
0.550	775	0.630	0.761	0.746	1.596
0.600	734	0.635	0.764	0.753	1.593
0.650	705	0.634	0.771	0.757	1.587
0.700	689	0.636	0.769	0.750	1.592
0.750	662	0.640	0.764	0.748	1.601
0.800	643	0.644	0.770	0.749	1.581
0.850	629	0.645	0.775	0.760	1.575
0.900	595	0.648	0.786	0.770	1.562
0.950	565	0.643	0.789	0.779	1.556
0.955	557	0.640	0.787	0.776	1.559
0.960	542	0.641	0.780	0.765	1.572
0.965	538	0.641	0.784	0.771	1.561
0.970	523	0.644	0.777	0.762	1.576
0.975	500	0.647	0.777	0.760	1.575
0.980	460	0.653	0.771	0.753	1.585
0.985	420	0.657	0.761	0.746	1.607
0.990	395	0.654	0.747	0.733	1.644
0.995	326	0.657	0.743	0.725	1.662
0.996	312	0.654	0.744	0.726	1.661
0.997	295	0.654	0.737	0.718	1.679
0.998	238	0.670	0.695	0.694	1.760
0.999	141	0.703	0.631	0.608	1.893

CONCLUSIONS

It has constantly been reported that the recently developed machine-learning-based scoring functions outperform conventional scoring functions on binding affinity prediction of protein−ligand interactions. However, careful investigation of the protein structural and sequence similarity between the training and test data shows that the improvement in the scoring power by these methods is attributed to the existence of highly similar proteins in the training set. After removal of such highly similar proteins from the training set and retraining of the scoring functions, the machine-learning-based methods are not better than the conventional scoring functions any more. On the contrary, the performance of conventional functions like X-Score is relatively stable no matter what training data are used for fitting the weights of its energy terms.

The development of new descriptors for improving the prediction of protein−ligand binding affinity is still in demand. We hope that this study will provide guidance for constructing new benchmarks for training and validation of scoring functions for protein−ligand binding affinity prediction.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00049.

Tables S1−S3 (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: yangjy@nankai.edu.cn.

ORCID

Jianyi Yang: 0000-0003-2912-7737

Notes

The authors declare no competing financial interest.

The pairwise similarities between samples in the training and test sets are available at http://yanglab.nankai.edu.cn/download/SF.

ACKNOWLEDGMENTS

This project was supported in part by the National Natural Science Foundation of China (11501306, 81520108019), the Ministry of Science and Technology 973 Project (2014CB542800), and the Thousand Youth Talents Plan of China.

DOI: 10.1021/acs.jcim.7b00049
J. Chem. Inf. Model. 2017, 57, 1007−1012

REFERENCES

(1) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based Scoring Function to Predict Protein-Ligand Interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.

(2) Joseph-McCarthy, D.; Baber, J. C.; Feyfant, E.; Thompson, D. C.; Humblet, C. Lead Optimization via High-Throughput Molecular Docking. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 264−274.

(3) Ding, B.; Wang, J.; Li, N.; Wang, W. Characterization of Small Molecule Binding. I. Accurate Identification of Strong Inhibitors in Virtual Screening. *J. Chem. Inf. Model.* **2013**, *53*, 114−122.

(4) 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.

(5) Volkamer, A.; Kuhn, D.; Rippmann, F.; Rarey, M. DoGSiteScorer: a Web Server for Automatic Binding Site Prediction, Analysis and Druggability Assessment. *Bioinformatics* **2012**, *28*, 2074−2075.

(6) Jorgensen, W. L. Efficient Drug Lead Discovery and Optimization. *Acc. Chem. Res.* **2009**, *42*, 724−733.

(7) Trosset, J.-Y.; Vodovar, N. Structure-Based Target Druggability Assessment. *Methods Mol. Biol.* **2013**, *986*, 141−164.

(8) Tanrikulu, Y.; Kruger, B.; Proschak, E. The Holistic Integration of Virtual Screening in Drug Discovery. *Drug Discovery Today* **2013**, *18*, 358−364.

(9) Huang, N.; Kalyanaraman, C.; Bernacki, K.; Jacobson, M. P. Molecular Mechanics Methods for Predicting Protein-Ligand Binding. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5166−5177.

(10) Mooij, W. T.; Verdonk, M. L. General and Targeted Statistical Potentials for Protein-Ligand Interactions. *Proteins: Struct., Funct., Genet.* **2005**, *61*, 272−287.

(11) Huang, S. Y.; Grinter, S. Z.; Zou, X. Scoring Functions and Their Evaluation Methods for Protein-Ligand Docking: Recent Advances and Future Directions. *Phys. Chem. Chem. Phys.* **2010**, *12*, 12899−12908.

(12) 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.

(13) Ain, Q. U.; Aleksandrova, A.; Roessler, F. D.; Ballester, P. J. Machine-Learning Scoring Functions to Improve Structure-based Binding Affinity Prediction and Virtual Screening. *WIREs Comput. Mol. Sci.* **2015**, *5*, 405−424.

(14) Liu, J.; Wang, R. Classification of Current Scoring Functions. *J. Chem. Inf. Model.* **2015**, *55*, 475−482.

(15) Li, Y.; Han, L.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results. *J. Chem. Inf. Model.* **2014**, *54*, 1717−1736.

(16) Ballester, P. J.; Mitchell, J. B. A Machine Learning Approach to Predicting Protein-Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics* **2010**, *26*, 1169−1175.

(17) Ballester, P. J.; Schreyer, A.; Blundell, T. L. Does a More Precise Chemical Description of Protein-Ligand Complexes Lead to More Accurate Prediction of Binding Affinity? *J. Chem. Inf. Model.* **2014**, *54*, 944−955.

(18) Zilian, D.; Sotriffer, C. A. SFCscore(RF): A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2013**, *53*, 1923−1933.

(19) Brylinski, M. Nonlinear Scoring Functions for Similarity-Based Ligand Docking and Binding Affinity Prediction. *J. Chem. Inf. Model.* **2013**, *53*, 3097−3112.

(20) Li, G. B.; Yang, L. L.; Wang, W. J.; Li, L. L.; Yang, S. Y. ID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein-Ligand Interactions. *J. Chem. Inf. Model.* **2013**, *53*, 592−600.

(21) 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.

(22) Ballester, P. J.; Mitchell, J. B. Comments on "Leave-Cluster-Out Cross-Validation Is Appropriate for Scoring Functions Derived from Diverse Protein Data Sets": Significance for the Validation of Scoring Functions. *J. Chem. Inf. Model.* **2011**, *51*, 1739−1741.

(23) Gabel, J.; Desaphy, J.; Rognan, D. Beware of Machine Learning-Based Scoring Functions-On the Danger of Developing Black Boxes. *J. Chem. Inf. Model.* **2014**, *54*, 2807−2815.

(24) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-wide Collection of Binding Data: Current Status of the PDBbind Database. *Bioinformatics* **2015**, *31*, 405−412.

(25) Wang, R.; Lu, Y.; Fang, X.; Wang, S. An Extensive Test of 14 Scoring Functions Using the PDBbind Refined Set of 800 Protein-Ligand Complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114−2125.

(26) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.

(27) CASF-2007 Data. http://www.pdbbind.org.cn/download/CASF-2007.tar.gz (accessed June 2016).

(28) Zhang, Y.; Skolnick, J. Scoring Function for Automated Assessment of Protein Structure Template Quality. *Proteins: Struct., Funct., Genet.* **2004**, *57*, 702−710.

(29) Zhang, Y.; Skolnick, J. TM-align: a Protein Structure Alignment Algorithm Based on the TM-score. *Nucleic Acids Res.* **2005**, *33*, 2302−2309.

(30) TM-align. http://zhanglab.ccmb.med.umich.edu/TM-align/ (accessed January 2017).

(31) Needleman, S. B.; Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **1970**, *48*, 443−453.

(32) Zhang, Y. NW-align. http://zhanglab.ccmb.med.umich.edu/NW-align/ (accessed December 2016).

(33) 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.

(34) Li, H.; Leung, K. S.; Wong, M. H.; Ballester, P. J. Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets. *Mol. Inf.* **2015**, *34*, 115−126.

(35) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(36) X-Score, version 1.3. http://www.sioc-ccbg.ac.cn/?p=42 (accessed December 2016).

(37) Xu, J.; Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **2010**, *26*, 889−95.

(38) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 33.

DOI: 10.1021/acs.jcim.7b00049
J. Chem. Inf. Model. 2017, 57, 1007−1012