ACS OMEGA

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

Article

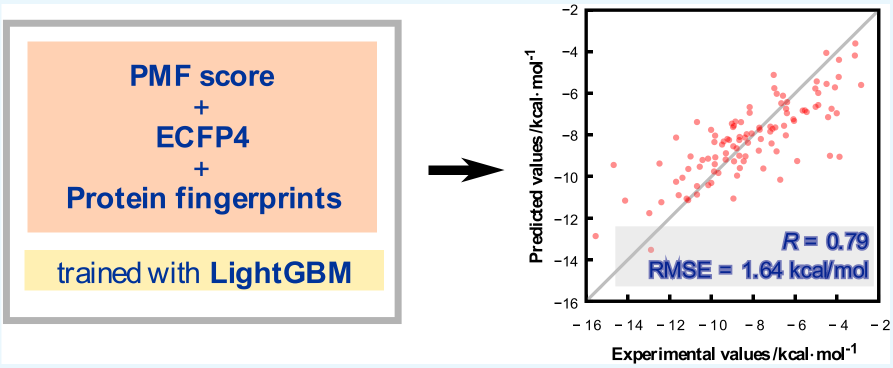
Machine-Learning- and Knowledge-Based Scoring Functions Incorporating Ligand and Protein Fingerprints

Kazuhiro J. Fujimoto,* Shota Minami, and Takeshi Yanai*

Cite This: *ACS Omega* 2022, 7, 19030−19039

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: We propose a novel machine-learning-based scoring function for drug discovery that incorporates ligand and protein structural information into a knowledge-based PMF score. Molecular docking, a simulation method for structure-based drug design (SBDD), is expected to reduce the enormous costs associated with conventional experimental methods in terms of rational drug discovery. Molecular docking has two main purposes: to predict ligand-binding structures for target proteins and to predict protein−ligand binding affinity. Currently available programs of molecular docking offer an accurate prediction of ligand binding structures for many systems. However, the accurate prediction of binding affinity remains challenging. In this study, we developed a new scoring function that incorporates fingerprints representing ligand and protein structures as descriptors in the PMF score. Here, regression analysis of the scoring function was performed using the following machine learning techniques: least absolute shrinkage and selection operator (LASSO) and light gradient boosting machine (LightGBM). The results on a test data set showed that the binding affinity delivered by the newly developed scoring function has a Pearson correlation coefficient of 0.79 with the experimental value, which surpasses that of the conventional scoring functions. Further analysis provided a chemical understanding of the descriptors that contributed significantly to the improvement in prediction accuracy. Our approach and findings are useful for rational drug discovery.

1. INTRODUCTION

The development of a new drug requires a long period of time and a large amount of money before approval and marketing.[1] To reduce these costs, there is a strong demand for rational drug discovery.[2] In the early stages of drug discovery, drug candidates are experimentally searched for from a vast array of compounds in terms of their binding affinity to the target protein (lead identification), and the compounds are repeatedly improved to further increase their binding affinity (lead optimization).[3] These processes are based on the idea of hit or miss and therefore bring about the enormous costs of drug discovery.

Computer-aided drug discovery (in silico drug discovery) has been attracting significant attention as a solution to such problems in conventional drug discovery.[4−6] Structure-based drug design (SBDD),[7,8] which uses the three-dimensional structure of target proteins, plays a central role in computer-aided drug discovery. If the protein pocket is regarded as a keyhole, SBDD is a method to design a matching key based on the shape of the keyhole. Therefore, computer-aided drug discovery by SBDD is expected to be more rational than the conventional approach in designing drug candidates with high activity and selectivity. The anti-influenza drug oseltamivir[9] and the chronic myelogenous leukemia drug imatinib[10] are representative examples of successful SBDD.

The background for the realization of SBDD is the accumulation of a large amount of data on three-dimensional protein structures due to the development of structural analysis

Received: May 6, 2022
Accepted: May 12, 2022
Published: May 25, 2022

https://doi.org/10.1021/acsomega.2c02822

infrastructure and the dramatic improvement in computational power. In addition, the development of a computational science method called molecular docking constitutes another major factor. It is a type of computer simulation and has become an indispensable tool for SBDD.[5,7]

The purpose of molecular docking is twofold: first, to predict the binding structure of the ligand in the binding pocket of the target protein and second, to predict the protein−ligand binding affinity. These accurate predictions are key to rational drug discovery. To this end, a number of molecular docking programs have been developed, ranging from commercial to academic use.[11−20] Simulations using these programs reproduce actual ligand-binding structures well in many systems. Thus, one of the goals of molecular docking, the prediction of ligand-binding structures, is now possible with considerable accuracy. In contrast, the other goal of molecular docking, the prediction of binding affinity (binding energy), has not yet been achieved to a satisfactory degree, and there is still a large discrepancy between the predicted and experimental values. Due to the inability to predict absolute binding affinities, the relative binding affinities of various ligands to a given target protein are often evaluated in drug discovery, but the effectiveness of this approach is highly system-dependent. For this reason, the accurate prediction of binding affinity is one of the major challenges in computer-aided drug discovery.

The poor prediction of binding affinities by current molecular docking is mainly attributed to inaccuracies of the scoring function. It has been reported that incorporating solvation and entropy effects into the scoring function improves the prediction accuracy,[21−26] but these incorporations are usually computationally expensive. On the other hand, one of the knowledge-based scoring functions, the PMF score,[27−30] can simply take into account these contributions. The PMF score expresses the interatomic potential in a protein−ligand complex as a frequency of occurrence of two atoms at a certain distance in the crystal structure data and hence gives large negative values at binding distances that often appear in the crystal structures. Thus, the PMF score has the advantage of implicitly accounting for many kinds of contributions to the binding, such as solvation, entropy, and enthalpy effects, at low computational cost.

This study aims to develop a new scoring function based on the PMF score that provides an accurate prediction of binding affinity. In recent years, the application of machine-learning-based regression algorithms has been proposed for the development of new scoring functions,[31−33] where random forests and support vector machines have been used successfully.[34−36] Some of the scoring functions developed have shown significant improvements over conventional scoring functions,[37−41] with Pearson correlation coefficients between predicted and experimental values exceeding 0.7 to 0.8. Based on these successes, machine learning is employed in this study to develop scoring functions. It should be noted, however, that we employ a different strategy from the one reported previously in the sense that the model function is constructed by adding fingerprints representing the ligand and protein structure into the PMF score. Incorporating ligand fingerprints has already been reported,[42] but there are no reports of scoring functions that take protein fingerprints into account. Regression analysis is then performed using machine learning algorithms such as LASSO[43] and LightGBM.[44] The scoring function developed in this study exhibits a Pearson correlation coefficient of 0.79, indicating a higher reproducibility to infer experimental binding affinities than the conventional methods. Further analysis reveals that ligand and protein fingerprints compensate for the poor description of the PMF score. Our approach and findings would be useful for the further development of computer-aided drug discovery.

2. METHODS

In this section, we briefly explain how to develop a new machine-learning-based scoring function in which ligand and protein structural information is incorporated into the knowledge-based PMF score.[27−30] First, descriptors constituting model functions are introduced, and then the model is optimized by machine learning.

2.1. PMF Score. The PMF score[27−30] $V_{ij}(r)$ represents the knowledge-based protein−ligand binding energy and is defined as the sum of the interaction energies $A_{ij}(r)$ of the protein−ligand atom pairs at distance r.

$$V_{ij}(r) = \sum_{\substack{i,j \\ r<r_{ij}^{c}}} A_{ij}(r) \tag{1}$$

Here, the indices i and j are protein and ligand atoms, respectively, and $r_{\mathrm{ij}}^{\mathrm{c}}$ is the cutoff distance defined for each pair of protein and ligand atoms. There are 17 types of protein atoms and 34 types of ligand atoms.[30] The interactions beyond $r_{\mathrm{ij}}^{\mathrm{c}}$ are not considered in the PMF score. The interatomic energy $A_{ij}(r)$ is calculated using the Boltzmann distribution as follows:

$$A_{ij}(r) = -k_{\mathrm{B}}T\ln\left[f_{j}(r)\frac{\rho_{ij}(r)}{\rho_{ij}^{0}}\right], \tag{2}$$

where k_{B} is Boltzmann's constant, T is temperature, $\rho_{ij}(r)$ is the number density of ligand−protein pair ij at distance r, and ρ_{ij}^{0} is the number density in the reference state where the atomic interaction of pair ij is zero. Therefore, $\frac{\rho_{ij}(r)}{\rho_{ij}^{0}}$ corresponds to the probability density distribution of pair ij in the structural data set of the protein−ligand complex. $f_j(r)$ is called the ligand volume correction factor and is used to remove the volume of ligand atoms from the number density.[29] In this study, the PMF score is used as the basis for developing the scoring function.

2.2. Ligand Fingerprints. To incorporate the structural information of the ligand in the PMF score, we employ the molecular fingerprint method. Molecular fingerprints are the representation of molecular structures based on certain rules using binary vectors representing presence or absence as 1 or 0, or count vectors representing the frequency of occurrence. It is one of the most common methods to convert a molecular structure into a computer-readable form and is widely used in chemoinformatics[45−47] and in silico drug discovery.[48] Hereafter, molecular fingerprints are called ligand fingerprints to distinguish them from fingerprints on proteins, which will be introduced later.

There are two major types of ligand fingerprints. The first one assigns a substructure in the molecule to each bit of the array. Here, substructures refer to fragments such as functional groups or carbon skeletons. This method can appropriately represent molecular structures composed of the substructures registered as keys in a bit array. However, if the substructures are not registered, an accurate representation of the molecular structure is not possible. The second fingerprint method constructs molecular structures using the information on the relative positions of atoms in the molecule. This method can generate ligand fingerprints that reflect any chemical structure. However, it is difficult to extract only the substructure information from the generated fingerprint. This is a drawback of this method in the analysis of machine-learning-based results. In this study, we have employed MACCSKeys[49] and extended connectivity fingerprint (ECFP)[50,51] as representative examples of the first and second fingerprint methods, respectively.

The MACCSKeys fingerprint is a bit array consisting of 166 keys representing different substructures. The concept of MACCSKeys is illustrated in Figure 1a. If a substructure in the molecule is among the types associated with the 166 keys, the corresponding bit is set to 1; otherwise, it is set to 0. It should be noted that even if a molecule contains several substructures of the same type, the number of substructures is not reflected in MACCSKeys.

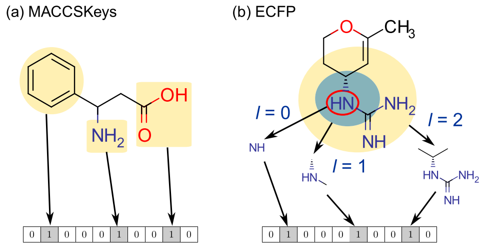


Figure 1. Representation of molecular structures using (a) MACCSKeys and (b) ECFP. The MACCSKeys bit is set to 1 if the molecule contains a registered substructure; otherwise, it is set to 0. ECFP searches for the relative position of each atom in the molecule using a bond layer *l* and records the structural information in a bit array.

The ECFP fingerprint is one of the circular fingerprints,[52] which represents molecular structures by considering the position of atoms within a bond layer *l* centered on each atom of the molecule. The concept of ECFP is illustrated in Figure 1b. First, the relative positions of atoms within a given bond layer are iteratively explored, and then the resulting atom environments are transformed to a bit array through a hash function. The size of the bond layer and bit array can be determined arbitrarily, but typically, the maximum bond layer of 2 (ECFP4) and 1024 bits are used, respectively. ECFP has the advantage of not requiring a predefined substructure data set, in contrast to MACCSKeys. In addition, ECFP can consider several functional groups of the same type separately, which makes it a more accurate representation than MACCSKeys.

2.3. Protein Fingerprints.

In addition to the structural information of the ligand, we have attempted to include the structural information of the protein in the scoring function using our original fingerprint method shown below. This method is named protein fingerprint (PF), and its concept is shown in Figure 2. We assign three types of moieties to each amino acid constituting the given protein: N-linked main chain, O-linked main chain, and side chain. To account for direct protein−ligand interactions, we only incorporate atoms constituting amino acids located within 5 Å of the ligand atoms in the binding pocket. Hydrogen atoms are not considered here. Our descriptors distinguish 20 types of amino acids including the three types of the associated moieties except for glycine, resulting in a total of 59 different descriptors, which were included in the scoring function. The PF method employs count vectors to record structural information, whereas MACCSKeys and ECFP use binary vectors. The following procedure is used to represent PF. First, all the atoms in the protein that are within 5 Å of the ligand are searched for, then the corresponding atoms are classified into the 59 different amino acid sites, and the number of their occurrences is stored in the corresponding array of descriptors. It should be noted that the PF method takes into account the numbers of amino acids of the same type, which are stored in the count vectors.

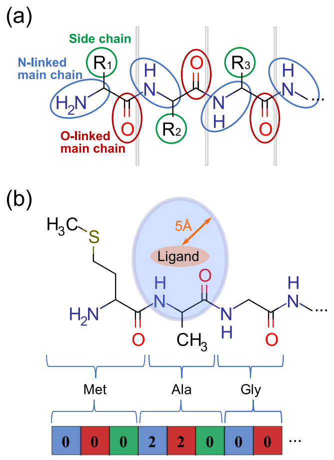


Figure 2. Protein fingerprints (PF) developed in this study. (a) Each amino acid is described by three descriptors: N-linked main chain, O-linked main chain, and side chain. (b) Example to illustrate PF. Here we consider a situation where only the main chain of alanine is within 5 Å of the ligand. The counter associated with the N-linked main chain descriptor of alanine (blue box) is incremented by 2 due to the presence of a nitrogen atom and an alpha carbon atom in the alanine main chain. Similarly, the counter associated with the O-linked main chain descriptor of alanine (red box) is incremented by 2 due to the carbon and oxygen atoms forming the peptide bond. It should be noted that the number of hydrogen atoms is not considered in PF.

2.4. Machine Learning with LASSO and LightGBM.

To develop a new scoring function, the PMF score modified with ligand and protein fingerprints is trained with the machine learning method. In this study, we employ two machine learning techniques: the least absolute shrinkage and selection operator (LASSO)[43] and the light gradient boosting machine (LightGBM).[44] Here, we attempt to construct the following relation for the objective variable y and m descriptors (explanatory variables) x_1, x_2, ...x_m:

$$y = f(x_1, x_2, \ldots x_m), \tag{3}$$

where $f(x_1, x_2, \ldots x_m)$ is the objective function. It should be noted that we assume the protein−ligand binding energy for the objective variable and the components of PMF score, ligand fingerprints, and protein fingerprints for the descriptors.

LASSO is a regression analysis technique that imposes regularization on the least squares process.[43] First, we consider a linear model with the following form:

$$f(x_1, x_2, \ldots x_m) = c_0 + c_1x_1 + c_2x_2 + \ldots + c_mx_m, \tag{4}$$

where c_l (for l = 0 to m) is the partial regression coefficient. The values of these coefficients are determined by the least

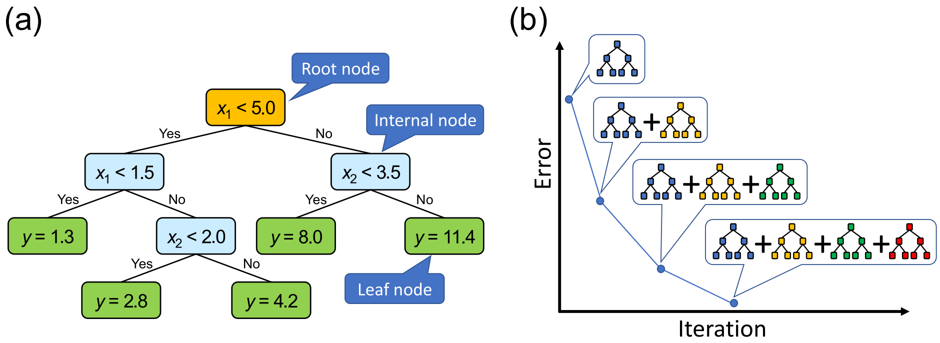


Figure 3. (a) Example of a decision tree for a regression problem. The characteristics of the descriptors of the data are diagnosed by the conditional expressions at the root and internal nodes, and the predicted value of the objective variable is returned at the leaf node. (b) Schematic illustration of LightGBM. For each iteration, a new decision tree is added and the error decreases accordingly.

squares method. To this end, the following loss function L is considered:

$$L = \sum_{k}^{n} (y^k - f(x_1^k, x_2^k, ... x_m^k))^2, \tag{5}$$

where y^k and x_l^k (for $l = 1$ to m) denote the objective variable and descriptors for training data k, respectively, and the summation runs over all n data. If the absolute values of the partial regression coefficients become large, the model fits too well to the training data and poorly to the test data. This phenomenon is called overfitting (overtraining). To circumvent this problem, LASSO adds a regularization (penalty) term to the loss function, which is the sum of the absolute values of the partial regression coefficients multiplied by a hyperparameter α.

$$L = \sum_{k}^{n} (y^k - f(x_1^k, x_2^k, ... x_m^k))^2 + \alpha \sum_{l}^{m} |c_l^k| \tag{6}$$

In this form, the loss function L increases due to the regularization term as the partial regression coefficient increases. In this way, LASSO prevents an increase in the absolute value of the partial regression coefficient in the least squares treatment, making overfitting less likely to occur. The suppression of overfitting strongly depends on the value of the hyperparameter α. Therefore, this study adopts the value of α that gives the highest prediction accuracy for the pretest data set.

LightGBM is a supervised machine learning method based on the gradient-boosting decision tree (GBDT) algorithm,[44] which has recently been applied in various fields such as life sciences,[53−55] engineering,[56,57] and economics.[58] An attempt to apply XGBoost, a type of GBDT, to protein−ligand binding affinity has also been reported.[59] Since LightGBM is a combination of the decision tree and gradient boosting methods, each is described here. First, the decision tree[60] is a tree-based data analysis method that can be applied to both classification and regression problems. In this study, the regression model of eq 3 is considered. As illustrated in Figure 3a, the properties of the data were evaluated at the root and internal nodes using conditional expressions on the descriptors, and finally, the value of the objective variable y is predicted based on the result at the leaf node. Another feature of LightGBM is gradient boosting,[61,62] which is a technique that attempts to iteratively improve the prediction accuracy of a model, $f(x_1, x_2, ...x_m)$, by adding a new estimator, $\delta(x_1, x_2, ...x_m)$.

$$f_\mu(x_1, x_2, ... x_m) = f_{\mu-1}(x_1, x_2, ... x_m) + \beta\delta_\mu(x_1, x_2, ... x_m), \tag{7}$$

where μ represents the iteration step and β is a hyperparameter called the learning rate. In LightGBM, $\delta_\mu(x_1, x_2, ...x_m)$ corresponds to a decision tree that represents the error between the actual (experimental) value y and the predicted value by the $\mu - 1$ step decision tree $f_{\mu-1}(x_1, x_2, ...x_m)$. A schematic illustration of LightGBM is shown in Figure 3b. Repetition of the operation according to eq 7 can gradually reduce the error. However, too much iteration leads to overfitting. To avoid this problem, we employ a method called early stopping, in which the number of iterations M (the number of trees) with the highest prediction accuracy is determined based on the results on the pretest data set conducted with different iteration counts.

2.5. Computational Details.

PDBbind[63] v2019 was used as the data set for developing new scoring functions. PDBbind is known as a high-quality protein−ligand structure data set that can be used to develop and validate scoring functions. In this study, we used 6271 data structures from PDBbind v2019 in which dissociation constant K_d values exist. Of the 6271 protein−ligand complex structures, 4933 were used as the training data set, 1234 as the pretest data set, and 104 as the test data set. The training data set was used in machine learning with LASSO and LightGBM, while the pretest data set was used to determine the value of the hyperparameter α for LASSO and the number of iterations M for LightGBM. The test data set corresponds to the core set in PDBbind (CASF-2016[64]), which is widely used as the benchmark data for evaluating the accuracy of scoring functions.[65−67] This study also used it only as an external test data set, not as a training data set, to evaluate the performance of the scoring function. All atomic coordinates of the protein and ligand, including hydrogens, were used without any optimization for the PDBbind structures.

In this study, we aimed to improve the prediction accuracy of the PMF score by incorporating the structural information of ligands and proteins using the fingerprint methods. In addition, we examined the effect of incorporating van der Waals (vdW) interactions into the PMF score using the 6−12 Lennard−Jones potentials[68] with the amber99 force field.[69] The pairwise interatomic potentials, PMF scores and vdW interactions, were calculated using the atomic coordinates of the PDBbind structures and included in the scoring function as descriptors. It should be noted that the PMF score in this study does not consider the interatomic potentials associated with

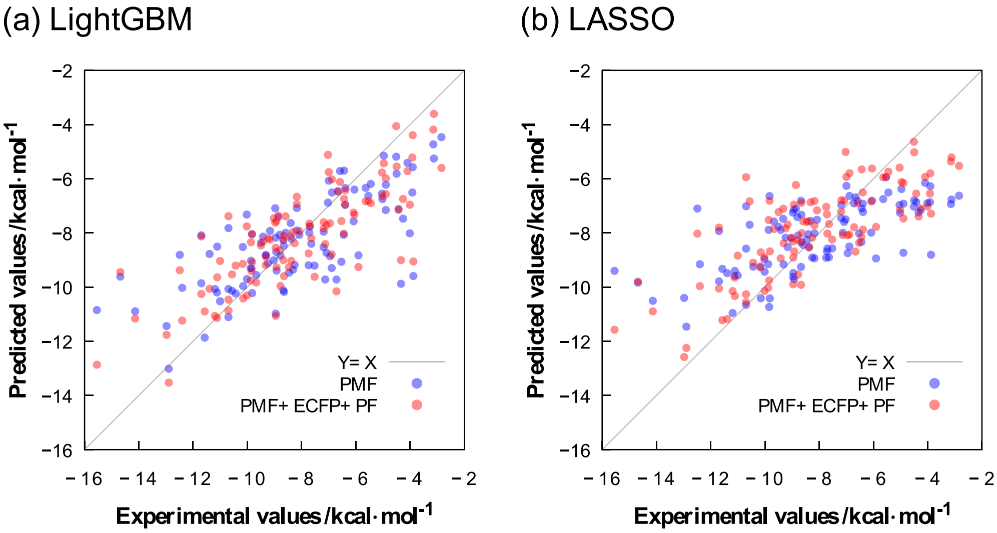


Figure 4. Scatter plots of experimental values versus predicted values by scoring functions trained with (a) LightGBM and (b) LASSO.

hydrogen, whereas the vdW interaction does. MACCSKeys and ECFP were applied to the ligand structure, and each bit comprising the bit array was used for a single descriptor. Protein fingerprints were applied to protein atoms within 5 Å of the ligand atoms, which were used as descriptors. Machine learning using LASSO or LightGBM was performed on the objective functions consisting of the PMF score, MACCSKeys, ECFP, PF, and vdW. The hyperparameter α for the LASSO regression was determined through machine learning on the pretest data set with α varying from 0.001 to 0.01 in 0.001 increments. The number of iterations M for LightGBM early-stopping was determined based on the results of machine learning on the pretest data set with M varying up to 10,000. It should be noted that the optimal values for the hyperparameter α and the number of iterations M are different for each model. The learning rate β of 0.1 was used for LightGBM.

The number of descriptors included in the scoring function is 448 for PMF scores, 166 for MACCSKeys, 1024 for ECFP, 59 for PF, and 493 for vdW. We validated a combination of these descriptors to improve the prediction accuracy of the scoring function.

The RDKit program package[70] is used for generating ligand fingerprints with MACCSKeys and ECFP. For machine learning of the models, the scikit-learn[71] and LightGBM[72] libraries were used for LASSO and LightGBM, respectively.

Two methods were employed in this study to assess the validity of the model. The first is the Pearson correlation coefficient R, given by

$$R = \frac{\sum_k^n (y_k - \bar{y})(f_k - \bar{f})}{\sqrt{\sum_k^n (y_k - \bar{y})^2}\sqrt{\sum_k^n (f_k - \bar{f})^2}}, \tag{8}$$

where y_k and f_k are the experimental and predicted binding affinities for the kth complex, respectively, and $\bar{y}$ and $\bar{f}$ are their average values. The second evaluation method is the root mean squared error (RMSE), which is given by

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_k^n (f_k - y_k)^2}. \tag{9}$$

Both R and RMSE were calculated using the test data set containing the 104 complexes (n = 104).

The training results vary depending on how the 6271 complex structures are divided into 4933 training and 1234 pretest data sets. To avoid such a dependence, the following steps were taken. First, we created 10 different combinations of training and pretest data sets with different seed values involved in the generation of random numbers and then performed machine learning on these data sets using LightGBM. Finally, we employed the data set combination that produced the correlation coefficient closest to the average of the resulting 10 correlation coefficients.

3. RESULTS AND DISCUSSION

3.1. Comparison of Binding Affinities Calculated by New Scoring Functions with Experimental Values. We applied our scoring functions to a test data set of 104 structures, which is the core data set of PDBbind, and examined the Pearson correlation coefficients between the experimental and predicted binding energies. The results of the scatter plots are shown in Figure 4, and the correlation coefficient R and RMSE values are summarized in Table 1. For comparison, the same analysis was also performed using various conventional scoring functions. The predicted values of

Table 1. Pearson Correlation Coefficient (R) and RMSE between the Experimental Values and the Predicted Values by the Newly Developed Scoring Functions

model	method	R	RMSE (kcal/mol)
PMF	LASSO	0.67	2.04
	LightGBM	0.73	1.85
PMF + MACCSKeys	LASSO	0.76	1.84
	LightGBM	0.69	1.93
PMF + ECFP	LASSO	0.77	1.76
	LightGBM	0.75	1.76
PMF + PF	LASSO	0.69	2.00
	LightGBM	0.75	1.77
PMF + vdW	LASSO	0.67	2.02
	LightGBM	0.74	1.82
PMF + MACCSKeys + PF	LASSO	0.75	1.83
	LightGBM	0.73	1.83
PMF + ECFP + PF	LASSO	0.77	1.77
	LightGBM	0.79	1.64

binding affinity by each scoring function were taken from the CASF-2016[64] data set. These results are listed in Table 2. The results of the Pearson correlation coefficients show that including ligand and protein fingerprints in the scoring function improves the prediction accuracy of the binding energy step by step from the PMF model. The most accurate result was obtained from the PMF + ECFP + PF model trained with LightGBM, with a Pearson correlation coefficient of 0.79. Pearson correlation coefficients calculated with the conventional scoring functions range from 0.24 to 0.74, with the best result being obtained for deltaVinaRF20.[35] These results indicate that our newly developed scoring function has a higher prediction accuracy than the conventional ones. The RMSE of the PMF + ECFP + PF model is 1.64 kcal/mol, which also shows better prediction accuracy of binding energies than the results of AutodockVina (2.42 kcal/mol), ASE@MOE (9.63 kcal/mol), and deltaVinaRF20 (10.88 kcal/mol). These results clearly show that the newly developed scoring functions are applicable not only to the relative comparison of protein—ligand binding energies but also to the evaluation of their absolute values.

Table 2. Pearson Correlation Coefficient (R) between the Experimental Values and the Predicted Values by the Conventional Scoring Functions

model	R	model	R
ASE@MOE	0.54	LigScore1@DS	0.24
Affinity-dG@MOE	0.55	LigScore2@DS	0.41
Alpha-HB@MOE	0.55	London-dG@MOE	0.33
AutodockVina	0.54	PLP1@DS	0.41
ChemScore@SYBYL	0.45	PLP2@DS	0.43
D-Score@SYBYL	0.40	PMF04@DS	0.43
DrugScore2018	0.43	PMF@DS	0.50
DrugScoreCSD	0.49	PMF@SYBYL	0.37
G-Score@SYBYL	0.40	X-Score	0.55
GBVI_WSA-dG@MOE	0.24	X-ScoreHM	0.50
LUDI1@DS	0.41	deltaSAS	0.53
LUDI2@DS	0.45	deltaVinaRF20	0.74
LUDI3@DS	0.41		

Overall, the accuracy in terms of the correlation coefficient and RMSE was found to be improved by adding ligand fingerprints as descriptors, although there were some cases where the prediction accuracy worsened, such as the PMF + MACCSKeys models trained by LightGBM. In addition, the inclusion of protein fingerprints in the scoring function further improved the correlation coefficient and RMSE results. In the case of LightGBM, the change from the PMF to the PMF + ECFP + PF models improved the correlation coefficient by 0.06 and rectified the RMSE by 0.21 kcal/mol. These results clearly show the importance of adding ligand and protein structural information as descriptors to the scoring function.

The best scoring function in our results was obtained from training with LightGBM, but in some models, LASSO gave better results. This indicates that the best machine learning method depends on the type of descriptor included in the scoring function. Interestingly, in the present results, the PMF + ECFP + PF model provided the best values of the correlation coefficient and RMSE for LightGBM as well as LASSO. This strongly suggests that the ECFP and PF fingerprints are suitable for descriptors to be included in the scoring function.

In the present calculations, the ECFP fingerprints gave better results than the MACCSKeys fingerprints. MACCSKeys contains 166 bits, each bit corresponding to a specific substructure such as carbon skeletons and functional groups. Therefore, if a test compound consists of substructures registered in MACCSKeys, the molecular structure information will be reflected well in the scoring function; otherwise, it cannot be accurately represented. On the other hand, the ECFP fingerprint can represent any molecular structure by considering the relative positions of atoms within the bond layer l centered on each atom of the compound. For this reason, ECFP probably gave better results than MACCSKeys in the data set used here.

The worst score in the present results was obtained from the PMF + vdW model trained with LASSO, with a correlation coefficient of 0.67 and a RMSE of 2.02 kcal/mol. This result is almost the same as that of the PMF model trained by LASSO, indicating that the addition of the vdW interaction does not improve the prediction accuracy. The PMF model implicitly deals with many types of interactions, such as desolvation effects and enthalpy, and hence the effect of the vdW interaction is already included. This means that the high dependency of the vdW interaction on the PMF model hinders the improvement of the prediction accuracy of the scoring function. Such a strong correlation between descriptors is called multicollinearity.[73]

3.2. Analysis of Descriptors Affecting the Scoring Function.

To gain a chemical understanding of the improvement in prediction accuracy, we analyzed the contribution of each descriptor in the scoring function. Here, we focus on the partial regression coefficients (eq 4) in the PMF + MACCSKeys model trained with LASSO regression. The

Table 3. MACCSKeys Descriptors with Large Contribution

descriptor			descriptor		
key no.	key description	contribution (kcal/mol)	key no.	key description	contribution (kcal/mol)
23	NC(O)O	−0.74	92	OC(N)C	0.44
159	O > 1	−0.60	28	QCH2Q	0.4
27	I	−0.56	57	O HETEROCYCLE	0.38
36	S HETEROCYCRE	−0.45	47	SAN	0.37
131	QH > 1	−0.40	162	AROMATIC	0.33
71	NO	−0.36	125	AROMATIC RING >1	0.31
87	X!A$A	−0.36	163	6M RING	0.25
70	QNQ	−0.35	49	CHARGE	0.24
127	A$A!O > 1	−0.34	160	CH3	0.23
116	CH3AACH2A	−0.32	82	ACH2QH	0.21

reason for choosing this model is that MACCSKeys provides a clear chemical interpretation due to the one-to-one correspondence of its components to specific substructures and that LASSO allows us to easily estimate the contribution of descriptors based on the magnitude of the regression coefficient.

Table 3 summarizes the descriptors with the large contributions to the binding energies. It should be noted that the positive and negative values of regression coefficients correspond to repulsive and attractive interactions, respectively. As for the negative regression coefficients, we can see that there are many large contributions from heteroatoms such as key numbers 23, 71, 131, and 159, which are related to hydrogen bonding. In addition, the large contributions from halogen atoms such as key numbers 27 and 87 are shown, suggesting that the consideration of halogen effects, which could not be incorporated by PMF alone, leads to a higher prediction accuracy. As for the positive coefficients, the large contributions came from ring structures and aromatics such as key numbers 125, 162, and 163 and hydrocarbons such as 28, 82, and 160. These substructures cause destabilization to protein−ligand binding due to their positive coefficients. Therefore, it was suggested that these substructures are not suitable for drug design.

A similar analysis was performed for the PF fingerprints. Table 4 summarizes the descriptors of PF that have made significant contributions. Although the PF fingerprints based on count vectors are not simply comparable to MACCSKeys with binary vectors, we can see that the overall contribution of PF to the scoring function was smaller than that of MACCSKeys. In this study, we have constructed the PF fingerprints using three descriptors for each amino acid except glycine: the N-linked and C-linked main chain, and its side chain. The analysis showed that the contributions of the main chains were larger than those of the side chains. This was contrary to our chemical intuition that the side chain properties would give a larger contribution than the main chain. The result also showed a significant contribution from the cysteine side chain other than the main chain. This descriptor may compensate for interactions involving sulfur atoms that are not adequately represented by the PMF score. Another possibility is the incorporation of the disulfide bond effect. Since disulfide bonds are not considered in the descriptors other than the PF fingerprints, the contribution of the cysteine side chain may have been greater.

Table 4. PF Descriptors with Large Contribution

descriptor[a]	contribution (kcal/mol)	descriptor[a]	contribution (kcal/mol)
Asn-N	−0.13	Asn-O	0.16
His-O	−0.12	Gln-O	0.12
Cys-S	−0.12	Ala-O	0.10
Phe-N	−0.11	Lys-O	0.10
Met-N	0.10	Pro-O	0.08
Ile-O	−0.09	His-N	0.08
Asp-N	−0.09	Trp-N	0.07
Ala-N	−0.08	Cys-O	0.06
Gly-N	−0.07	Asp-O	0.06
Tyr-O	−0.07	Asn-O	0.06

[a] N-linked and O-linked main chain, and side chain are abbreviated as N, O, and S, respectively.

3.3. Systems with Improved and Unimproved Scores.

The above analysis succeeded in finding descriptors that contribute significantly to the new scoring function. We then turned our attention to specific systems in which the incorporation of ligand fingerprints greatly improved the computational results. For ease of interpretation, we again analyzed the PMF + MACCSKeys model trained with LASSO. Table 5 summarizes the calculated binding energies for the PMF and PMF + MACCSKeys models, and their chemical structures are illustrated in Figure 5. The results clearly show that the addition of MACCSKeys improves the binding energy by 1.84 and 0.50 kcal/mol for the systems with PDB ID of 4J3L and 3ARP, respectively. We further analyzed the MACCSKeys descriptors and found significant contributions of the halogen atom (key number 87: −0.36 kcal/mol) and functional groups involved in hydrogen bonding (key numbers 159 and 131: −0.60 and −0.40 kcal/mol). This result indicates that the PMF score alone cannot sufficiently describe these effects for compound **1**. The analysis also showed that MACCSKeys for compound **2** provides a large destabilization of the charged functional group (key number 49: 0.24 kcal/mol) due to its positive regression coefficient. This is a counterintuitive result because electrostatic interactions generally contribute significantly to protein−ligand interactions. The reason for this could be that the MACCSKeys descriptor of the charged functional group ameliorates the overestimation of the implicit electrostatic effect in the PMF score.

Table 5. Binding Energies for the PMF and PMF + MACCSKeys Models Along with the Experimental Values, Exemplifying Change in Predicted Value by Adding MACCSKeys (kcal/mol)

	model		
PDB ID	PMF	PMF + MACCSKeys	exptl.
4J3L	−8.58	−10.42	−10.67
3ARP	−10.41	−9.91	−9.82
3LKA	−8.88	−7.50	−3.87

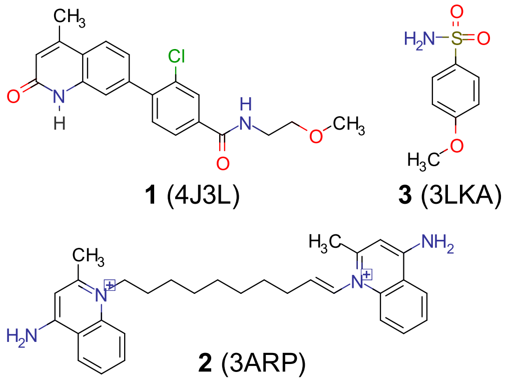


Figure 5. Ligand structures included in the test data set. For **1** and **2**, our scoring function showed improved prediction accuracy with the incorporation of ligand fingerprints. For **3**, the improvement over PMF score was not observed. The corresponding PDB IDs are shown in parentheses.

We also investigated systems with large deviations from the experimental values. As also shown in Table 5, the PMF + MACCSKeys model for a system with PDB ID of 3LKA gave a large error (3.63 kcal/mol) from the experimental value. Such a large error was also observed in the PMF + ECFP model trained with LightGBM. The error calculated by the PMF model alone was 5.01 kcal/mol, which means that the interaction is not correctly described at the stage by the PMF model. From these results, we can see that the inclusion of ligand fingerprints in the PMF score did not sufficiently improve the prediction accuracy for this system. In the PMF score, the nitrogen atom in compound **3** is classified as the atom type of ND. However, the environment of the nitrogen atom in sulfonamides would be different from that of most nitrogen atoms with the ND type because it is bound to a sulfonyl group. For this reason, its classification into ND is considered to be inappropriate for the nitrogen atom in compound **3**, and the error of the PMF model is likely to be large. As shown above, a model including ligand fingerprints can greatly improve the prediction accuracy for many systems, but no such effect was observed for 3LKA. The reason for this may lie in the training data. In this study, the PDBbind structures were used as training data without any structural optimization. In addition, no evaluation of the protonation state was performed for both the ligands and the proteins. Such lack of structural refinement may have caused the poor prediction accuracy for 3LKA. If we can prepare a large amount of training data with structural optimization and protonation state evaluation in the future, the prediction accuracy may be further improved.

4. CONCLUSIONS

In this study, we developed a new scoring function for the protein−ligand binding energy prediction by introducing the ligand fingerprints and protein fingerprints into the knowledge-based PMF function. Here, the machine learning techniques LASSO and LightGBM were used to train the scoring functions. The PMF + ECFP + PF model trained with LightGBM showed the best results, with a Pearson correlation coefficient of 0.79 and a RMSE of 1.64 kcal/mol, indicating a higher prediction accuracy for binding energy than the conventional scoring functions. Further analysis revealed that the ECFP and PF fingerprints are suitable for combination with PMF regardless of the two machine learning methods used in this study. To the best of our knowledge, this is the first example of introducing protein fingerprints into a scoring function. Although there are scoring functions that show better correlation coefficients than our results,[37−41] the results of this study successfully demonstrate the effectiveness of protein fingerprints as well as ligand fingerprints.

We also analyzed the components of the binding energy obtained from our newly developed scoring function and found that the inclusion of the ligand fingerprints into the PMF score improves prediction accuracy, especially for halogen atoms and functional groups involved in hydrogen bonding. On the other hand, we found a specific system in which the use of ligand fingerprints was not effective. This analysis strongly suggests the need for improving the quality of the training data with structural optimization and protonation state evaluation.

Incorporating ligand and protein structural information using the fingerprint methods into the scoring function is a promising approach for predicting binding energies. The PF fingerprints developed in this study are based on a representation using only three descriptors for each amino acid. Even such a simple handling of the protein structure significantly contributed to the improvement in prediction accuracy. A more appropriate representation of the protein structure would possibly lead to better results of the binding energies. The development of advanced protein fingerprints is a future challenge. The proposed methods and findings in this study will be useful for the development of computational drug discovery.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c02822.

Details on hyperparameter values in LASSO and the number of iterations for early stopping in LightGBM; list of PDB IDs used for the training data set, pretest data set, and test data set (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Kazuhiro J. Fujimoto − *Institute of Transformative Bio-Molecules (WPI-ITbM), Nagoya University, Nagoya 464-8601, Japan; Department of Chemistry, Graduate School of Science, Nagoya University, Nagoya 464-8601, Japan;* orcid.org/0000-0003-0286-3646; Email: fujimotok@chem.nagoya-u.ac.jp

Takeshi Yanai − *Institute of Transformative Bio-Molecules (WPI-ITbM), Nagoya University, Nagoya 464-8601, Japan; Department of Chemistry, Graduate School of Science, Nagoya University, Nagoya 464-8601, Japan;*
Email: yanait@chem.nagoya-u.ac.jp

Author

Shota Minami − *Department of Chemistry, Graduate School of Science, Nagoya University, Nagoya 464-8601, Japan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c02822

Notes

The authors declare no competing financial interest.
The atomic coordinates of protein−ligand complexes can be obtained from the PDBbind-CN database: http://www.pdbbind.org.cn/. The PDB IDs used for the training data set, pretest data set, and test data set are given in the Supporting Information. The python libraries (RDKit: https://www.rdkit.org/, LASSO: https://scikit-learn.org/stable/, and LightGBM: https://github.com/microsoft/LightGBM/) employed in our calculations are publicly available. The hyperparameter values in LASSO and the number of iterations for early termination in LightGBM are presented in the Supporting Information. Additional data are available from the authors upon request.

■ ACKNOWLEDGMENTS

This study was supported by JSPS KAKENHI (Grant 20K05430 to K.J.F. and 21H01881 and JP21K18931 to T.Y.) and JST, CREST (Grant JPMJCR21O5 to K.J.F.).

■ REFERENCES

(1) Wouters, O. J.; McKee, M.; Luyten, J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA* **2020**, *323*, 844−853.

(2) Kapetanovic, I. M. Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach. *Chem.-Biol. Interact.* **2008**, *171*, 165−176.

(3) Hughes, J. P.; Rees, S.; Kalindjian, S. B.; Philpott, K. L. Principles of early drug discovery. *Br. J. Pharmacol.* **2011**, *162*, 1239−1249.

(4) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. DOCKING AND SCORING IN VIRTUAL SCREENING FOR DRUG DISCOVERY: METHODS AND APPLICATIONS. *Nat. Rev. Drug Discov.* **2004**, *3*, 935−949.

(5) Meng, X.-Y.; Zhang, H.-X.; Mezei, M.; Cui, M. Molecular Docking: A powerful approach for structure-based drug discovery. *Curr. Comput.-Aided Drug Des.* **2011**, *7*, 146−157.

(6) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Jr. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66*, 334−395.

(7) Ferreira, L. G.; Dos Santos, R. N.; Oliva, G.; Andricopulo, A. D. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* **2015**, *20*, 13384−13421.

(8) Congreve, M.; Murray, C. W.; Blundell, T. L. Keynote review: Structural biology and drug discovery. *Drug Discovery Today* **2005**, *10*, 895−907.

(9) von Itzstein, M.; Wu, W.-Y.; Kok, G. B.; Pegg, M. S.; Dyason, J. C.; Jin, B.; Van Phan, T.; Smythe, M. L.; White, H. F.; Oliver, S. W.; Colman, P. M.; Varghese, J. N.; Ryan, D. M.; Woods, J. M.; Bethell, R. C.; Hotham, V. J.; Cameron, J. M.; Penn, C. R. Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **1993**, *363*, 418−423.

(10) Druker, B. J.; Lydon, N. B. Lessons learned from the development of an Abl tyrosine kinase inhibitor for chronic myelogenous leukemia. *J. Clin. Invest.* **2000**, *105*, 3−7.

(11) 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.

(12) 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.

(13) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(14) Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S. B.; Johnson, A. P. eHiTS: A new fast, exhaustive flexible ligand docking system. *J. Mol. Graphics Modell.* **2007**, *26*, 198−212.

(15) 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.

(16) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(17) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488−506.

(18) Corbeil, C. R.; Williams, C. I.; Labute, P. Variability in docking success rates due to dataset preparation. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 775−786.

(19) Jain, A. N. Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499−511.

(20) Uehara, S.; Fujimoto, K. J.; Tanaka, S. Protein-ligand docking using fitness learning-based artificial bee colony with proximity stimuli. *Phys. Chem. Chem. Phys.* **2015**, *17*, 16412−16417.

(21) 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.

(22) Chia-en, A. C.; Chen, W.; Gilson, M. K. Ligand configurational entropy and protein binding. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 1534−1539.

(23) Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.* **2007**, *28*, 1145−1152.

(24) Chen, J.; Brooks, C. L., III; Khandogin, J. Recent advances in implicit solvent-based methods for biomolecular simulations. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140−148.

(25) Huang, S.-Y.; Zou, X. Inclusion of solvation and entropy in the knowledge-based scoring function for protein− ligand interactions. *J. Chem. Inf. Model.* **2010**, *50*, 262−273.

(26) Kar, P.; Lipowsky, R.; Knecht, V. Importance of polar solvation and configurational entropy for design of antiretroviral drugs targeting HIV-1 protease. *J. Phys. Chem. B* **2013**, *117*, 5793−5805.

(27) 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.

(28) Muegge, I. A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. *Perspect. Drug Discovery Des.* **2000**, *20*, 99−114.

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

(30) Muegge, I. PMF scoring revisited. *J. Med. Chem.* **2006**, *49*, 5895−5902.

(31) 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. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2015**, *5*, 405−424.

(32) 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.* **2020**, *10*, No. e1429.

(33) Li, H.; Sze, K.-H.; Lu, G.; Ballester, P. J. Machine-learning scoring functions for structure-based drug lead optimization. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *10*, No. e1465.

(34) 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.

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

(36) Guedes, I. A.; Barreto, A.; Marinho, D.; Krempser, E.; Kuenemann, M. A.; Sperandio, O.; Dardenne, L. E.; Miteva, M. A. New machine learning and physics-based scoring functions for drug discovery. *Sci. Rep.* **2021**, *11*, 1−19.

(37) Cang, Z.; Mu, L.; Wei, G.-W. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Comput. Biol.* **2018**, *14*, No. e1005929.

(38) Meng, Z.; Xia, K. Persistent spectral−based machine learning (PerSpect ML) for protein-ligand binding affinity prediction. *Sci. Adv.* **2021**, *7*, eabc5329.

(39) Sánchez-Cruz, N.; Medina-Franco, J. L.; Mestres, J.; Barril, X. Extended connectivity interaction features: improving binding affinity prediction through chemical description. *Bioinformatics* **2021**, *37*, 1376−1382.

(40) Wang, Z.; Zheng, L.; Liu, Y.; Qu, Y.; Li, Y.-Q.; Zhao, M.; Mu, Y.; Li, W. OnionNet-2: A Convolutional Neural Network Model for Predicting Protein-Ligand Binding Affinity based on Residue-Atom Contacting Shells. *Front. Chem.* **2021**, 913.

(41) 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. *Brief. Bioinform.* **2022**, *23*, bbab527.

(42) Boyles, F.; Deane, C. M.; Morris, G. M. Learning from the ligand: using ligand-based features to improve binding affinity prediction. *Bioinformatics* **2020**, *36*, 758−764.

(43) Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. Ser. B* **1996**, *58*, 267−288.

(44) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.

(45) Brown, R. D.; Martin, Y. C. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(46) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.

(47) Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *J. Mol. Graphics Modell.* **2010**, *29*, 157−170.

(48) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screen.* **2005**, *10*, 682−686.

(49) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.

(50) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(51) Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J. Chem. Doc.* **1965**, *5*, 107−113.

(52) Engel, T.; Gasteiger, J., *Chemoinformatics: Basis Concepts and Methods*. Wiley-VCH: Weinheim, 2018.

(53) Zhang, K.; Liu, X.; Shen, J.; Li, Z.; Sang, Y.; Wu, X.; Zha, Y.; Liang, W.; Wang, C.; Wang, K.; Ye, L.; Gao, M.; Zhou, Z.; Li, L.; Wang, J.; Yang, Z.; Cai, H.; Xu, J.; Yang, L.; Cai, W.; Xu, W.; Wu, S.; Zhang, W.; Jiang, S.; Zheng, L.; Zhang, X.; Wang, L.; Lu, L.; Li, J.; Yin, H.; Wang, W.; Li, O.; Zhang, C.; Liang, L.; Wu, T.; Deng, R.; Wei, K.; Zhou, Y.; Chen, T.; Lau, J. Y.-N.; Fok, M.; He, J.; Lin, T.; Li, W.; Wang, G. Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography. *Cell* **2020**, *181*, 1423−1433.

(54) Kopitar, L.; Kocbek, P.; Cilar, L.; Sheikh, A.; Stiglic, G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci. Rep.* **2020**, *10*, 11981.

(55) Bar, N.; Korem, T.; Weissbrod, O.; Zeevi, D.; Rothschild, D.; Leviatan, S.; Kosower, N.; Lotan-Pompan, M.; Weinberger, A.; Le Roy, C. I.; Menni, C.; Visconti, A.; Falchi, M.; Spector, T. D.; The IMI DIRECT consortium; Adamski, J.; Franks, P. W.; Pedersen, O.; Segal, E. A reference map of potential determinants for the human serum metabolome. *Nature* **2020**, *588*, 135−140.

(56) Ma, X.; Sha, J.; Wang, D.; Yu, Y.; Yang, Q.; Niu, X. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electron. Commer. Res. Appl.* **2018**, *31*, 24−39.

(57) Chun, P.-j.; Izumi, S.; Yamane, T. Automatic detection method of cracks from concrete surface imagery using two-step light gradient boosting machine. *Comput.-Aided Civ. Infrastruct. Eng.* **2021**, *33*, 61−72.

(58) Xiaolei, S.; Mingxi, L.; Zeqian, S. A novel cryptocurrency price trend forecasting model based on LightGBM. *Finance Res. Lett.* **2020**, *32*, 101084.

(59) Li, H.; Peng, J.; Sidorov, P.; Leung, Y.; Leung, K.-S.; Wong, M.-H.; Lu, G.; Ballester, P. J. Classical scoring functions for docking are unable to exploit large volumes of structural and interaction data. *Bioinformatics* **2019**, *35*, 3989−3995.

(60) Loh, W. Y. Classification and regression trees. *Data Min. Knowl. Discov.* **2011**, *1*, 14−23.

(61) Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367−378.

(62) Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **2021**, *54*, 1937−1967.

(63) Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J. D.; Zardecki, C. The protein data bank. *Acta Crystallogr. D Biol. Crystallogr.* **2002**, *58*, 899−907.

(64) 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.

(65) Li, Y.; Su, M.; Liu, Z.; Li, J.; Liu, J.; Han, L.; Wang, R. Assessing protein-ligand interaction scoring functions with the CASF-2013 benchmark. *Nat. Protoc.* **2018**, *13*, 666−680.

(66) Li, Y.; Liu, Z.; Li, J.; Han, L.; Liu, J.; Zhao, Z.; Wang, R. Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set. *J. Chem. Inf. Model.* **2014**, *54*, 1700−1716.

(67) 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.

(68) Jones, J. E. On the determination of molecular fields.-II. From the equation of state of a gas. *Proc. R. Soc. London, Ser. A* **1924**, *106*, 463−477.

(69) Wang, J.; Cieplak, P.; Kollman, P. A. How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? *J. Comput. Chem.* **2000**, *21*, 1049−1074.

(70) https://www.rdkit.org/ (13 April 2021),

(71) https://scikit-learn.org/stable/ (14 April 2021),

(72) https://github.com/microsoft/LightGBM/releases/tag/v3.2.1 (31 May 2021),

(73) Farrar, D. E.; Glauber, R. R. Multicollinearity in Regression Analysis: The Problem Revisted. *Rev. Econ. Stat.* **1967**, *49*, 92−107.