ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

XLPFE: A Simple and Effective Machine Learning Scoring Function for Protein−Ligand Scoring and Ranking

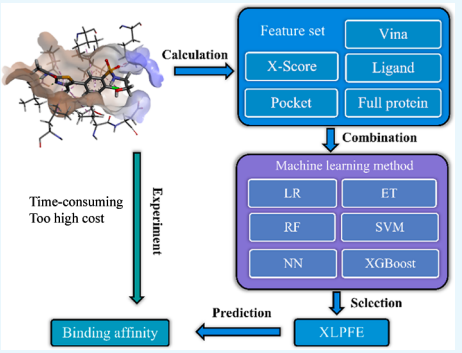
Lina Dong, Xiaoyang Qu, and Binju Wang*

Cite This: *ACS Omega* 2022, 7, 21727−21735

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Prediction of protein−ligand binding affinities is a central issue in structure-based computer-aided drug design. In recent years, much effort has been devoted to the prediction of the binding affinity in protein−ligand complexes using machine learning (ML). Due to the remarkable ability of ML methods in nonlinear fitting, ML-based scoring functions (SFs) can deliver much improved performance on a selected test set, such as the comparative assessment of scoring functions (CASF), when compared to the classical SFs. However, the performance of ML-based SFs heavily relies on the overall similarity of the training set and the test set. To improve the performance and transferability of an SF, we have tried to combine various features including energy terms from X-score and AutoDock Vina, the properties of ligands, and the statistical sequence-related information from either the binding site or the full protein. In conjunction with extreme trees (ET), an ML model, we have developed XLPFE, a new SF. Compared with other tested methods such as X-score, AutoDock Vina, ΔvinaXGB, PSH-ML, or CNN-score, XLPFE achieves consistently better scoring and ranking power for various types of protein−ligand complex structures beyond the CASF, suggesting that XLPFE has superior transferability. In particular, XLPFE performs better with metalloenzymes. With its faster speed, improved accuracy, and better transferability, XLPFE could be usefully applied to a diverse range of protein−ligand complexes.



1. INTRODUCTION

Computer-aided drug design (CADD) can accelerate the process of drug development, saving much time and cost compared to experimental procedures.[1] Drug screening is important in the discovery both of new drugs and new applications for old drugs and is among the most important tasks of CADD. To maximize the effects of drugs while minimizing their side effects, the interactions between the drug and the target should be fully understood. Accordingly, an accurate prediction of binding affinity between drugs and their target proteins is the key to drug screening.[2]

To date, many theoretical methods, such as quantum mechanics/molecular mechanics,[3,4] free energy perturbation,[5,6] and thermodynamic integration,[7,8] have been developed to accurately predict the binding affinity of ligands for proteins. Unfortunately, the routine applications of these methods in high-throughput compound screening are hampered by high computational costs. Scoring functions (SFs) however have much lower computational costs and have found wide application in the prediction of binding affinity in protein−ligand complexes.[9] Classical SFs can be force field-based,[10,11] empirical,[12−15] or knowledge-based.[16,17] Force field-based SFs are usually based on calculated energies, while empirical SFs are based on a hypothetical equation with linear regression (LR) parameters. In knowledge-based SFs, the energy terms are derived from the statistics of protein−ligand interactions. In all these SFs, a predetermined functional form is assumed to characterize the relationship between binding affinities and the relevant parameters.

With the recent rapid development of artificial intelligence, much effort has been devoted to develop the machine learning[18−21] (ML) based SFs for binding affinity prediction of the protein−ligand complexes. Compared to the classical SFs, ML can automatically learn to use generalized nonlinear functional forms and feature information from training data, which can improve the accuracy of binding affinity prediction.[22] ML can be classified into two categories: traditional ML-based methods and deep learning (DL)-based methods.[23] In traditional ML-based methods, the correlations between the binding affinity and the selected features are calculated via nonlinear regression using classical ML algorithms. For example, Zilian and Sotriffer proposed a method named SFCscoreRF, in which a random forest (RF) method was used to perform the nonlinear fitting of the SFCscore descriptors.[24] Ballester and co-workers developed a series of SFs using an RF algorithm and energy terms taken mainly from RF-Score.[25−27] Wang and Zhang

Received: March 22, 2022
Accepted: May 30, 2022
Published: June 13, 2022

introduced a new method named ΔvinaRF20, which was trained by RF and is based on 20 features, including five AutoDock Vina interaction terms, five ligand-dependent terms, and 10 buried solvent-accessible surface area-related features.[28] Very recently, Xia and co-workers have developed a persistent spectral hypergraph (PSH) model-based ML SF (PSH-ML), which achieved a high Pearson's correlation coefficient (R_p) of 0.855 for the scoring power in a CASF-2016 benchmark test set.[29] Unlike ML, DL-based methods usually do not require feature engineering, which can directly convert the original structural data into the high-dimensional neural network for subsequent regression. In 2017, Koes and co-workers developed a convolutional neural network (CNN) SF, in which the input is based on a comprehensive three-dimensional representation of protein−ligand interactions.[30] Other successful applications of DL are TopologyNet,[31] KDEEP,[32] Pafnucy,[33] InteractionGraphNet,[34] and others.[35−40] Compared with the DL methods that rely on complex models, the traditional ML method has the advantages of simplicity, fast training speed, and diminished dependence on data and computing equipment. Therefore, many current efforts still rely on the use of traditional ML to improve the scoring power of SFs.[27,41−48]

Though the ML-based SFs can achieve much improved performance on the selected test set such as the comparative assessment of scoring functions (CASF),[49] the performance can be significantly diminished as the overall similarity thresholds between the training set and the test set decrease. In some tests, the performances of ML-based SFs are even inferior to those of the conventional SFs.[50] Specifically, the binding affinity prediction in metalloenzymes is challenging due to the complex interactions between the ligand, metals, and the protein environment.[15,51] As such, extensive efforts have been devoted to improve the prediction accuracy for metallocomplexes.[51,52] In terms of these key issues, we wonder if the selection of more suitable features and ML methods could significantly improve the performance and transferability of SFs when compared to the previous results. To pursue this, we combined diverse features, including energy terms from X-score[12] and AutoDock Vina,[13] properties of the ligand, and the sequence information from either the active site or the full protein. Through this elaborate selection and training, we have developed a new SF, XLPFE. Compared with other tested methods such as X-score,[12] AutoDock Vina,[13] ΔvinaXGB,[53] PSH-ML,[29] and CNN-Score,[30] XLPFE consistently achieves scoring power and ranking power better than that of CASF[49] for various types of protein−ligand complex structures, suggesting that XLPFE has much-improved transferability. In particular, XLPFE can achieve robust performance in the scoring and ranking of binding affinity for metalloproteins.

2. METHODS

2.1. Data Sets. The PDBbind database (http://www.pdbbind-cn.org/),[54−56] developed and maintained by Wang and co-workers, provides a comprehensive collection of the experimentally measured binding affinity data for complexes and the PDB structures. It has been widely used for the development and validation of SFs. The refined set provided by the PDBbind database prior to 2018 was used as the training set, and the refined post-2018 data set was selected as test set 1.[57] CASF-2016[49] was used as our test set 2. The number of complexes contained in each set is given in Table 1. In 2020, Wang and co-workers compiled data sets based on the PDBbind refined set by removing redundant samples using various similarity thresholds.[50] We built the standard training sets (the second column of Table 2) similarly to evaluate and select different models with the CASF-2016 benchmark. In order to reduce the similarity between the training set and test sets, we also have reduced the size of the test set but maintained the size of the training set. Also, the numbers of standard test sets are listed in the third column of Table 2.

Table 1. Summary of the Data Sets

	source	numbers
training set	PDBbind refined set (before 2018)	4190
test set 1	PDBbind refined set (after 2018)	394
test set 2	CASF-2016	285

Table 2. Sample Size of the Nonredundant Set under Different Similarity Thresholds

sequence similarities (%)	numbers of training sets	numbers of test sets
100	4190	285
95	3949	158
90	3390	57
85	2824	23
80	2318	

2.2. Feature Sets. Our feature set includes five subsets derived from the energy term of AutoDock Vina (V), the energy term of X-score (X), the statistic feature related to the ligand (L), the statistic sequence-based feature related to the pocket (P), and the statistic sequence-based feature related to the full protein (F) (Table 3).

Table 3. Summary of the Feature Sets

feature set	terms	dimension
AutoDock Vina	58 terms from the Vina source code	58
X-score	VDW, HB, RT, HS, HM, and HP	6
ligand	charge; C, N, O, H, F, P, S, Cl, Br, and I element numbers; and 1, 2, 3, am, and ar bond numbers	16
pocket	20 amino acid numbers and crystal H_2O number	21
full protein	20 amino acid numbers	20

For the V subset, 58 features from the source code of Vina were selected,[13] including protein−ligand interaction terms and a set of ligand properties. Besides the Gaussian, repulsion, hydrogen bonding (HB), and hydrophobic terms included in AutoDock Vina, some other terms, such as simple property counts, electrostatic interactions, AutoDock4 desolvation effects, nonhydrophobic contacts, and Lennard Jones 4−8 van der Waals interactions (Table S1), were also included.

X-score,[12] developed by Wang and co-workers, is an empirical SF, which is composed of four major energy terms with respect to van der Waals interactions (VDW), HB, deformation effects (RT), and hydrophobic effects. According to the different approaches to its calculation, the hydrophobic effect can be further sectioned into hydrophobic pairs (HP), hydrophobic matching (HM), and hydrophobic surface (HS). HP calculates the hydrophobic energy by counting the hydrophobic contact pairs between the protein and ligand, whereas HM and HS compute this energy using an HM algorithm and an HS algorithm, respectively (see the Part S1 section for more details).[58−60] These six terms from X-score constitute a subset of X.

https://doi.org/10.1021/acsomega.2c01723
ACS Omega 2022, 7, 21727−21735

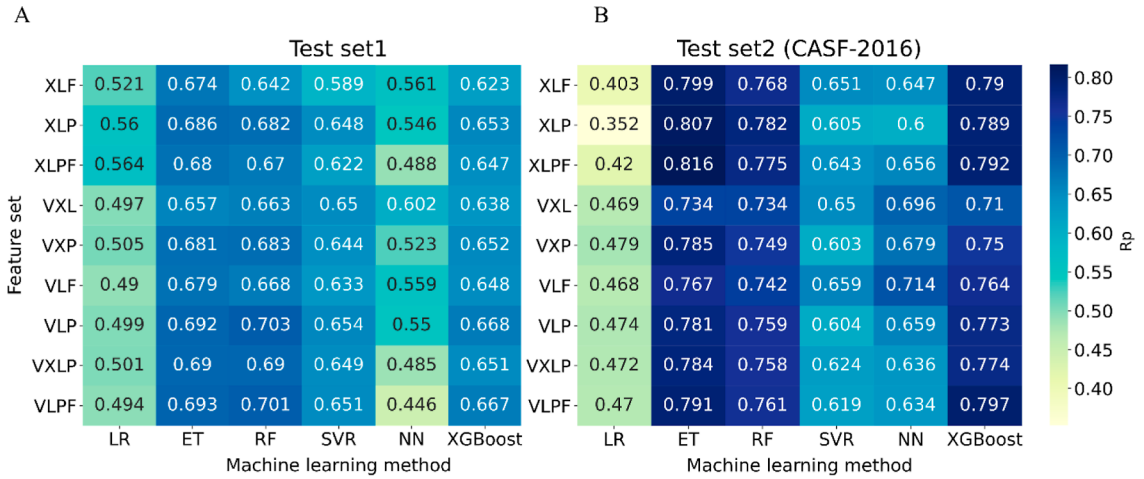


Figure 1. Pearson correlation coefficients between the experimental data and the predicted binding affinities on (A) test set 1 and (B) test set 2 (CASF-2016) for combinations of different feature sets and ML methods. The dimension of features increases from the top to the bottom. The darker the color (blue), the higher the correlation, and the lighter the color (yellow), the lower the correlation.

For the information of the ligand (L), a total of 12 features were considered, including the total charge of the ligand, the number of atoms of each element (C, N, O, F, P, S, Cl, Br, I, and H), and the number of each type of bond—single bond, double bond, triple bond, amide bond, and aromatic bond.

For the information from the binding site (P), a total of 21 features were selected, including the quantity of 20 amino acids and the number of water molecules in the site. For the information of the full protein (F), the quantitative distribution of 20 amino acids was taken into account.

In order to eliminate the influence of unit and scale differences between features, each feature was subsequently standardized using the sklearn.preprocessing.StandardScaler class.

2.3. ML Methods. Scikit-Learn[61] 0.24.1 was used to generate machine-learning models. LR, tree-related models including extra trees (ET),[62] RF,[63] extreme gradient boosting (XGBoost),[64] support vector regressions (SVR),[65] and neural networks (NN)[66] were selected. The ML models can be used to predict the binding affinity according to

$$F_i(x_1, x_2, ..., x_n) = \mathrm{pre}_i \tag{1}$$

where $(x_1,x_2,...,x_n)$ is the vector of input features and n is the number of features. F is the machine-learning model that adopts a nonlinear function. The output is the predicted binding affinity for protein−ligand complex i. Different models obtain the minimum mean absolute error (MAE) in different ways according to

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{pre}_i - \mathrm{exp}_i) \tag{2}$$

where pre_i and exp_i are the predicted and experimental binding affinities of the protein−ligand complex, respectively, and i and N are the number of samples in the training set. Fivefold cross-validation was used to efficiently search the hyperparameter space for each model. After training a call (five calls in total), the cost function (the MAE of predictions on the subset) across folds is returned to the estimator, which in turn chooses a new hyperparameter configuration for the next call using its acquisition function (eq 1) to further decrease the model's cost function (eq 2). A brief description and tuned hyperparameters of each ML method are shown in Table S2. The results from all models are the average of 10 repeated experiments.

2.4. Performance Evaluation. The R_p is a measure of the linear dependence of the predicted binding affinity values on the experimental values according to eq 3. The Spearman correlation coefficient (R_s) can measure the strength of an association between the predicted and experimental binding affinity values according to a monotonic function (eq 4). RMSE is the root mean square error between the predicted binding affinity and the experimental value (eq 5). These are calculated as follows

$$R_p = \frac{\sum_{i=1}^{N}(\mathrm{pre}_i - \mathrm{pre}_{ave})(\mathrm{exp}_i - \mathrm{exp}_{ave})}{\sqrt{\sum_{i=1}^{N}(\mathrm{pre}_i - \mathrm{pre}_{ave})^2 \sum_{i=1}^{N}(\mathrm{exp}_i - \mathrm{exp}_{ave})^2}} \tag{3}$$

$$R_s = 1 - \frac{6\sum_{i=1}^{N}(r\mathrm{pre}_i - r\,\mathrm{exp}_i)}{N(N^2-1)} \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{pre}_i - \mathrm{exp}_i)^2} \tag{5}$$

pre_i is the binding affinity from the given SF on the ith complex in the test set; exp_i is the experimental binding constant (in logarithm units, $\log K_a$) of this complex; pre_{ave} and exp_{ave} are the corresponding averages; $r\mathrm{pre}_i$ is the rank of the binding affinity of the ith complex; $r\mathrm{exp}_i$ is the rank of the experimental binding affinity of this complex; and N is the total number of samples. All predicted values were based on the crystal structures of the protein−ligand complexes.

3. RESULTS

3.1. Selection of Feature Set Combinations. Nine feature sets with different combinations (XLF, XLP, XLPF, VXL, VXP, VLF, VLP, VXLP, and VLPF) were applied to the multiple LR, and five ML models (ET, RF, SVR, NN, and XGBoost) were used. Figure 1 compares the performance of different feature sets and different ML models on the test set. As the dimension of features increases from the top to the bottom, it can be seen that the performance of R_p fails to improve steadily. However, it can be seen that the tree-related ML methods (ET, RF, and XGBoost) in general have a performance that is better

https://doi.org/10.1021/acsomega.2c01723
ACS Omega 2022, 7, 21727−21735

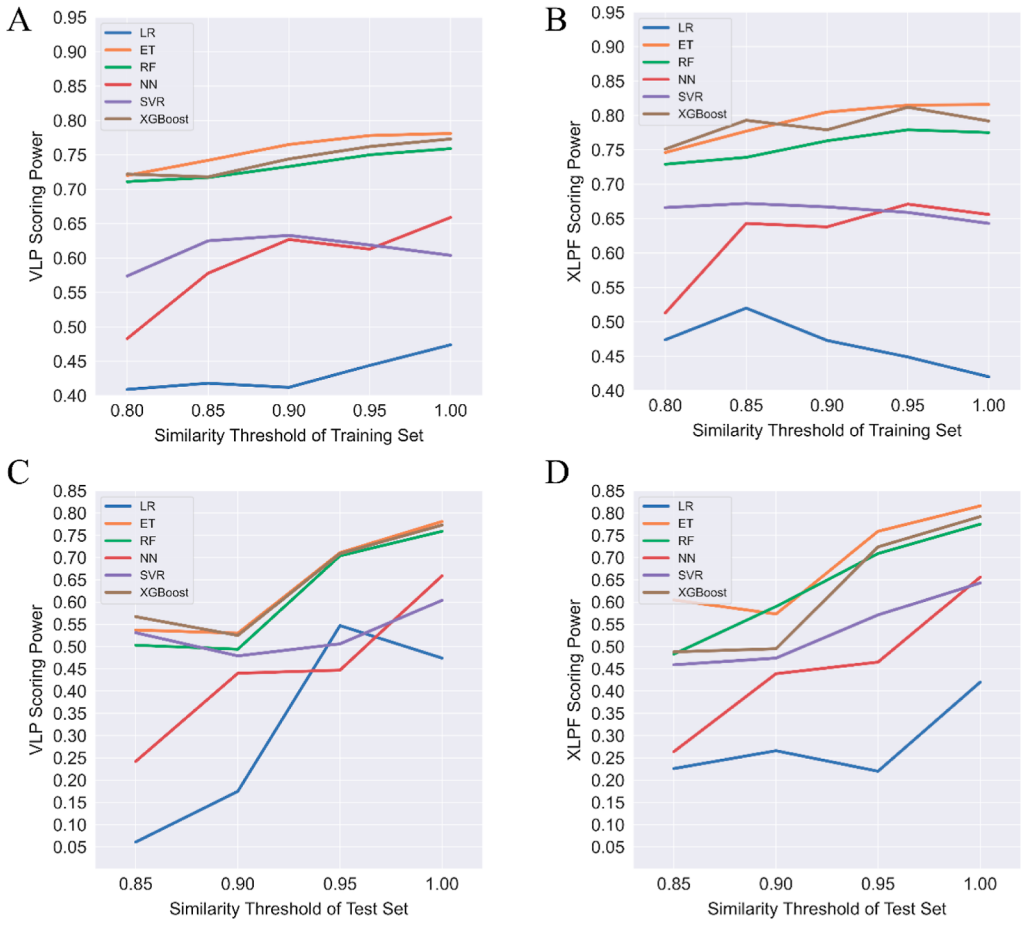


Figure 2. Pearson correlation coefficients between the experimental data and the predicted binding affinities for different sequence similarity thresholds of the training set using(A) VLP feature set and (B) XLPF feature set and different sequence similarity thresholds of the test set using(C) VLP feature set and (D) XLPF feature set. Different ML methods are shown in different colors: LR in blue, ET in orange, RF in green, NN in red, SVR in purple, and XGBoost in brown.

than that of the other two (SVR and NN). On test set 1, VLP achieves the highest R_p value of 0.703, while XLPF has the highest R_p value of 0.816 on test set 2 (CASF-2016). Consistent with the previous observation, the highest R_p also corresponds to the lowest RMSE on the same feature set (Table S4). In particular, the VLP and XLPF feature sets achieve the low RMSE values of 1.30 and 1.36, respectively. Thus, both the VLP and XLPF feature sets were selected for further testing. It can be seen that in most combinations, performances in test set 2 are better than in test set 1. The different performance may be related to data diversity and similarity between the training set and test set.

3.2. Comparison of Different ML Methods. According to the different similarity thresholds in the training set, the stability of LR and ML methods was assessed on the VLP and XLPF feature sets. Figure 2A, B shows the performance of different machine models on both feature sets with different similarity thresholds of the training set. In order to reduce the similarity between the training set and test sets, we also have reduced the size of the test set but maintained the size of the training set (Table 2). Figure 2C,D shows the performance of different machine models on both feature sets with different similarity thresholds of the test set. Consistent with our test study described in the previous section, the tree-related methods, ET, RF, and XGBoost, perform better than the other two, SVR and NN, albeit the performance decreases with the decrease in the similarity.[50] Overall, ET is the best model in view of its stable and near-best performance. As shown in Figure 2, the R_p obtained from ET using the XLPF feature set is generally higher than that produced using VLP, and XLPF with 63 features has lower feature dimensions than VLP with 95 features. Thus, the combination of XLPF feature sets and the ET model (labeled as the XLPFE model) was selected in our subsequent studies.

As summarized in Figure 1 and Tables S3 and S4, XLPFE achieves R_p = 0.68 and RMSE = 1.34 on test set 1, while the corresponding values are 0.816 and 1.41 with test set 2 (CASF-2016). In terms of the CASF-2016 data set, we further tested and compared more than 30 common SFs included in the data set. As summarized in part 1, model XLPFE achieved a Pearson correlation coefficient of 0.816 and a Spearman correlation coefficient of 0.66, which identifies it as one of the best SFs in the CASF-2016 data set. The additional tests on the performance of XLPFE on CASF-2016 and its subsets can be found in Table S6.

3.3. Evaluation of Feature Importance. Figure 3A shows the calculated feature correlation of XLPFE. It can be seen that there is no high correlation between most features. The highest correlation is X-Score-related HP, HS, and HM. The three hydrophobic algorithms are highly related to the energy calculation of hydrophobic action, but they do not exceed 0.9.[67] The correlation between features and experimental values is basically proportional to the importance of features (Figure 3B). When the correlation between features and experimental values is higher, its importance is often higher. The features

https://doi.org/10.1021/acsomega.2c01723
ACS Omega 2022, 7, 21727−21735

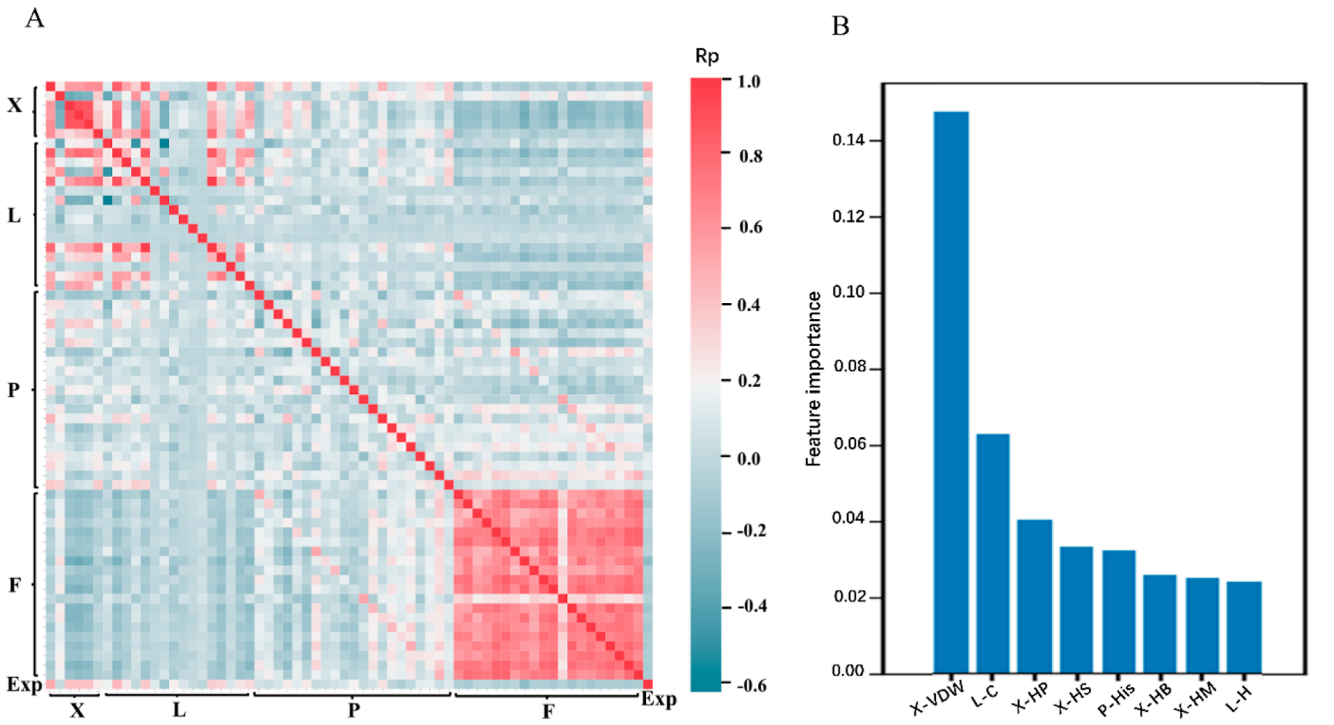


Figure 3. (A) Correlation matrix of features and experimental values. (B) Feature importance. Feature importance values are calculated based on the number of times a feature is used to split the data across all trees. Here, the eight most significant features are shown.

Table 4. Performance of XLPFE, X-Score, AutoDock Vina, ΔvinaXGB, CNN-Score, PSH-ML, and Lin_F9 Evaluated against a Set Consisting of 15 Selected Diverse Biological Targets[a]

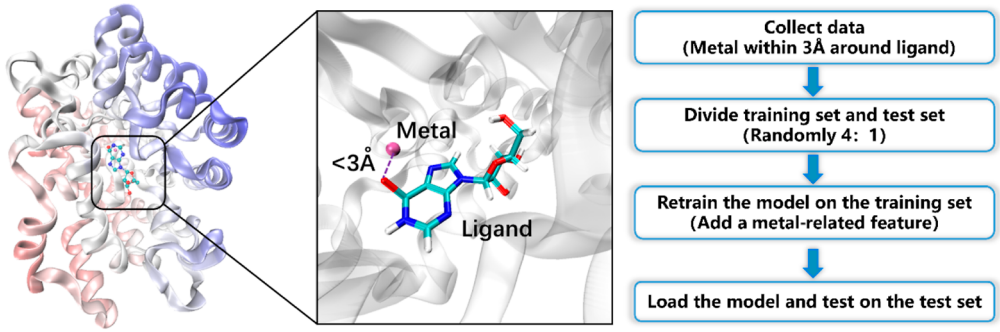
		XLPFE			X-Score			AutoDock Vina			Δvina XGB			CNN-Score			PSH-ML			Lin_F9		
Test target	Number	Rp	Rs	RMSE	Rp	Rs	RMSE	Rp	Rs	RMSE	Rp	Rs	RMSE	Rp	Rs	RMSE	Rp	Rs	RMSE	Rp	Rs	RMSE
BACE-1	92	0.819	0.738	0.98	0.806	0.751	1.15	0.671	0.734	1.77	0.731	0.713	1.75	0.830	0.783	1.61	0.824	0.761	1.09	0.813	0.746	1.46
CHK1	21	0.862	0.758	1.53	0.783	0.684	1.60	0.868	0.799	1.58	0.785	0.695	1.23	0.719	0.639	1.08	0.912	0.891	1.35	0.874	0.771	1.60
DPP4	27	0.467	0.315	1.72	0.413	0.289	1.90	0.355	0.435	2.02	0.526	0.412	1.34	0.675	0.534	0.90	0.733	0.500	1.23	0.607	0.402	1.67
ER	16	0.527	0.570	2.37	0.448	0.418	1.23	0.387	0.422	1.38	0.447	0.458	1.09	0.510	0.471	0.82	0.370	0.443	1.25	0.546	0.455	1.62
GluR2	11	0.384	0.418	1.18	0.273	0.400	1.60	0.340	0.441	1.93	0.201	0.368	1.35	0.591	0.841	1.01	0.220	0.106	1.07	0.522	0.621	2.20
HIV PR	16	0.560	0.590	1.95	0.426	0.461	1.08	0.377	0.549	2.59	0.770	0.820	1.27	0.476	0.505	1.33	0.584	0.652	1.00	0.527	0.543	1.78
HSP90	25	0.773	0.788	1.12	0.874	0.880	0.92	0.445	0.471	2.12	0.499	0.477	2.08	0.743	0.706	2.22	0.768	0.685	1.38	0.811	0.735	0.94
LTA-4H	25	0.756	0.839	2.24	0.770	0.821	2.35	0.628	0.618	1.90	0.763	0.842	1.77	0.658	0.772	2.43	0.723	0.805	2.15	0.750	0.809	1.58
P38a	22	0.759	0.711	1.72	0.668	0.679	0.97	0.668	0.598	1.26	0.739	0.618	0.93	0.588	0.438	1.04	0.812	0.550	0.79	0.704	0.539	1.08
PDE4B	28	0.794	0.804	1.65	0.799	0.816	1.70	0.789	0.800	1.86	0.748	0.711	1.35	0.752	0.724	1.12	0.766	0.665	1.57	0.748	0.766	1.90
PDK1	11	0.801	0.845	0.95	0.836	0.845	1.04	0.403	0.436	1.66	0.421	0.436	1.72	0.169	0.236	1.93	0.675	0.664	1.31	0.245	0.455	1.80
PTP1B	23	0.667	0.653	1.01	0.659	0.689	1.05	0.140	0.187	1.31	0.355	0.283	1.19	0.567	0.550	1.24	0.552	0.584	1.04	0.512	0.486	1.15
Renin	22	0.523	0.389	2.44	0.398	0.463	1.16	0.187	0.213	2.19	0.403	0.460	1.33	0.676	0.662	0.77	0.517	0.395	1.15	0.301	0.375	1.50
SRC	30	0.891	0.750	1.47	0.858	0.736	1.61	0.662	0.611	3.19	0.755	0.663	2.41	0.889	0.785	1.36	0.863	0.705	1.34	0.769	0.757	1.92
Thrombin	21	0.795	0.692	1.31	0.725	0.683	1.55	0.180	0.265	2.80	0.534	0.568	1.81	0.687	0.652	1.78	0.795	0.653	1.23	0.668	0.647	1.64
Average		0.692	0.657	1.58	0.649	0.641	1.39	0.473	0.505	1.97	0.578	0.568	1.51	0.635	0.620	1.38	0.674	0.604	1.26	0.626	0.607	1.59

[a]BACE-1, β-secretase 1; CHK1, serine/threonine-protein kinase chk1; DPP4, dipeptidyl peptidase 4; ER, estrogen receptor; GluR2, glutamate receptor 2; HIV PR, hiv-1 protease; HSP90, heat shock protein 90; LTA-4H, leukotriene A-4 hydrolase; P38a, mitogen-activated protein kinase 14; PDE4B, camp-specific3′,5′-cyclic phosphodiesterase 4b; PDK1, 3-phosphoinositide-dependent protein kinase 1; PTP1B, protein tyrosine phosphatase 1B; and SRC, proto-oncogene tyrosine protein kinase src. The more intensely red the table, the smaller the value, and the more intensely green the table the larger the value.

mentioned with high correlation—HP, HS, and HM also have similar importance.

The feature importance was based on the number of times a feature is used to split the data across all trees. Figure 3B shows the feature importance analysis of model XLPFE. It is seen that the VDW item generated using X-score ranks first in feature importance. Among the top eight features selected from feature importance, five items are from subset X-score, two are from ligand, and one is from the protein active site. Among the top 30 features in Figure S3, only one feature, the occurrence of Ser, is from the full protein. This is consistent with previous studies, which showed that the use of active site sequences can improve

https://doi.org/10.1021/acsomega.2c01723
ACS Omega 2022, 7, 21727−21735



Type	number of training sets	number of test sets	XLPFE			PSH-ML			Lin_F9			X-Score		
			Rp	Rs	RMSE	Rp	Rs	RMSE	Rp	Rs	RMSE	Rp	Rs	RMSE
Zn	1034	259	0.732	0.745	1.30	0.735	0.742	1.31	0.422	0.444	2.39	0.360	0.374	1.91
Mg	328	83	0.448	0.569	1.28	0.518	0.507	1.22	0.194	0.174	2.13	0.290	0.262	1.42
Mn	223	57	0.654	0.620	1.20	0.630	0.590	1.21	0.240	0.294	2.55	0.211	0.266	1.68
Ca	107	26	0.639	0.701	1.22	0.571	0.559	1.34	0.477	0.359	2.61	0.570	0.357	1.34
Fe	56	14	0.764	0.735	1.12	0.757	0.719	1.08	0.625	0.630	1.63	0.672	0.722	1.24
Average			0.647	0.674	1.22	0.642	0.623	1.23	0.392	0.377	2.26	0.421	0.396	1.52

Figure 4. Predicted affinities in complexes containing different types of metal atoms. Performances of XLPFE, PSH-ML, Lin_F9, and X-Score evaluated against a set consisting of five kinds of selected metal contained targets are listed in the table at the bottom of the figure.

the prediction of the affinity.[68,69] For the occurrence of amino acids in the protein active site, it can be seen that His, Gly, Tyr, and Trp rank in the top position (Figure S3), probably because they can form hydrogen bonds with ligands and thus play key roles in ligand binding. Leu and Phe also rank quite high, possibly due to their contribution to the formation of hydrophobic pockets.

3.4. Performance of XLPFE on Diverse Biological Targets. In this section, the performance of XLPFE with different biological targets is evaluated. For this purpose, we collected an expanded test set covering 15 different biological targets, including various kinases, secretory enzymes, and hydrolases. For each target, we collected all the relevant crystal structures of complexes that contain small molecule ligands with relatively reliable bioactivity data available from the PDBbind general set (for the PDB ID list of all the collected crystal structures of the 15 selected targets, see Table S7). In order to maintain a reasonable data size, targets including less than 10 complexes are not selected in data sets. For each target, the performance of XLPFE on the collected complexes was evaluated and characterized in terms of R_p, R_s, and RMSE. Table 4 shows the calculated R_p, R_s, and RMSE values for each target. The predicted R_p values for the 15 targets are in the range 0.384−0.891, and the average value is 0.692. The predicted R_s values are between 0.315 and 0.845, and the average is 0.657. The RMSE values are between 0.90 and 2.44, and the average is 1.58. For comparison, we also tested other popular SFs, including the traditional SFs X-score, AutoDock Vina, and Lin_F9, as well as the ML-based SFs ΔvinaXGB[53] and PSH-ML[29] and DL-based SF CNN-Score.[30,70] In ΔvinaXGB, a feature set of Vina's 58 energy terms and other interaction-related terms have been applied to the XGBoost algorithm, which achieved a scoring power of 0.796 on CASF-2016.[53] PSH-ML uses the graph and spectrum theory to generate additional features and has an outstanding scoring power on the benchmark set CASF-2016 with an obtained R_p value of 0.855.[29] It is noted that the results of CNN-Score we used herein were calculated from CNN affinity of GNINA.[70]

As shown in Table 4, XLPFE outperforms AutoDock Vina, ΔvinaXGB, and Lin_F9 in terms of average R_p, R_s, and RMSE. Compared to those in X-score, CNN-score, and PSH-ML, the predicted average R_p and R_s values are improved in XLPFE, although the average RMSE is lower than that in X-score, CNN-score, and PSH-ML. Overall, XLPFE achieved the highest average values of both R_p and R_s among these seven SFs, suggesting that it has robust scoring and ranking power. These results clearly indicate that XLPFE can achieve good and consistent performance on all the test targets, implying that it can be used for a broad range of biological targets. The sensitivity of our method in predicting the binding affinity of similar molecules for the same target in drug design was also confirmed.[71] Detailed discussions are listed in Part S2 of the Supporting Information. This shows that XLPFE has a considerable ability to correctly sort structurally similar ligands that have the same target.

3.5. Performance with Complexes Containing Different Types of Metal Atoms. To further demonstrate the performance of XLPFE, we performed binding affinity predictions in various metalloproteins. For this purpose, we collected the metalloproteins from PDBbind and classified them according to metal types (Figure 4). For Zn-containing metalloproteins, the training set and test set used in the previous study[15] were used for pretraining and testing (see Part S3 in the Supporting Information for detailed results), and our method performed well. For metalloproteins, the selection criterion is based on the distance between the metal and any atom of the ligand, which should be less than 3.0 Å. Accordingly, we have selected over 50 metalloproteins containing Zn, Fe, Mg, Mn, or Ca. For each metal type, the corresponding metalloproteins were randomly divided into a training set (80%) and a test set (20%).

In the feature set, one additional feature corresponding to the number of metals within 3 Å of any atoms of the ligand was

https://doi.org/10.1021/acsomega.2c01723
ACS Omega 2022, 7, 21727−21735

added (Table S10). For comparison, we also tested the ML-based SF PSH-ML[29] and the classical SF Lin_F9[14]. PSH-ML uses the graph and spectrum theory to generate additional features and has outstanding scoring power on the benchmark set CASF-2016 with an obtained R_p value of 0.855.[29] Lin_F9 is based on a linear combination of nine empirical terms, including one energy term to describe the metal−ligand bonding interactions.[14] It was found that Lin_F9 achieves better performance than Vina and X-score with metalloprotein complexes. For a consistent comparison, the same training and test sets as those used for XLPFE were applied to PSH-ML, and all results are summarized in Figure 4.

Among the four SFs, XLPFE achieves the best overall performance among all indexes, with the highest average R_p and R_s values, as well as the lowest average RMSE. In particular, the performance of XLPFE is much better than that of Lin_F9 or X-score. For Zn-metalloproteins, XLPFE shows a similar performance to PSH-ML, but for Mn, Ca, and Fe, XLPFE shows a better performance than PSH-ML. However, PSH-ML outperforms XLPFE in terms of Mg. XLPFE is much faster than PSH-ML. With PDB ID:1SZM as an example, PSH-ML needs 1370.62 s to generate features using a single-core CPU, while XLPFE only needs 0.91 s to generate features. In terms of Mg metalloproteins, our count shows that a considerable proportion of them contains more than one Mg cofactor and if the effects of multinuclear effects are accounted for in the future, the predictive performance of XLPFE could be further improved. Overall, XLPFE can achieve robust performance in the scoring and ranking of binding affinity for metalloproteins and is among one of the best SFs for metalloproteins.

4. DISCUSSION AND CONCLUSIONS

In this work, we have combined various feature sets and ML methods to improve the scoring performance of SFs. The five feature sets include energy terms from X-score (X) and AutoDock Vina (V), the properties of the ligand (L), and the statistic sequence-related information from either the binding site (P) or the full protein (F), while the ML methods consist of ET, RF, XGBoost, SVR, and NN. Among the various combinations of feature sets and ML methods, we found that the combination of XLPF feature sets and the ET model (labeled as the XLPFE model) achieves the best and most stable performance. On the benchmark set CASF-2016, XLPFE shows outstanding scoring power (0.816) and ranking power (0.66), which is one of the best SFs in the CASF-2016 data set. XLPFE also shows consistently better scoring power and ranking power than other SFs for various expanded test sets beyond the CASF, including traditional SFs X-score and AutoDock Vina, ML-based SF ΔvinaXGB, and DL-based SF CNN-score. In particular, XLPFE achieved the best overall performance for metalloproteins. All these findings suggest that an appropriate selection of feature sets can render much improved transferability in ML-based SFs. Unlike some state-of-the-art SFs based on the complex topological features or structural features, the combination of simple energy terms and other properties related to protein−ligand binding may achieve better scoring performance and transferability.[42,47] Thus, this study gives some information that can lead to the further improvement of ML-based SFs.

DATA AND SOFTWARE AVAILABILITY

All protein−ligand structural data and related experimental binding affinity are from PDBbind (http://www.pdbbind-cn.org/). Our training and test sets and the code of XLPFE are open access (https://github.com/LinaDongXMU/XLPFE). Other data and results are all listed in the Supporting Information.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c01723.

Introduction of relevant SFs, ML methods, and corresponding hyperparameters, details of the performance of different features and methods on the test sets, and performance of various SFs on the test set CASF-2016-associated data (PDF)

AUTHOR INFORMATION

Corresponding Author

Binju Wang − *State Key Laboratory of Physical Chemistry of Solid Surfaces and Fujian Provincial Key Laboratory of Theoretical and Computational Chemistry, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 360015, P. R. China;* orcid.org/0000-0002-3353-9411; Email: wangbinju2018@xmu.edu.cn

Authors

Lina Dong − *State Key Laboratory of Physical Chemistry of Solid Surfaces and Fujian Provincial Key Laboratory of Theoretical and Computational Chemistry, iChEM, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 360015, P. R. China*

Xiaoyang Qu − *State Key Laboratory of Physical Chemistry of Solid Surfaces and Fujian Provincial Key Laboratory of Theoretical and Computational Chemistry, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 360015, P. R. China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c01723

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (2019YFA0906400) and NSFC (no. 22073077, 21933009, and 21907082). We thank Dr. Jianing Lu, Dr. Junjie Hu, and Xiaoxu Guo for valuable discussions.

ABBREVIATIONS

CADD, computer-aided drug design; CNN, convolutional neural network; DL, deep learning; ET, extra trees; HB, hydrogen bond interaction; HM, hydrophobic matching; HP, hydrophobic pairs; HS, hydrophobic surface; LR, linear regressions; MAE, mean absolute error; ML, machine learning; NN, neural network; RF, random forests; RMSE, root mean square error; R_p, Pearson correlation coefficient; R_s, Spearman correlation coefficient; RT, rotatable bonds; SF, scoring function; SVR, support vector regressions; VDW, van der Waals interaction; XGBoost, extreme gradient boosting

https://doi.org/10.1021/acsomega.2c01723
ACS Omega 2022, 7, 21727−21735

REFERENCES

(1) Gorgulla, C.; Boeszoermenyi, A.; Wang, Z.-F.; Fischer, P. D.; Coote, P. W.; Padmanabha Das, K. M.; Malets, Y. S.; Radchenko, D. S.; Moroz, Y. S.; Scott, D. A.; Fackeldey, K.; Hoffmann, M.; Iavniuk, I.; Wagner, G.; Arthanari, H. An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **2020**, *580*, 663−668.

(2) Wang, D. D.; Zhu, M.; Yan, H. Computationally predicting binding affinity in protein-ligand complexes: free energy-based simulations and machine learning-based scoring functions. *Briefings Bioinf.* **2021**, *22*, bbaa107.

(3) Senn, H. M.; Thiel, W. QM/MM methods for biomolecular systems. *Angew. Chem., Int. Ed. Engl.* **2009**, *48*, 1198−1229.

(4) Brunk, E.; Rothlisberger, U. Mixed Quantum Mechanical/Molecular Mechanical Molecular Dynamics Simulations of Biological Systems in Ground and Electronically Excited States. *Chem. Rev.* **2015**, *115*, 6217−6263.

(5) Chen, J.; Wang, X.; Pang, L.; Zhang, J. Z. H.; Zhu, T. Effect of mutations on binding of ligands to guanine riboswitch probed by free energy perturbation and molecular dynamics simulations. *Nucleic Acids Res.* **2019**, *47*, 6618−6631.

(6) Scheen, J.; Wu, W.; Mey, A. S. J. S.; Tosco, P.; Mackey, M.; Michel, J. Hybrid Alchemical Free Energy/Machine-Learning Methodology for the Computation of Hydration Free Energies. *J. Chem. Inf. Model.* **2020**, *60*, 5331−5339.

(7) Steinbrecher, T.; Mobley, D. L.; Case, D. A. Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations. *J. Chem. Phys.* **2007**, *127*, 214108.

(8) Bhati, A. P.; Wan, S.; Wright, D. W.; Coveney, P. V. Rapid, Accurate, Precise, and Reliable Relative Free Energy Prediction Using Ensemble Based Thermodynamic Integration. *J. Chem. Theory Comput.* **2017**, *13*, 210−222.

(9) Wang, Z.; Sun, H.; Yao, X.; Li, D.; Xu, L.; Li, Y.; Tian, S.; Hou, T. Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. *Phys. Chem. Chem. Phys.* **2016**, *18*, 12964−12975.

(10) Wang, E.; Sun, H.; Wang, J.; Wang, Z.; Liu, H.; Zhang, J. Z. H.; Hou, T. End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chem. Rev.* **2019**, *119*, 9478−9508.

(11) Wang, E.; Fu, W.; Jiang, D.; Sun, H.; Wang, J.; Zhang, X.; Weng, G.; Liu, H.; Tao, P.; Hou, T. VAD-MM/GBSA: A Variable Atomic Dielectric MM/GBSA Model for Improved Accuracy in Protein-Ligand Binding Free Energy Calculations. *J. Chem. Inf. Model.* **2021**, *61*, 2844−2856.

(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) 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.

(14) Yang, C.; Zhang, Y. Lin_F9: A Linear Empirical Scoring Function for Protein-Ligand Docking. *J. Chem. Inf. Model.* **2021**, *61*, 4630−4644.

(15) Bao, J.; He, X.; Zhang, J. Z. H. Development of a New Scoring Function for Virtual Screening: APBScore. *J. Chem. Inf. Model.* **2020**, *60*, 6355−6365.

(16) 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.

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

(18) Yang, X.; Wang, Y.; Byrne, R.; Schneider, G.; Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem. Rev.* **2019**, *119*, 10520−10594.

(19) Lim, S.; Lu, Y.; Cho, C. Y.; Sung, I.; Kim, J.; Kim, Y.; Park, S.; Kim, S. A review on compound-protein interaction prediction methods: Data, format, representation and model. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 1541−1556.

(20) Ghislat, G.; Rahman, T.; Ballester, P. J. Recent progress on the prospective application of machine learning to structure-based virtual screening. *Curr. Opin. Chem. Biol.* **2021**, *65*, 28−34.

(21) Di Filippo, J. I.; Cavasotto, C. N. Guided structure-based ligand identification and design via artificial intelligence modeling. *Expert Opin. Drug Discovery* **2022**, *17*, 71−78.

(22) Xiong, G. L.; Shen, C.; Yang, Z. Y.; Jiang, D. J.; Liu, S.; Lu, A. P.; Chen, X.; Hou, T. J.; Cao, D. S. Featurization strategies for protein-ligand interactions and their applications in scoring function development. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021**, *12*, No. e1567.

(23) Shen, C.; Ding, J.; Wang, Z.; Cao, D.; Ding, X.; Hou, T. From machine learning to deep learning: Advances in scoring functions for protein−ligand docking. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2019**, *10*, No. e1429.

(24) 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.

(25) 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.

(26) 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.

(27) 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.

(28) Wang, C.; Zhang, Y. Improving scoring-docking-screening powers of protein-ligand scoring functions using random forest. *J. Comput. Chem.* **2017**, *38*, 169−177.

(29) Liu, X.; Feng, H.; Wu, J.; Xia, K. Persistent spectral hypergraph based machine learning (PSH-ML) for protein-ligand binding affinity prediction. *Briefings Bioinf.* **2021**, *22*, bbab127.

(30) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942−957.

(31) Cang, Z.; Wei, G.-W. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Comput. Biol.* **2017**, *13*, No. e1005690.

(32) Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; De Fabritiis, G. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* **2018**, *58*, 287−296.

(33) Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P. Development and evaluation of a deep learning model for protein-ligand binding affinity prediction. *Bioinformatics* **2018**, *34*, 3666−3674.

(34) Jiang, D.; Hsieh, C.-Y.; Wu, Z.; Kang, Y.; Wang, J.; Wang, E.; Liao, B.; Shen, C.; Xu, L.; Wu, J.; Cao, D.; Hou, T. InteractionGraphNet: A Novel and Efficient Deep Graph Representation Learning Framework for Accurate Protein-Ligand Interaction Predictions. *J. Med. Chem.* **2021**, *64*, 18209−18232.

(35) Bao, J.; He, X.; Zhang, J. Z. H. DeepBSP-a Machine Learning Method for Accurate Prediction of Protein-Ligand Docking Structures. *J. Chem. Inf. Model.* **2021**, *61*, 2231−2240.

(36) Zheng, L.; Fan, J.; Mu, Y. OnionNet: a Multiple-Layer Intermolecular-Contact-Based Convolutional Neural Network for Protein-Ligand Binding Affinity Prediction. *ACS Omega* **2019**, *4*, 15956−15965.

(37) Jones, D.; Kim, H.; Zhang, X.; Zemla, A.; Stevenson, G.; Bennett, W. F. D.; Kirshner, D.; Wong, S. E.; Lightstone, F. C.; Allen, J. E. Improved Protein-Ligand Binding Affinity Prediction with Structure-Based Deep Fusion Inference. *J. Chem. Inf. Model.* **2021**, *61*, 1583−1592.

(38) Kwon, Y.; Shin, W. H.; Ko, J.; Lee, J. AK-Score Accurate Protein-Ligand Binding Affinity Prediction Using an Ensemble of 3D-Convolutional Neural Networks. *Int. J. Mol. Sci.* **2020**, *21*, 8424−8440.

(39) Karlov, D. S.; Sosnin, S.; Fedorov, M. V.; Popov, P. graphDelta: MPNN Scoring Function for the Affinity Prediction of Protein-Ligand Complexes. *ACS Omega* **2020**, *5*, 5150−5159.

https://doi.org/10.1021/acsomega.2c01723
ACS Omega 2022, 7, 21727−21735

(40) Meli, R.; Anighoro, A.; Bodkin, M. J.; Morris, G. M.; Biggin, P. C. Learning protein-ligand binding affinity with atomic environment vectors. *J. Cheminf.* **2021**, *13*, 59.
(41) Shen, C.; Hu, Y.; Wang, Z.; Zhang, X.; Zhong, H.; Wang, G.; Yao, X.; Xu, L.; Cao, D.; Hou, T. Can machine learning consistently improve the scoring power of classical scoring functions? Insights into the role of machine learning in scoring functions. *Briefings Bioinf.* **2021**, *22*, 497−514.
(42) 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.
(43) Nguyen, D. D.; Wei, G.-W. AGL-Score: Algebraic Graph Learning Score for Protein-Ligand Binding Scoring, Ranking, Docking, and Screening. *J. Chem. Inf. Model.* **2019**, *59*, 3291−3304.
(44) Zhang, X.; Shen, C.; Guo, X.; Wang, Z.; Weng, G.; Ye, Q.; Wang, G.; He, Q.; Yang, B.; Cao, D.; Hou, T. ASFP (Artificial Intelligence based Scoring Function Platform): a web server for the development of customized scoring functions. *J. Cheminf.* **2021**, *13*, 6.
(45) Meng, Z.; Xia, K. Persistent spectral-based machine learning (PerSpect ML) for protein-ligand binding affinity prediction. *Sci. Adv.* **2021**, *7*, No. eabc5329.
(46) Wang, D. D.; Ou-Yang, L.; Xie, H.; Zhu, M.; Yan, H. Predicting the impacts of mutations on protein-ligand binding affinity based on molecular dynamics simulations and machine learning methods. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 439−454.
(47) Dong, L.; Qu, X.; Zhao, Y.; Wang, B. Prediction of Binding Free Energy of Protein-Ligand Complexes with a Hybrid Molecular Mechanics/Generalized Born Surface Area and Machine Learning Method. *ACS Omega* **2021**, *6*, 32938−32947.
(48) Jiang, P.; Chi, Y.; Li, X. S.; Liu, X.; Hua, X. S.; Xia, K. Molecular persistent spectral image (Mol-PSI) representation for machine learning models in drug design. *Briefings Bioinf.* **2022**, *23*, bbab527.
(49) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.* **2019**, *59*, 895−913.
(50) Su, M.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Tapping on the Black Box: How Is the Scoring Power of a Machine-Learning Scoring Function Dependent on the Training Set? *J. Chem. Inf. Model.* **2020**, *60*, 1122−1136.
(51) Zhang, J.; Yang, W.; Piquemal, J.-P.; Ren, P. Modeling Structural Coordination and Ligand Binding in Zinc Proteins with a Polarizable Potential. *J. Chem. Theory Comput.* **2012**, *8*, 1314−1324.
(52) Soni, A.; Bhat, R.; Jayaram, B. Improving the binding affinity estimations of protein-ligand complexes using machine-learning facilitated force field method. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 817−830.
(53) Lu, J.; Hou, X.; Wang, C.; Zhang, Y. Incorporating Explicit Water Molecules and Ligand Conformation Stability in Machine-Learning Scoring Functions. *J. Chem. Inf. Model.* **2019**, *59*, 4540−4549.
(54) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.
(55) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind database: methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.
(56) 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.
(57) Sheridan, R. P. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J. Chem. Inf. Model.* **2013**, *53*, 783−790.
(58) 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.
(59) 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.
(60) Böhm, 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.
(61) Pedregosa, F. e. a. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(62) Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3−42.
(63) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.
(64) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Krishnapuram, B., Shah, A., Aggarwal, C., Shen, D., Rastogi, R., Eds.; ACM: San Francisco, CA, 2016; pp 785−794.
(65) Brereton, R. G.; Lloyd, G. R. Support vector machines for classification and regression. *Analyst* **2010**, *135*, 230−267.
(66) Sainath, T. N.; Vinyals, O.; Senior, A.; Sak, H. Ieee, Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks. *2015 Ieee International Conference on Acoustics, Speech, and Signal Processing*; Ieee: New York, 2015; pp 4580−4584.
(67) Yoo, W.; Mayberry, R.; Bae, S.; Singh, K.; Peter He, Q.; Lillard, J. W., Jr. A Study of Effects of MultiCollinearity in the Multivariable Analysis. *Int. J. Appl. Sci. Technol.* **2014**, *4*, 9−19.
(68) Born, J.; Huynh, T.; Stroobants, A.; Cornell, W. D.; Manica, M. Active Site Sequence Representations of Human Kinases Outperform Full Sequence Representations for Affinity Prediction and Inhibitor Generation: 3D Effects in a 1D Model. *J. Chem. Inf. Model.* **2022**, *62*, 240−257.
(69) Wang, K.; Zhou, R.; Li, Y.; Li, M. DeepDTAF: a deep learning method to predict protein-ligand binding affinity. *Briefings Bioinf.* **2021**, *22*, bbab072.
(70) McNutt, A. T.; Francoeur, P.; Aggarwal, R.; Masuda, T.; Meli, R.; Ragoza, M.; Sunseri, J.; Koes, D. R. GNINA 1.0: molecular docking with deep learning. *J. Cheminf.* **2021**, *13*, 43.
(71) Shang, E.; Yuan, Y.; Chen, X.; Liu, Y.; Pei, J.; Lai, L. De novo design of multitarget ligands with an iterative fragment-growing strategy. *J. Chem. Inf. Model.* **2014**, *54*, 1235−1241.

https://doi.org/10.1021/acsomega.2c01723
ACS Omega 2022, 7, 21727−21735