Machine Learning on Ligand-Residue Interaction Profiles to Significantly Improve Binding Affinity Prediction

Beihong Ji, Xibing He, Jingchen Zhai, Yuzhao Zhang, Viet Hoang Man, Junmei Wang*

Department of Pharmaceutical Sciences and Computational Chemical Genomics Screening Center, School of Pharmacy, University of Pittsburgh, Pittsburgh, PA 15261, USA.

* Corresponding author, junmei.wang@pitt.edu

Beihong Ji is a PhD student in the School of Pharmacy, University of Pittsburgh. Her research interest is the development of tools for drug discovery and target identification in drug abuse.

Xibing He is a Research Scientist in the School of Pharmacy, University of Pittsburgh. His research interests include force-field development, free-energy calculations and computer-aided drug design.

Jingchen Zhai is a MS student in the School of Pharmacy, University of Pittsburgh. Her research interests include molecular dynamic simulation, free energy calculations, large-scale virtual screening and PBPK modeling.

Yuzhao Zhang is a MS student in the School of Pharmacy, University of Pittsburgh. Her research interest focuses on computational chemistry and medicinal chemistry.

Viet Hoang Man is a Research Scientist in the School of Pharmacy, University of Pittsburgh. His research focuses on amyloid aggregation, protein-ligand/protein interactions and force field development,

Junmei Wang is an associate professor in Department of Pharmaceutical Sciences and Computational Chemical Genomic Screening Center, School of Pharmacy, University of Pittsburg.

His research interest includes computer modeling and simulation of protein-ligand interactions and other biological events, and pharmacometrics and computational systems pharmacology.

Abstract

Structure-based virtual screenings (SBVSs) play an important role in drug discovery projects. However, it is still a challenge to accurately predict the binding affinity of an arbitrary molecule binds to a drug target and prioritize top ligands from a SBVS. In this study, we developed a novel method, using ligand-residue interaction profiles (IPs) to construct machine learning (ML)-based prediction models, to significantly improve the screening performance in SBVSs. Such a kind of the prediction model is called an IP scoring function (IP-SF). We systematically investigated how to improve the performance of IP-SFs from many perspectives, including the sampling methods before interaction energy calculation and different ML algorithms. Using six drug targets with each having hundreds of known ligands, we conducted a critical evaluation on the developed IP-SFs. The IP-SFs employing a gradient boosting decision tree (GBDT) algorithm in conjunction with the MIN+GB simulation protocol achieved the best overall performance. Its scoring power, ranking power and screening power significantly outperformed the Glide SF. First, compared to Glide, the average values of mean absolute error and root mean square error of GBDT/MIN+GB decreased about 38% and 36%, respectively. Second, the mean values of squared correlation coefficient and predictive index increased about 225% and 73%, respectively. Third, more encouragingly, the average value of the areas under the curve of receiver operating characteristic (ROC) for six targets by GBDT, 0.87, is significantly better than that by Glide, which is only 0.71. Thus, we expected IP-SFs to have broad and promising applications in SBVSs.

Key Words: Structure-Based Virtual Screening (SBVS), Machine Learning (ML); Scoring Function (SF); Binding Affinity Prediction; Scoring Power; Machine Learning-Based Scoring Function (ML-based SF)

Introduction

In the past decades, computer-aided drug design (CADD) has proven to be a powerful technique to reduce the cost and accelerate the process in the development of drug candidates [1, 2]. As a crucial step in CADD, structure-based virtual screenings (SBVSs) play an important role in drug discovery projects. In a SBVS, a large library of compounds is screened by computational techniques like molecular docking for the recognition of potential bioactive ligands [3-5]. Necessarily, to assist the process of compounds screening in SBVSs, a wide variety of scoring functions (SFs) have been developed these years [6-26]. The SFs typically fulfill several different desirable tasks and hence are usually evaluated by the following four criteria [27]: (1) the docking power (the capability to detect near-native binding modes from computer-generated decoy poses), (2) the screening power (the capacity to distinguish active binders from inactive or decoy compounds), (3) the ranking power (the ability of ranking ligands of the same receptor correctly according to their binding affinities), and (4) the scoring power (the capability to achieve a high correlation between predicted and experimental binding affinities of different receptor-ligand complexes). The former three powers (docking, screening, and ranking) are inherently correlated to the scoring power. However, the predictive accuracy of SFs is still one of the most challenging problems remained in the field of computational chemistry [9, 28, 29]. Typically, classical scoring functions (CSFs) are classified into three categories: force field (FF)-based SF, knowledge-based SF and empirical SF [30]. FF-based SFs are described as a sum of non-bonded and bonded interactions energies computed from given FFs [14]. Since those FFs were mainly generated for the modeling of intermolecular energies, some terms such as entropy and desolvation energy are excluded in FF-based SFs [31]. Knowledge-based SFs use the 3D coordinates of a large set of protein-ligand complex structures as a knowledge base to characterize a complex based on some

features, which is usually the occurrence frequencies of atom-atom pairwise contacts [32, 33]. Empirical SFs can be considered to compute the protein-ligand binding through the sum of contribution terms [27]. They utilize similar function forms as FF-based SFs but contain additional ones such as solvent accessibility surface area (SASA). Another difference between empirical SFs and FF-based SFs is that for the former one, each energy term has its own weight, which can be derived by linearly fitting terms to measured binding affinities. A fact is that CSFs are developed based on some simplifications for the sake of efficiency, so it is difficult for them to completely account for certain processes of protein-ligand recognition, which further affect their ability of molecule ranking and selection [34]. The major limitations of CSFs include the lack of structural flexibility of target protein and effects of solvation [35]. In addition to those SFs, there are other more accurate computational methods based on the combination of molecular dynamics (MD) and free-energy sampling algorithms, such as rigorous alchemical free energy methods (TI, FEP, etc.) [36, 37] or end-point methods (MM-GBSA, MM-PBSA, LIE, etc.) [38, 39]. Without doubt, ones can obtain the prediction of binding affinity with more accuracy by using these approaches. However, considering the massive demands on computational resource and time, it is not practical to screen large libraries of compounds by utilizing those expensive techniques.

Recently, an alternative methodology for the binding affinity prediction has gradually emerged and gained more and more attention. Unlike CSFs with predetermined functional form, non-parametric machine learning (ML)-based SFs can learn feature information from training data and adjust the parameters to further improve the performance of prediction. Since ML-based SFs are able to capture binding interactions between ligands and target receptors that are hard to be modeled by CSFs, they are considered to generate more accurate prediction of binding affinity. In addition, recently, a wide variety of ML algorithms have been developed, such as support vector

machine (SVM) [40], gradient boosting decision tree (GBDT) [41], adaptive boosting (Adaboost) [42], random forest (RF) [43] and artificial neural network (ANN) [44, 45], etc. In the last few years, there have been a surge to apply ML techniques in building SFs and researchers have made their efforts to explore the predictive effects of a number of ML algorithms by using different feature representations [14-20]. One of the common features used to characterize the protein-ligand complex in a ML-based SF is geometrical features. For example, Ballester and Mitchell presented a ML-based SF, RF-Score, with the application of RF to predict protein-ligand binding affinity. In their study, the relationship between the structure of complex and the binding affinity was characterized by the intermolecular interaction features, which comprise the number of occurrences of interacting atom pairs within a specific distance threshold. The features they used to characterize the structure of complexes included several binding and geometrical characteristics. Knowledge-based pairwise potentials, such as SVR-KB and SVM-SP SFs developed by Li et al, are also a kind of feature descriptor that is commonly used. Another important feature representation that is considered in the binding affinity prediction can be categorized as physico-chemical descriptors. SVR-EP, SFCscore^{RF} and B2BScore SFs were all developed by using physico-chemical features like ligand molecular weight, van der Waals energy, hydrogen bonds, hydrophobic effects, etc. There are also a number of ML-based SFs which were developed based on more complex features from a variety of categories. For instance, Khamis and Gomaa proposed ML-based SFs that depend on a wide range of features, including geometrical features, energy terms and pharmacophore features [20].

Traditionally, the feature sources used to characterize the protein-ligand interactions in ML-based SFs are energy terms from CSFs [18, 46, 47], basic structural features [14, 48] or structural interaction fingerprints [49]. However, as a key step in the development of the ML-based SFs, the

study of feature descriptors is still on the way and which feature representations can better improve the performance remains a burning question. Here, to improve the scoring power of CSFs, we proposed a novel ML-based SF, IPscore, by incorporating the interaction profile (IP) described by ligand-residue interaction energies as the feature descriptor to predict the protein-ligand binding affinity. To derive this unique feature, our computational protocol consists of molecular docking, geometry minimization and/or molecular dynamics relaxation, and molecular mechanics-Generalized Born surface area (MM-GBSA) free energy decomposition. Previously, Li and Hou at al. have improved the performance of MIEC-SVM SF by utilizing similar molecular modeling technique to derive the feature representation [50]. But they only emphasized to improve the screening power of SF, i.e., the ability to discriminate active molecules from non-actives [51]. Besides, their good performance relied on different combinations of energy components under the framework of MM-GBSA. More importantly, in our work, we evaluated the performance of seven ML algorithms under different scenarios, which were created based on different processes of free energy calculation, including molecular mechanics (MM) minimization using either an implicit water model of Generalized Born (GB) [52] or an explicit water model of TIP3P [53] to account for the solvent effect, and the application of MD to relax the docking complex prior to the minimization or not. By this way, we identified the best structure processing condition before binding profile generation to maximize the performance of our IP-SFs measured by both the scoring power and the ranking power. The essential steps of developing IP-SF for a protein target were demonstrated in **Figure 1**. In the following sessions, the detail of the model establishment and the performance evaluation were described.

Materials and Methods

Dataset preparation

To train and test the developed ML-based SFs in our work, six drug receptors that come from two large protein categories, G protein-coupled receptors (GPCR) and kinases, were cherry-picked due to not only their importance in drug discovery, but also their large amounts of reported compound activities in the binding assays from the ChEMBL database (https://www.ebi.ac.uk/chembl/) [54, 55]. Among the six drug targets, serotonin 2A receptor (5HT2AR), adenosine A2A receptor (A2AR), cannabinoid receptor 1 (CB1), and muscarinic acetylcholine M1 receptor (M1R) belong to GPCRs, while extracellular signal-regulated kinase 2 (ERK2) and vascular endothelial growth factor receptor 2 (VEGFR2), are members of kinases. The X-ray crystal structures of all targets and their co-crystallized ligands were obtained from Protein Data Bank (https://www.rcsb.org/) [56, 57]. The crystal structures selected to participate in ligand-residue IP calculations meet the following conditions, listed in descending order of their importance: (1) the structure is for a protein that belongs to human (Homo sapiens); (2) the structure is in complex with a co-crystallized modulator; (3) the structure has no missing segments; and (4) the X-ray diffraction resolution of the structure is low. The detail information and structures of receptors were listed in **Table S1** and shown in **Figure S1**.

For each drug target, up to 1200 compounds whose inhibition constants (K_i) collected by ChEMBL were randomly selected. To ensure the distribution of K_i values was not altered during the random selection, we grouped each compound into four activity categories: potent (K_i < 10 nM), less potent (10 nM ≤ K_i < 1 μM), weak (1 μM ≤ K_i < 100 μM) and very weak (100 μM ≤ K_i < 1 M). For each category, up to 300 compounds were randomly selected by using numpy.random.choice built in Python 3.7 program [58]. If the number of compounds in a category is less than 300, then all compounds in that category are collected. For the collected compounds with two or more K_i activities from different binding assays, the averaged K_i values were adopted. To evaluate the

screening power of IP-SF, we manually classified the selected compounds into the active set and decoy set by the cutoff of K_i = 100 nM, which is lower than normal threshold (10 μM), in order to better balance the numbers of compounds in the active and decoy sets. The selected compounds were then randomly allocated into the training set and test set with the ratio of 4:1. The number of compounds in the active and decoy sets as well as the training and test sets were listed in **Table 1**.

Molecular docking

Selected compounds were docked to their receptors using the Glide [59, 60] docking program built in the Schrodinger software (Maestro 11.2). Before docking, each protein target was first treated utilizing the *Protein Structure Preparation Wizard* module following the standard procedures, i.e., removing water and co-crystal solvent, filling in missing hydrogen atoms, etc. It is noted that only the added hydrogen was movable during the geometry optimized using the OPLS force field. Next, a grid file was generated for the drug target using the cocrystal ligand to define the center of the binding pocket, and no rotatable or constraint group was defined. We applied the default option to define the enclosing box size and the generated grid file is suitable to dock ligands that have similar sizes as that of the cocrystal ligand. Last, Glide flexible docking using the Glide SP scoring function was performed with the following settings: the partial charge cutoff for ligands was 0.15, van der Waal radius scaling factor was 0.80 and the formation of intramolecular hydrogen bonds was rewarded. For each ligand, the top docking pose that has the best docking score was subjected to minimizations and/or MD simulations, and subsequent ligand-residue interaction profile calculation as detailed below.

System setup for MM calculations

Before MM minimization and MD simulations, AM1-BCC charges [61, 62] were computed for all ligands by the SQM and Antechamber modules in AMBER 18 [63]. The reason we chose AM1-BCC instead of RESP charge method [64] was to reduce the computational cost. Other force field parameters for ligands came from GAFF [65], while the AMBER FF14SB [66] force field was assigned for the proteins. The residue topology of a ligand was generated by utilizing the Antechamber module [67] in the AMBER software package [68, 69]. For 5HT2A receptor, a Zn^{2+} complex was formed with two HIS and two GLU residues. We developed FF14SB/GAFF-compatible force field parameters for the Zn^{2+} complex and the details was provided in the supporting information.

MM minimization and MD simulations were performed using either an explicit water model or an implicit water model to account for solvent effect. For the formal, the protein complex was first immersed in a rectangle TIP3P water box with the shortest distance between any atom of the complex and box borders is at least 14 Å; then a certain number of Cl^- and Na^+ were added to the solvent box so that their concentration is about 0.15 M and the whole system was neutralized. The detailed computation protocol of adding ions was provided in the supporting information. Particle mesh Ewald (PME) was used to treat the long-range electrostatic interactions [70] and the nonbonded interaction cutoff was set to be 10 Å to handle the real-space interactions. For the latter, the "standard" pairwise generalized Born model by Hawkins, Cramer, and Truhlar [71] was employed. No count ions were added to neutralize the MD system.

MM minimization and MD simulation

We have two strategies to prepare the complex structures before the binding profile calculations. First, a complex was relaxed and optimized only through a set of minimization procedures. We constrained the main chain atoms of the receptor and the bound ligand by using a harmonic

potential. The harmonic potential force constant was reduced from 10 to 5, 2, 1 and 0 (kcal/mol)/Å^2 for the receptor, while it was decreased from 100 to 50, 20, 10 and 0 (kcal/mol)/Å^2 for the ligand, progressively in five 1000-step minimizations. Second, after the 5000-step minimizations, the system was further relaxed by a set of 10,000-step MD simulations with the restraint settings the same as those for minimizations. All MD simulations were performed at 298 K and 1 atm with a time step of 2 femtoseconds. Finally, the whole system was relaxed without any restraint for 1000 steps. All minimization and MD simulations were conducted using the PMEMD module in the AMBER 18 [72, 73].

In summary, we consider four computational protocols for processing a complex structure before the interaction profile generation: (a) MIN+GB (minimization only, using the GB model to account for solvent effect); (b) MIN+TIP3P (minimization only, using explicit the TIP3P water model); (c) MD+GB (minimization plus MD simulation, using the GB model); and (d) MD+TIP3P (minimization plus MD simulation, using the explicit TIP3P water model).

MM-GBSA residue-ligand free energy decomposition

For the two minimization protocols, the optimized complex at the end of a series of stepwise minimizations was employed for the residue-ligand binding free energy decomposition analysis. However, for the two MD protocols, we added an extra step to fully minimize the last MD snapshot, so that a high-quality conformation was used for the interaction profile calculation. We calculated the MM-GBSA residue-ligand interaction energies using an internal program which analyzes the outputs of Sander and conduct statistics on each component of MM-GBSA free energy. The interior and exterior dielectric constants were set to 1.0 and 80, respectively. For each system, if a residue has its interaction energy with any ligand is lower than -0.1 kcal/mol, the residue is recognized as a key residue for charactering the binding profile. The MM-GBSA interaction

energies between all key residues and a ligand constitutes the binding profile of this ligand. The numbers of key residues using different structure processing protocols were listed in **Table 2.** The and the corresponding interaction profiles of each drug receptor are shown in **Figures S2-S6.** For the sake of convenience, we adopted the new residue IDs generated by AMBER to label those key residues, and the correspondence between the new and the original residue IDs in the PDB files was presented in **Table S2**.

ML algorithms

Seven popular ML algorithms were applied and compared in this study. They are ordinary least squares (LS) regression [74], Bayesian regression [75], support vector regression (SVR) [76], RF [43], Adaboost [42], GBDT [41] and the neural network model, multi-layer perceptron (MLP) [77]. ML regression models were constructed utilizing the Scikit-learn [78] package implemented in the Python program, except for MLP which was performed using the Keras [79] module in Python. The feature matrix consisted of MM-GBSA interaction energies between ligands and key residues (< -0.1 kcal/mol) served as the input data for each ML algorithm. No extra processing was taken to the data as all input data was extracted following the same protocol and was at the same scale. The inputs, i.e., feature matrixes or IPs for training ML models were provided in the supplemental material **(Tables S3-S6)**. The description of ML algorithms and their corresponding hyperparameters are displayed in **Table S7**. An example code for the training process is shown in **Figure S6**

Additionally, the performance of deep neural network (DNN) [80], an effective and advanced technique evolved from ANN was also studied. However, considering it is more time-consuming to construct DNN model and it requires a large amount of data to avoid the overfitting issue, we only used it in the construction of SF models for the MIN+GB structure processing protocol. DNN

models were constructed using the Keras sequential model [79] in Python. For each target, 20 % of the training data was randomly selected to form a validation set. For each epoch, the performance of the model on the validation set was evaluated. The loss function during the learning process was mean squared error (MSE). Mean absolute error (MAE) was also calculated as a metric of performance during the training processes. The losses and MAE changes during processes are shown in **Figures S7** and **S8**. The architectures of the DNN models for each system were described in **Table S8**.

Performance metrics

To evaluate the scoring and ranking power of IP-SFs, four metrics were adopted, i.e. root mean square error (RMSE), MAE, squared correlation coefficient (CORR2) and predictive index (PI) [81-83], to estimate the difference between predictive and experimental binding affinity of compounds in testing sets for each target. RMSE and MAE are measure of scoring power, while CORR2 and PI are measure of ranking power. The performance of Glide docking score was taken as the standard to evaluate whether the developed IP-SFs can improve the docking performance. Lower RMSE and MAE, and higher CORR2 and PI indicate better predictive performance.

As for the assessment of screening power, we used the area under the curve (AUC) of receiver operating characteristic (ROC) curve, Accuracy, F1-score and enrichment factor (EF) at 10% ($EF_{10\%}$) and 40% levels ($EF_{40\%}$) [84] as the performance metrics. All of these three metrics range from 0 to 1, with 0 indicating the worst and 1 indicating the best scenarios. However, for ROC, a random model theoretically has an AUC value of 0.5 and a good classifier typically has AUC value of 0.8 or larger. EF is one of the commonly used metrics in virtual screening studies. It is defined as the proportion of the true active binders in the sampled subset relative to the proportion of true

active binders in the total dataset [85, 86]. Similar to the evaluation of scoring and ranking power, we adopted the performance of Glide SF as the reference.

Results and Discussion

With the combination of different solvent models (implicit GB or explicit TIP3P) and structure optimization approaches (minimization only or minimization plus MD), in total four structure processing protocols were investigated. For each protocol, all developed ML-based SFs were first trained on the training sets and then tested on the testing sets. The performance of Glide docking score was utilized as the reference. The docking performance of newly developed ML-based SFs by different ML algorithms under four scenarios were compared and discussed below.

Comparison of performance of ML-based scoring functions on different targets

The performance of several ML-based SFs versus CSF (Glide docking) was presented by the radar chart (**Figure 2**) and **Table S9**. Their scoring power was evaluated by RMSE, MAE, and ranking power by CORR2 and PI. Generally, the ranking power of Glide SP scoring function was poor and it varies from target to target, e.g., CORR2 is only 0.05 for CB1 receptor. Even ERK has the best CORR2 value, 0.27, its ranking power is still not satisfactory. Our IPscore significantly improved the ranking power in most cases, regardless of different ML methods and structure processing protocols. As shown in **Table S9**, there are 28 PI-SFs constructed for one drug target, so totally there were 168 PI-SFs built in the study. Among all of 168 constructed models, only 18 models have at least one metric worse than the Glide SF, and most of them were constructed using LS and Bayesian algorithms. As far as the drug target is concerned, CB1 receptor and M1R have 7 and 9 less-satisfactory models. For all the drug targets, most PI-SFs outperformed the Glide SF in terms of scoring and ranking power. The performance of best models for each metric and drug target is

summarized in **Table 3**. The mean values calculated from the best models for the six drug targets are 1.11 kcal/mol, 1.40 kcal/mol, 0.53 and 0.72, for MAE, RMSE, CORR2 and PI, respectively. Compared to Glide SP docking SF, the MAE and RMSE dropped about 40% and 38%, respectively, while CORR2 and PI increased about 231% and 76%, respectively. Of note, this excellent performance was achieved using different machine learning methods with different structure processing protocols. In the next section, we will identify the best ML method and computational protocol which produce the best overall performance.

Interestingly, the performance of our developed models varied on different targets. Taking the performance of GBDT models for the six targets as an example, the CORR2 values using GBDT models varied from 0.29 to 0.74 according to **Table 3**. Such a difference in performance of GBDT for different targets may be caused by the quality of Glide docking poses varied from a system to another. For example, the CORR2 for 5HT2AR using Glide SF is 0.06, whereas it is much higher (0.27) for ERK2. In our approach, traditional docking SF is utilized as the initial step, followed by a series of molecular modeling and ML/DL methods. Because the best docking pose for each compound was selected as the starting structure for the following MM minimizations or MD simulations, the performance of GBDT models can be affected by the initial results of docking method. However, as we emphasized previously, the essential performance of our models was significantly superior to Glide SF on all tested protein targets. It is pointed out that the differences in the GBDT performance may also be caused by other various reasons, such as the quality of bioassays, the quality of crystal structures used for study, and the range of measured activities of the ligands.

Impacts of ML algorithms on performance

To compare the performance of different ML methods, we ranked them according to four metrics. **Figure 3** shows the ranking results of seven ML approaches for all receptors with four structure preparation protocols. With the scoring power changing from best to worst, the color of grid turned from red to green. An overview from **Figure 3** indicates that the GBDT model contains most red grids, illustrating its best predictive performance among all ML algorithms. Other than GBDT, most red grids concentrated at RF, AdaBoost and MLP, demonstrating their relatively better performance than the rest of ML methods. Meanwhile, LS and Bayesian have most green grids, indicating IP-SFs constructed using linear regression algorithms perform poorly. These results suggested that the relationship between ligand-residue interaction energies and binding affinity is not simple additive, as such machine learning algorithms which can bring nonlinearity to the input data can generate better predictive models.

Based on the ranking results, we created a frequency distribution histogram for each structure processing protocol. The rank of a ML algorithm, which varies from 1 to 7, represents the ML algorithm's ranking measured by a specific metric for a specific drug target. In total, there are 24 ranks for each ML algorithm as there are six drug targets and four measuring metrics. The distribution of the ranks of all ML algorithms are demonstrated in **Figure 4**. The histograms were colored according to the following rule: the smaller the rank, the darker the color it is. Thus, according to **Figure 4**, the SF that was built based on GBDT obtained the best predictive performance among all ML algorithms as it has the darkest columns no matter what structure processing protocol was applied.

Impacts of simulation protocols on performance

Four different structure processing protocols were tested to generate ideal complex structures prior MM-GBSA binding free energy decomposition analysis. To better explore the influence of

different structure processing protocols on model performance, we evaluated the performance of IP-SFs generated by using GBDT, the ML algorithm that was found to have the best overall performance. Then we compared the performance of four protocols, MD+GB, MD+TIP3P, MIN+GB, MIN+TIP3P by using the average performance of six targets. See **Figure 5**. Overall, there was no dramatic difference observed among those four protocols. Interestingly, we found that the performance of MIN+GB and MIN+TIP3P were slightly better than that of MD+GB and MD+TIP3P, suggesting that the additional structural relaxation using MD simulations may not be necessary. However, it is pointed out that for some drug targets, especially those do not have high resolution structures, additional structural relaxation using MD simulations may still be needed. As to the two types of water models, there was no big difference in the performance of IP-SFs between MIN+GB and MIN+TIP3P either. As shown in **Figure 5**, MIN+GB outperformed MIN+TIP3P slightly on the metrics of CORR2 and PI, while MIN+TIP3P has slightly better performance on MAE and RMSE.

Besides the performance, the computational efficiency of the four structure processing protocols is also a key criterion to be considered in practice. For docking scoring function, we conducted Glide docking in this study using Central Processing Unit (CPU). For MM minimization and MD simulation processes, to balance the calculated accuracy and timing, we used GTX 1080 Graphical Processing Unit (GPU) to accelerate the calculation procedure. In recent years, GPU MD is routinely conducted in both academia and industry. Considering the two minimization protocols perform better than the ones utilizing MD simulations, here we only compared the timings of the MIN+GB and MIN+TIP3P protocols. For instance, the average CPU time for docking each compound to the CB1 target with Glide is 11.1 s. The average GPU time for MM minimization for each compound in CB1 is 11.9 s with GB implicit water model, and 33.8 s with TIP3P explicit

water model. The average time cost by using GB model is less than using more complexed TIP3P water model, and more similar to Glide docking. Similarly, as to the ERK2 target, the average time for each ligand is 8.5 s for Glide docking, but 15.3 s and 61.6 s for minimizations with GB implicit water and TIP3P explicit water, respectively. Overall, the wall time of running Glide docking and MIN+GB are comparable, suggesting the IP-SF calculation using the MIN+GB protocol is very efficient.

Taken together, we concluded that the simplest structure processing protocol, MIN+GB has not only the best overall performance but also the best computational efficiency. Hence it should be first applied to generate ligand-residue interactions profiles for most drug targets. Therefore, the ideal combination of ML-based IP-SF and structure processing protocol, GBDT in conjunction with MIN+GB, was identified. Compared to Glide SF, averagely the MAE decreased around 38% and 36%, while CORR2 and PI can increase about 225% and 73%.

Assessment the performance of IP-SFs constructed by using deep learning

Deep learning is a part of a big family of ML techniques. DNN evolves from the traditional ANN, where multiple layers are involved in the construction of models to extract higher level of features hidden in the input data [87]. Recently, DNN has gained considerable attention from computer-aided drug design community and has been widely applied in a variety of fields, such as drug repurposing [88, 89], structure prediction [90], etc. It is of interest to evaluate the performance of IP-SFs constructed using DNN. We only assessed and compared the performance of DNN-based IP-SFs following the MIN+GB structure processing protocol due to relatively large computational cost. As shown in **Table 3** and **Figure 6**, the docking performance of DNN-based IP-SFs is superior to Glide docking SF, but inferior to GBDT-based IP-SFs. This result although

disappointing, is understandable. The interaction profile data used to train predictive models is quite small and DNN usually performs better with big data.

Screening power of the best-performed IP-SF in SBVSs

Besides ranking power, another widely used metric for evaluating virtual screening performance is the screening power, which measures the ability of a SF to recognize active ligands as hits. As we have identified GBDT with MIN+GB has the best overall performance on scoring power and ranking power, the screening power evaluation was only conducted for this IP-SF. Several commonly-used metrics, including AUC of ROC curves, Accuracy, F1-score and EF, were applied to compare the screening power of GBDT and Glide SF. For EF, we calculated the $EF_{10\%}$ and $EF_{40\%}$, for screenings which enriched 10% and 40% of the total compounds, respectively. It is obvious that GBDT achieved better performance than Glide SF as measured by all the metrics as listed in **Table 4**. Particularly, the mean AUC of ROCs, 0.87, is significantly better than that of Glide SF, which is only 0.71. On average the AUC of IP-SF for six targets improved approximately 22.5% in comparison with the AUC of Glide docking, however, if we excluded ERK2 target for which the difference between two SFs is minimal due to the imbalance between the numbers of actives (758) and inactives (92), the improvement of AUC value increased to 31%. The ROC plots were illustrated in **Figure 7**. It is worth to pointing out that the inactives defined in this study are those having measured activities larger than 0.1 μM. We used a relatively low threshold to achieve a good balance between the numbers of actives and inactives. If we introduced real decoys, the screening power measured by the above metrics, especially EF and AUC, could be much better.

Conclusion

To improve the screening performance of traditional SFs, we applied a novel approach to construct ML-based SFs using residue-ligand interaction energies as input. The construction process and evaluation of IP-SFs were presented. The interaction profile of a ligand binding to a receptor was calculated after a set of structure preparation processes including molecular docking, molecular mechanics minimization and/or molecular dynamics simulation to relax and optimize the complex, and MM-GBSA free energy decomposition. Key residues were identified for each drug target by analyzing the interaction profiles of its ligands. Predictive models were then constructed using a variety of machine learning methods. Several evaluation tests on six protein targets were conducted to assess the IP-SPs. 150 out of 168 IP-SPs outperformed the original Glide docking SF measured by all four metrics, and only 18 IP-SPs models constructed using LS, Bayesian and SVR performed less ideal with the value of at least one measuring metric worse than that of Glide SF, nevertheless, the overall performance of those 18 IP-SPs may still be better than Glide SF. We have also identified an ideal combination of ML-based IP-SF and structure processing protocol, i.e., GBDT with MIN+GB, which achieved the best overall performance for all the six drug targets. Compared to Glide SF, on average the MAE and RMSE decreased about 38% and 36%, respectively; while CORR2 and PI increased about 225% and 73%. As to AUC, the mean value increased from 0.71 for Glide SF to 0.87 for the best IP-SF. Therefore, for a given drug target, we recommend to first construct predictive model using the GBDT method after the docking complexes are minimized using the implicit GB model. In conclusion, we have successfully developed a novel method to construct the predictive model with high scoring, ranking and screening power using ligand-residue interaction profiles by machine learning methods. We expect this approach to have broad and promising applications in SBVSs.

Key Points

- To improve the performance of traditional scoring functions (SFs), we developed a novel approach by incorporating ligand-residue interaction profile (IP) into machine learning (ML) algorithms to construct SFs. A SF developed using IP is therefore called an IP-SF.
- Several evaluation tests were conducted to assess the IP-SPs for six protein targets, and the newly developed SFs significantly outperformed Glide SF in terms of scoring power, ranking power and screening power.
- An ideal computational protocol of preparing protein-ligand structures for IP calculations was developed. The performance of the IP-SF generated by employing GBDT algorithm in conjunction with the MIN+GB simulation protocol achieved the best overall performance for six drug targets.
- Besides high scoring, ranking and screening powers, the calculating cost of the IP-SF is minimal compared to more rigorous methods including the end-point MM-PBSA and the path-based methods like free energy perturbation and thermodynamic integration. Thus, the developed IP-SFs are expected to have broad and promising applications in SBVSs.

Supplementary Data

The descriptions about targets and compounds, feature representations, the processes of model construction and values of performance metrics for each model were listed in the supplementary data. **Table S1** describes entry codes, resolutions, released dates and deposition authors for six receptors. **Table S2** shows the correspondence between the AMBER-generated and the original residue IDs in PDB for six receptors. **Tables S3-S6** show the input feature representations for each target using four simulation protocols. **Table S7** is the description and hyperparameters of ML algorithms. **Table S8** shows the architectures of the DNN models for each system using MIN+GB

protocol. **Table S9** lists RMSE, MAE, CORR2 and PI values for IP-SFs versus Glide SF for each target using different ML algorithms and simulation protocols. **Table S10** lists the residue topology and force field parameters for the Zn^{2+} complex in 5HT2A receptor. **Figure S1** shows the structures and binding sites of six targets. **Figures S2-S5** show the distribution of key residues in six targets using four simulation protocols. **Figure S6** demonstrates an example code of ML training process for 5HT2AR target using MD+GB protocol. **Figures S7-S8** display the losses and MAE changes for each system during processes.

Acknowledgements

This work was supported by the following funds from the National Science Foundation (NSF) and National Institutes of Health (NIH): NSF 1955260, NIH R01GM079383 and NIH P30DA035778. The authors also thank the computing resources provided by the Center for Research Computing (CRC) at University of Pittsburgh.

Conflict of Interest

There are no conflicts to declare.

References

1. Jorgensen WL. Efficient Drug Lead Discovery and Optimization. *Acc Chem Res* 2009;**42**:724-733.
2. Sliwoski G, Kothiwale S, Meiler J et al. Computational methods in drug discovery. *Pharmacol rev* 2013;**66**:334-395.
3. Rifaioglu AS, Atas H, Martin MJ et al. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Briefings Bioinf* 2019;**20**:1878-1912.

4. da Silva Rocha SFL, Olanda CG, Fokoue HH et al. Virtual Screening Techniques in Drug Discovery: Review and Recent Applications. *Curr Top Med Chem* 2019;**19**:1751-1767.

5. Wang Z, Sun H, Shen C et al. Combined strategies in structure-based virtual screening. *Phys Chem Chem Phys* 2020;**22**:3149-3159.

6. Deng W, Breneman C, Embrechts MJ. Predicting protein-ligand binding affinities using novel geometrical descriptors and machine-learning methods. *J Chem Inf Comput Sci* 2004;**44**:699-703.

7. Zhang S, Golbraikh A, Tropsha A. Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces. *J Med Chem* 2006;**49**:2713-2724.

8. Artemenko N. Distance dependent scoring function for describing protein-ligand intermolecular interactions. *J Chem Inf Model* 2008;**48**:569-574.

9. Cheng T, Li X, Li Y et al. Comparative assessment of scoring functions on a diverse test set. *J Chem Inf Model* 2009;**49**:1079-1093.

10. Sotriffer CA, Sanschagrin P, Matter H et al. SFCscore: scoring functions for affinity prediction of protein-ligand complexes. *Proteins* 2008;**73**:395-419.

11. Durrant JD, McCammon JA. NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein−Ligand Complexes. *J Chem Inf Model* 2010;**50**:1865-1871.

12. Das S, Krein MP, Breneman CM. Binding affinity prediction with property-encoded shape distribution signatures. *J Chem Inf Model* 2010;**50**:298-308.

13. Ouyang X, Handoko SD, Kwoh CK. CScore: a simple yet effective scoring function for protein-ligand binding affinity prediction using modified CMAC learning architecture. *J Bioinf Comput Biol* 2011;**9 Suppl 1**:1-14.

14. Ballester PJ, Mitchell JBO. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinf (Oxford, England)* 2010;**26**:1169-1175.

15. Wang DD, Zhu M, Yan H. Computationally predicting binding affinity in protein–ligand complexes: free energy-based simulations and machine learning-based scoring functions. *Briefings Bioinf* 2020.

16. Li L, Wang B, Meroueh SO. Support Vector Regression Scoring of Receptor–Ligand Complexes for Rank-Ordering and Virtual Screening of Chemical Libraries. *Journal of Chem Inf Model* 2011;**51**:2132-2138.

17. Liu X, Zhu F, Ma X et al. The Therapeutic Target Database: an internet resource for the primary targets of approved, clinical trial and experimental drugs. *Expert Opin Ther Targets* 2011;**15**:903-912.

18. Zilian D, Sotriffer CA. SFCscoreRF: A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein–Ligand Complexes. *J Chem Inf Model* 2013;**53**:1923-1933.

19. Liu Q, Kwoh CK, Li J. Binding affinity prediction for protein-ligand complexes based on β contacts and B factor. *J Chem Inf Model* 2013;**53**:3076-3085.

20. Khamis MA, Gomaa W. Comparative assessment of machine-learning scoring functions on PDBbind 2013. *Engineering Applications of Artificial Intelligence* 2015;**45**:136-151.

21. Li H, Leung KS, Wong MH et al. Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets. *Mol Inform* 2015;**34**:115-126.

22. Li GB, Yang LL, Wang WJ et al. 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.

23. Kinnings SL, Liu N, Tonge PJ et al. A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *J Chem Inf Model* 2011;**51**:408-419.

24. Durrant JD, McCammon JA. NNScore 2.0: A Neural-Network Receptor–Ligand Scoring Function. *J Chem Inf Model* 2011;**51**:2897-2903.

25. Li L, Khanna M, Jo I et al. Target-specific support vector machine scoring in structure-based virtual screening: computational validation, in vitro testing in kinases, and effects on lung cancer cell proliferation. *J Chem Inf Model* 2011;**51**:755-759.

26. Ding B, Wang J, Li N et al. Characterization of small molecule binding. I. Accurate identification of strong inhibitors in virtual screening. *J Chem Inf Model* 2013;**53**:114-122.

27. Guedes IA, Pereira FSS, Dardenne LE. Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Frontiers in pharmacology* 2018;**9**:1089-1089.

28. Leach AR, Shoichet BK, Peishoff CE. Prediction of Protein−Ligand Interactions. Docking and Scoring: Successes and Gaps. *J Med Chem* 2006;**49**:5851-5855.

29. Li Y, Su M, Liu Z et al. Assessing protein–ligand interaction scoring functions with the CASF-2013 benchmark. *Nature Protoc* 2018;**13**:666-680.

30. Liu J, Wang R. Classification of current scoring functions. *J Chem Inf Model* 2015;**55**:475-482.

31. Kitchen DB, Decornez H, Furr JR et al. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* 2004;**3**:935-949.

32. Mooij WT, Verdonk ML. General and targeted statistical potentials for protein-ligand interactions. *Proteins* 2005;**61**:272-287.

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

34. Ain QU, Aleksandrova A, Roessler FD et al. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip Rev Comput Mol Sci* 2015;**5**:405-424.

35. Li H, Peng J, Sidorov P et al. Classical scoring functions for docking are unable to exploit large volumes of structural and interaction data. *Bioinf* 2019;**35**:3989-3995.

36. Wang L, Wu Y, Deng Y et al. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J Am Chem Soc* 2015;**137**:2695-2703.

37. He X, Liu S, Lee T-S et al. Fast, Accurate, and Reliable Protocols for Routine Calculations of Protein–Ligand Binding Affinities in Drug Design Projects Using AMBER GPU-TI with ff14SB/GAFF. *ACS Omega* 2020;**5**:4611-4619.

38. Wang E, Sun H, Wang J et al. End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chem Rev* 2019;**119**:9478-9508.

39. He X, Man VH, Ji B et al. Calculate protein-ligand binding affinities with the extended linear interaction energy method: application on the Cathepsin S set in the D3R Grand Challenge 3. *J Comput Aided Mol Des* 2019;**33**:105-117.

40. Cortes C, Vapnik V. Support-vector networks. *Machine Learning* 1995;**20**:273-297.

41. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Statist* 2001;**29**:1189-1232.

42. Rätsch G, Onoda T, Müller KR. Soft Margins for AdaBoost. *Machine Learning* 2001;**42**:287-320.

43. Breiman L. Random Forests. *Machine Learning* 2001;**45**:5-32.

44. Jain AK, Jianchang M, Mohiuddin KM. Artificial neural networks: a tutorial. *Computer* 1996;**29**:31-44.

45. Xin Y. Evolving artificial neural networks. *Proceedings of the IEEE* 1999;**87**:1423-1447.

46. Ashtawy HM, Mahapatra NR. A Comparative Assessment of Predictive Accuracies of Conventional and Machine Learning Scoring Functions for Protein-Ligand Binding Affinity Prediction. *IEEE/ACM Tran Comput Biol Bioinf* 2015;**12**:335-347.

47. Li H, Leung K-S, Wong M-H et al. Substituting random forest for multiple linear regression improves binding affinity prediction of scoring functions: Cyscore as a case study. *BMC Bioinf* 2014;**15**:291.

48. Wallach I, Dzamba M, Heifets A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery 2015.

49. Ballester PJ, Schreyer A, Blundell TL. 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.

50. Sun H, Pan P, Tian S et al. Constructing and Validating High-Performance MIEC-SVM Models in Virtual Screening for Kinases: A Better Way for Actives Discovery. *Sci Rep* 2016;**6**:24817.

51. Yan Y, Wang W, Sun Z et al. Protein–Ligand Empirical Interaction Components for Virtual Screening. *J Chem Inf Model* 2017;**57**:1793-1806.

52. Nguyen H, Roe DR, Simmerling C. Improved Generalized Born Solvent Model Parameters for Protein Simulations. *J Chem Theory Comput* 2013;**9**:2020-2034.

53. Mark P, Nilsson L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J Phys Chem A* 2001;**105**:9954-9960.

54. Gaulton A, Hersey A, Nowotka M et al. The ChEMBL database in 2017. *Nucleic Acids Res* 2017;**45**:D945-d954.

55. Davies M, Nowotka M, Papadatos G et al. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic acids research* 2015;**43**:W612-W620.

56. Berman HM, Westbrook J, Feng Z et al. The Protein Data Bank. *Nucleic Acids Res* 2000;**28**:235-242.

57. Burley SK, Berman HM, Bhikadiya C et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res* 2019;**47**:D464-d474.

58. Sanner MF. Python: a programming language for software integration and development. *J Mol Graph Model* 1999;**17**:57-61.

59. Friesner RA, Banks JL, Murphy RB et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J Med Chem* 2004;**47**:1739-1749.

60. Halgren TA, Murphy RB, Friesner RA et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J Med Chem* 2004;**47**:1750-1759.

61. Jakalian A, Bush BL, Jack DB et al. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method 2000;**21**:132-146.

62. Jakalian A, Jack DB, Bayly CI. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation 2002;**23**:1623-1641.

63. Case DA, Betz RM, Cerutti DS et al. AMBER 2016. University of California, San Francisco, 2016.

64. Bayly CI, Cieplak P, Cornell W et al. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J Phys Chem* 1993;**97**:10269-10280.

65. Wang J, Wolf RM, Caldwell JW et al. Development and testing of a general amber force field. *J Comput Chem* 2004;**25**:1157-1174.

66. Maier JA, Martinez C, Kasavajhala K et al. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* 2015;**11**:3696-3713.

67. Wang J, Wang W, Kollman PA et al. Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graphics Model* 2006;**25**:247-260.

68. Case DA, Cheatham TE, 3rd, Darden T et al. The Amber biomolecular simulation programs. *J Comput Chem* 2005;**26**:1668-1688.

69. Salomon-Ferrer R, Case DA, Walker RC. An overview of the Amber biomolecular simulation package 2013;**3**:198-210.

70. Darden T, York D, Pedersen L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems 1993;**98**:10089-10092.

71. Hawkins GD, Cramer CJ, Truhlar DG. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J Phys Chem* 1996;**100**:19824-19839.

72. Götz AW, Williamson MJ, Xu D et al. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J Chem Theory Comput* 2012;**8**:1542-1555.

73. Salomon-Ferrer R, Götz AW, Poole D et al. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J Chem Theory Comput* 2013;**9**:3878-3888.

74. Scott AJ, Holt D. The Effect of Two-Stage Sampling on Ordinary Least Squares Methods. *J Am Stat Assoc* 1982;**77**:848-854.

75. Zellner A. Bayesian and Non-Bayesian Analysis of the Regression Model with Multivariate Student-t Error Terms. *J Am Stat Assoc* 1976;**71**:400-405.

76. Noble WS. What is a support vector machine? *Nat Biotechnol* 2006;**24**:1565-1567.

77. Dawson CW, Wilby R. An artificial neural network approach to rainfall-runoff modelling. *Hydrol Sci J* 1998;**43**:47-66.

78. Pedregosa F, Varoquaux G, Gramfort A et al. Scikit-learn: Machine learning in Python 2011;**12**:2825-2830.

79. Chollet F. Keras 2015. GitHub. https://github.com/fchollet/keras

80. Liu W, Wang Z, Liu X et al. A survey of deep neural network architectures and their applications. *Neurocomputing* 2017;**234**:11-26.

81. Pearlman DA, Charifson PS. Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 MAP kinase protein system. *J Med Chem* 2001;**44**:3417-3423.

82. Luccarelli J, Michel J, Tirado-Rives J et al. Effects of Water Placement on Predictions of Binding Affinities for p38α MAP Kinase Inhibitors. *J Chem Theory Comput* 2010;**6**:3850-3856.

83. Michel J, Verdonk ML, Essex JW. Protein-Ligand Binding Affinity Predictions by Implicit Solvent Simulations: A Tool for Lead Optimization? *J Med Chem* 2006;**49**:7427-7439.

84. Jain AN, Nicholls A. Recommendations for evaluation of computational methods. *J Comput Aided Mol Des* 2008;**22**:133-139.

85. Li H, Zhang H, Zheng M et al. An effective docking strategy for virtual screening based on multi-objective optimization algorithm. *BMC Bioinf* 2009;**10**:58.

86. Venkatraman V, Pérez-Nueno VI, Mavridis L et al. Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Data set Reveals Limitations of Current 3D Methods. *J Chem Inf Model* 2010;**50**:2079-2093.

87. Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Briefings Bioinf* 2017;**18**:851-869.

88. Aliper A, Plis S, Artemov A et al. Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. *Mol Pharmaceutics* 2016;**13**:2524-2530.

89. Zeng X, Zhu S, Liu X et al. deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinf* 2019;**35**:5191-5198.

90. Senior AW, Evans R, Jumper J et al. Improved protein structure prediction using potentials from deep learning. *Nature* 2020;**577**:706-710.

Tables

Table 1. The number of actives and decoys and number of compounds in the training and test sets.

Receptor	Actives	Decoys	GB		TIP3P	
			Training	Test	Training	Test
5HT2AR	377	395	617	155	570	143
A2AR	342	547	711	178	700	176
CB1	265	383	518	130	501	126
M1R	217	364	464	117	438	110
ERK2	758	92	680	170	664	167
VEGFR2	287	343	504	126	500	126

Table 2. Numbers of Key Residues Participating Modeling with Machine Learning Methods

Receptor	Total Residues	No. of Key Residues			
		MD + GB	MD + TIP3P	MIN + GB	MIN + TIP3P
5HT2AR	370	90	92	90	91
A2AR	448	84	84	84	83
CB1	281	86	90	90	93
M1R	451	80	83	81	83
ERK2	337	82	83	83	87
VEGFR2	299	85	86	88	92

Table 3. The performance of IP-SFs in docking screenings. Each drug target has 28 IP-SFs models constructed using 7 different machine learning algorithms with four structure processing protocols. "Best" represents the best value achieved by 28 IP-SFs for a specific metric and a given drug target; "GBDT" and "DNN" represent the IP-SFs constructed using GBDT and DNN methods in conjunction with the MIN+GB structure processing protocol.

Metrics/Receptor		5HT2AR	A2AR	CB1	M1R	ERK2	VEGFR2	Average
MAE (kcal/mol)	Glide	2.57	1.83	1.77	1.52	1.24	2.11	1.84
	Best	1.35	1.37	1.13	1.30	0.49	1.00	1.11
	GBDT	1.35	1.37	1.24	1.34	0.54	1.00	1.14
	DNN	1.58	1.55	1.59	1.53	0.77	1.21	1.37
RMSE (kcal/mol)	Glide	3.01	2.25	2.17	1.95	1.61	2.65	2.27
	Best	1.68	1.70	1.41	1.63	0.73	1.26	1.40
	GBDT	1.75	1.70	1.53	1.65	0.81	1.26	1.45
	DNN	1.99	1.88	1.91	1.91	1.03	1.51	1.71
CORR2	Glide	0.06	0.10	0.05	0.26	0.27	0.22	0.16
	Best	0.29	0.36	0.44	0.55	0.81	0.72	0.53
	GBDT	0.29	0.36	0.44	0.54	0.74	0.72	0.52
	DNN	0.12	0.27	0.14	0.42	0.60	0.58	0.36
PI	Glide	0.22	0.32	0.26	0.45	0.69	0.50	0.41
	Best	0.55	0.64	0.65	0.75	0.88	0.86	0.72
	GBDT	0.55	0.58	0.65	0.74	0.85	0.86	0.71
	DNN	0.33	0.52	0.33	0.66	0.78	0.77	0.57

Table 4. Comparison of AUC, Accuracy, F1-score, $EF_{10\%}$ and $EF_{40\%}$ between Glide docking and IP-SF with the ideal combination of ML method and simulation protocol (GBDT with MIN+GB) on six targets.

Metrics/Receptor		5HT2AR	A2AR	CB1	M1R	ERK2	VEGFR2
Glide	AUC	0.60	0.68	0.55	0.68	0.97	0.76
	Accuracy	0.51	0.65	0.50	0.61	0.90	0.68
	F1-score	0.53	0.60	0.44	0.53	0.65	0.69
	$EF_{10\%}$	1.27	1.70	1.20	2.72	7.22	1.91
	$EF_{40\%}$	1.10	1.54	1.15	1.45	2.50	1.41
GBDT	AUC	0.81	0.79	0.86	0.86	0.98	0.94
	Accuracy	0.72	0.72	0.78	0.77	0.93	0.84
	F1-score	0.73	0.68	0.74	0.70	0.74	0.85
	$EF_{10\%}$	1.85	2.54	2.60	2.72	9.44	1.91
	$EF_{40\%}$	1.85	2.51	2.50	2.49	2.50	1.91

Figures

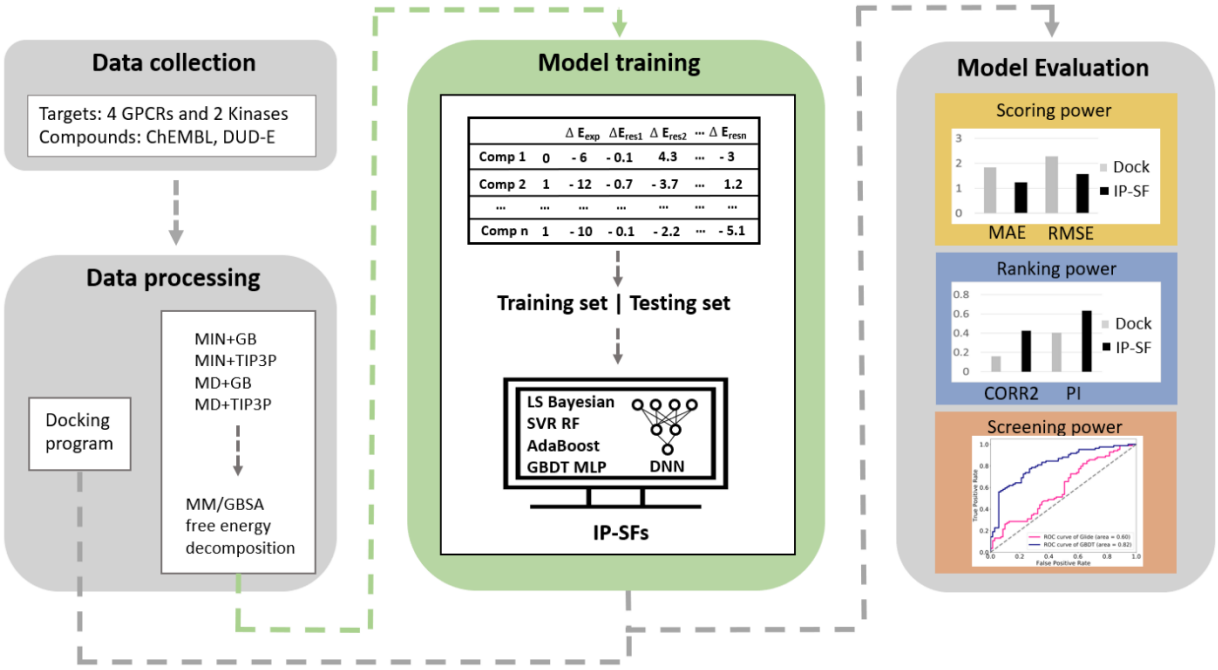

Figure 1. The workflow of the IP-SFs model construction and model evaluation.

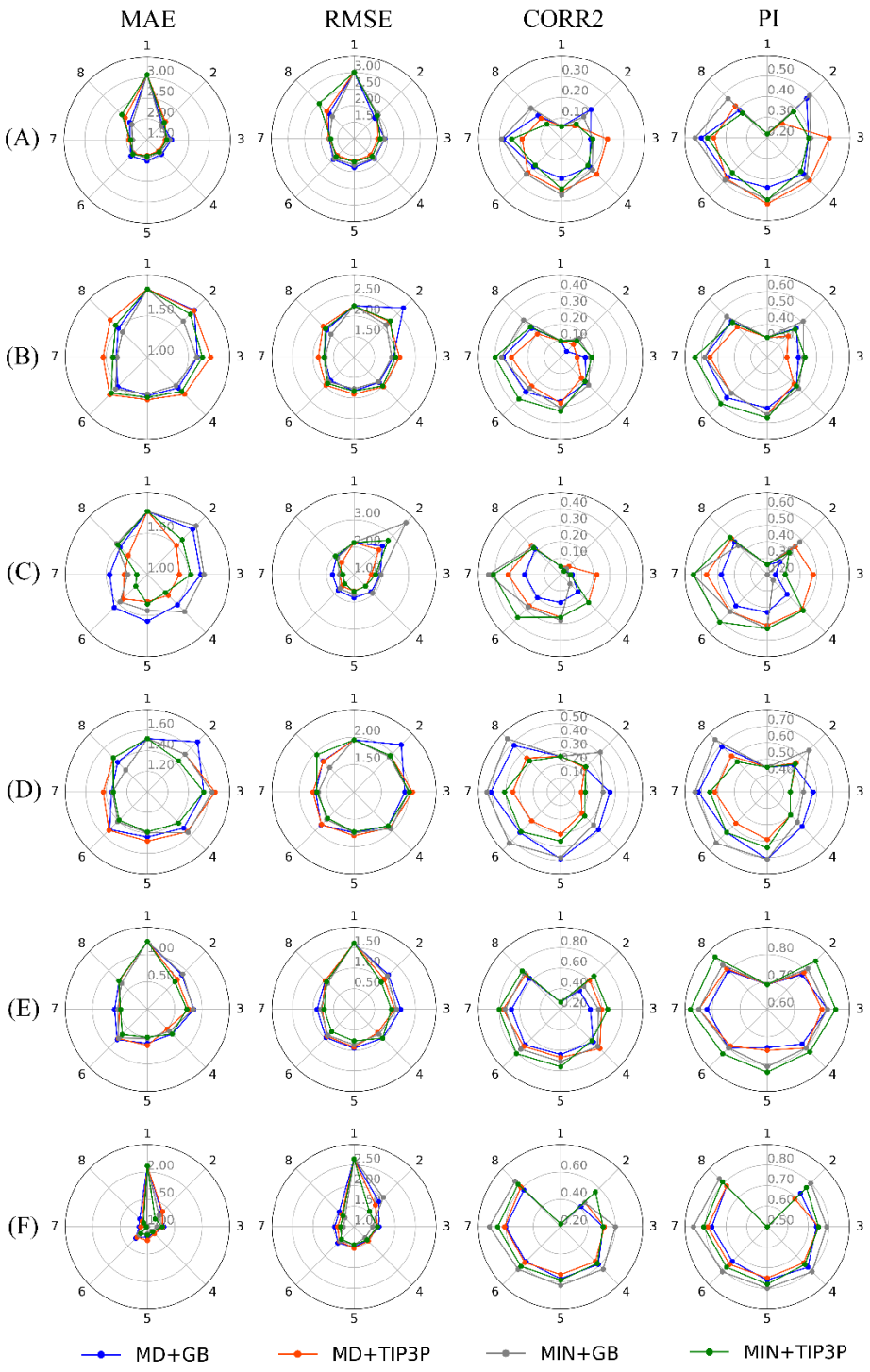

Figure 2. The performance of ML-based SFs versus Glide docking based on different simulation protocols. (A) 5HT2AR; (B) A2AR; (C) CB1; (D) M1R; (E) ERK2; (F) VEGFR2; 1-Glide; 2-LS; 3-Bayesian; 4-SVR; 5-RF; 6-AdaBoost; 7-GBDT; 8-MLP.

(A) MD+GB		5HT2AR	A2AR	CB1	M1R	ERK2	VEGFR2
LS	MAE	4	7	7	7	7	7
	RMSE	4	7	7	7	7	7
	CORR2	2	7	7	7	7	7
	PI	2	5	6	7	5	7
Bayesian	MAE	6	6	6	6	6	5
	RMSE	6	6	6	6	6	6
	CORR2	7	6	6	6	6	6
	PI	6	7	7	6	1	5
SVR	MAE	2	5	3	4	3	2
	RMSE	2	4	4	5	4	1
	CORR2	3	5	5	5	3	2
	PI	4	5	5	5	5	1
RF	MAE	3	2	4	3	2	2
	RMSE	3	3	3	2	2	3
	CORR2	3	3	4	2	4	3
	PI	5	3	4	2	7	4
AdaBoost	MAE	5	4	4	5	5	5
	RMSE	5	2	2	4	3	5
	CORR2	3	2	3	4	1	5
	PI	2	2	3	4	3	5
GBDT	MAE	1	1	1	1	1	1
	RMSE	1	1	1	1	1	1
	CORR2	1	1	1	1	2	1
	PI	1	1	1	1	1	3
MLP	MAE	6	3	2	2	4	4
	RMSE	7	5	5	3	4	4
	CORR2	6	4	1	3	5	3
	PI	7	3	1	3	3	1

(B) MD+TIP3P		5HT2AR	A2AR	CB1	M1R	ERK2	VEGFR2
LS	MAE	6	7	7	4	7	7
	RMSE	6	7	7	6	7	7
	CORR2	7	6	7	5	6	7
	PI	7	6	7	5	5	7
Bayesian	MAE	3	6	5	7	6	6
	RMSE	3	6	6	7	5	6
	CORR2	4	7	6	7	6	6
	PI	2	7	6	7	3	4
SVR	MAE	1	3	4	6	1	3
	RMSE	2	4	2	5	2	3
	CORR2	2	5	5	6	2	4
	PI	3	5	5	6	3	4
RF	MAE	1	1	2	3	3	4
	RMSE	1	2	2	3	3	5
	CORR2	1	2	3	3	5	5
	PI	1	1	3	3	7	4
AdaBoost	MAE	5	5	6	5	5	5
	RMSE	3	3	4	4	4	4
	CORR2	3	3	2	4	3	3
	PI	4	3	2	4	5	3
GBDT	MAE	3	2	1	1	1	1
	RMSE	5	1	1	1	1	1
	CORR2	5	1	1	1	1	1
	PI	5	1	1	1	1	1
MLP	MAE	7	3	2	2	4	2
	RMSE	7	5	5	2	5	2
	CORR2	6	4	4	1	4	1
	PI	6	4	3	2	2	2

(C) MIN+GB		5HT2AR	A2AR	CB1	M1R	ERK2	VEGFR2
LS	MAE	4	7	7	5	7	7
	RMSE	7	7	7	5	7	7
	CORR2	6	7	6	5	7	7
	PI	3	4	4	5	4	5
Bayesian	MAE	7	6	6	7	6	5
	RMSE	5	6	6	7	6	5
	CORR2	6	6	6	7	6	6
	PI	7	7	7	7	3	7
SVR	MAE	2	4	5	6	3	3
	RMSE	3	4	5	6	2	3
	CORR2	4	5	5	6	3	3
	PI	5	6	6	6	6	3
RF	MAE	3	3	2	3	1	4
	RMSE	2	2	2	4	2	4
	CORR2	2	3	2	4	3	5
	PI	2	2	2	4	4	5
AdaBoost	MAE	4	5	3	4	5	6
	RMSE	4	5	2	3	4	6
	CORR2	3	4	2	3	1	3
	PI	3	4	3	2	6	3
GBDT	MAE	1	1	1	2	2	1
	RMSE	1	1	1	2	1	1
	CORR2	1	1	1	2	2	1
	PI	1	1	1	2	1	1
MLP	MAE	6	2	4	1	4	2
	RMSE	6	3	4	1	5	2
	CORR2	4	2	4	1	5	2
	PI	5	2	5	1	2	2

(D) MIN+TIP3P		5HT2AR	A2AR	CB1	M1R	ERK2	VEGFR2
LS	MAE	6	7	7	4	5	6
	RMSE	6	7	7	5	5	6
	CORR2	6	7	7	5	5	6
	PI	6	6	6	5	3	5
Bayesian	MAE	4	6	6	7	6	7
	RMSE	6	6	5	7	4	7
	CORR2	5	6	6	7	5	7
	PI	5	7	7	7	3	7
SVR	MAE	3	4	3	4	3	3
	RMSE	3	5	3	4	7	5
	CORR2	3	5	4	6	7	5
	PI	4	5	5	6	7	6
RF	MAE	1	2	4	2	2	4
	RMSE	1	2	4	3	2	4
	CORR2	1	3	3	3	3	4
	PI	1	3	3	3	5	3
AdaBoost	MAE	5	5	2	2	4	5
	RMSE	4	3	2	1	2	3
	CORR2	3	2	2	1	1	3
	PI	3	2	2	1	5	3
GBDT	MAE	2	1	1	1	1	1
	RMSE	2	1	1	1	1	1
	CORR2	1	1	1	1	2	1
	PI	2	1	1	1	1	1
MLP	MAE	7	3	5	6	7	2
	RMSE	7	4	6	6	6	2
	CORR2	6	4	5	4	4	2
	PI	7	4	4	4	2	1

Figure 3. Ranks of seven ML approaches for all receptors with four simulation protocols: (A) MD+GB, (B) MD+TIP3P, (C) MN+GB, (D) MIN+GB.

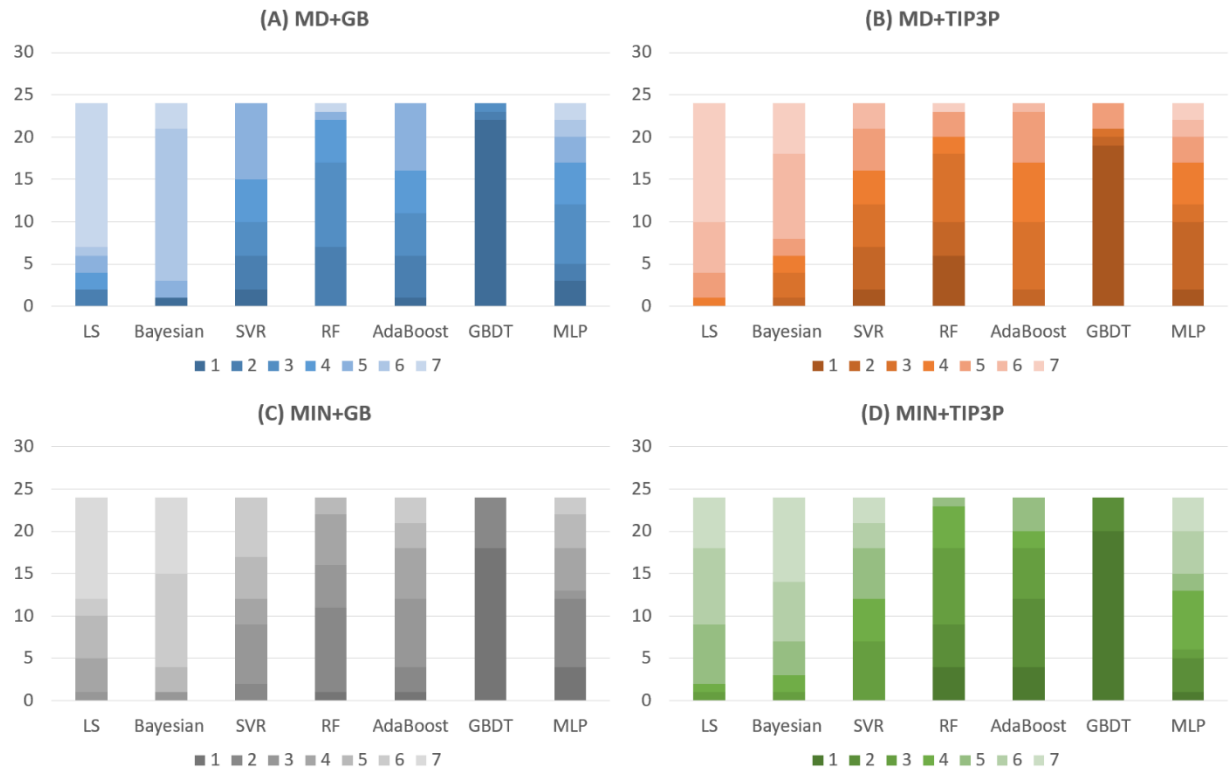

Figure 4. The frequency distribution histogram on ranks for all ML algorithms for all drug targets. (A) MD+GB, (B) MD+TIP3P, (C) MN+GB, (D) MIN+GB.

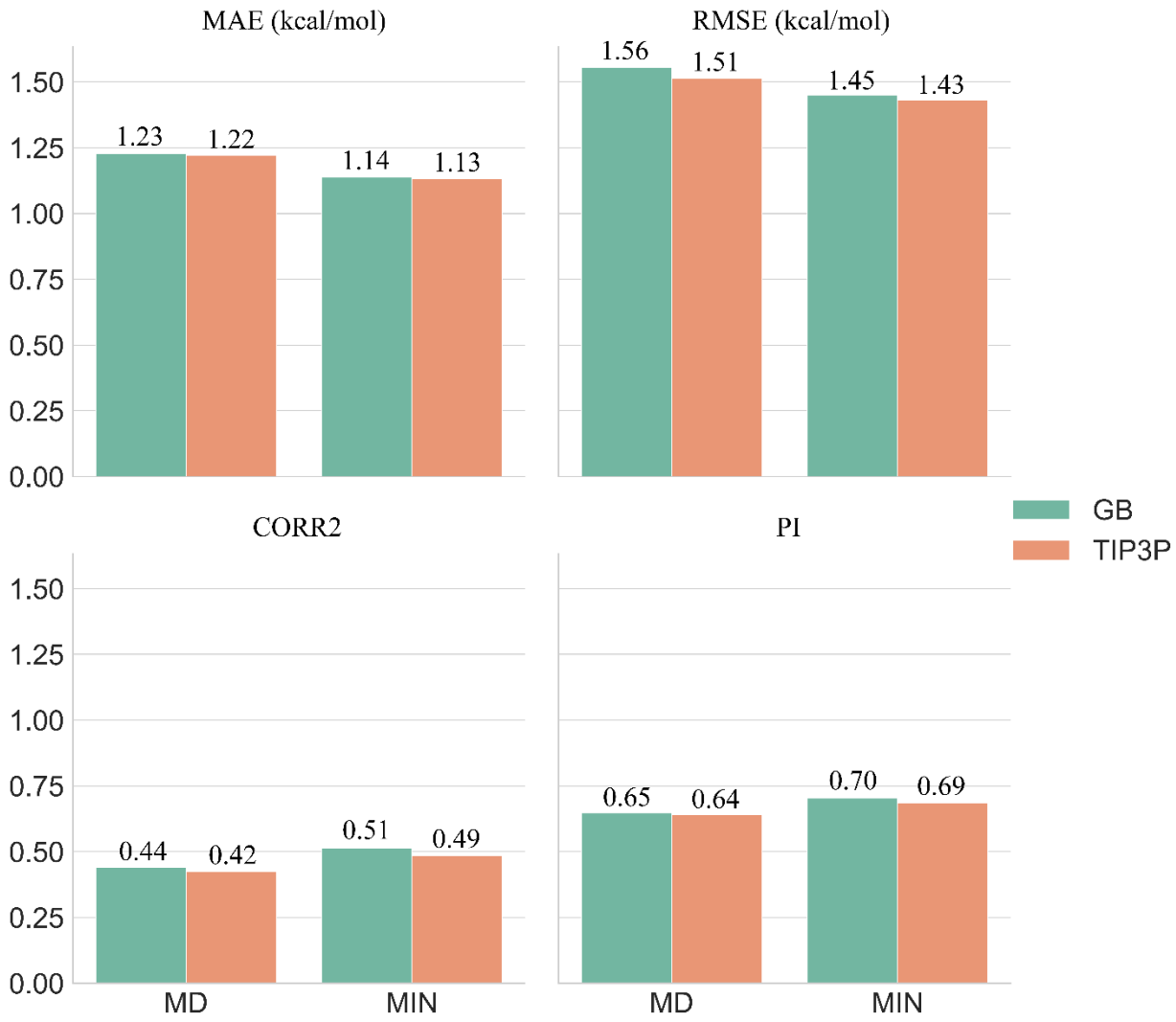

Figure 5. The comparison of the scoring power of GBDT-based SFs following four simulation protocols, MD+GB, MD+TIP3P, MIN+GB, MIN+TIP3P by using the average performance of six targets.

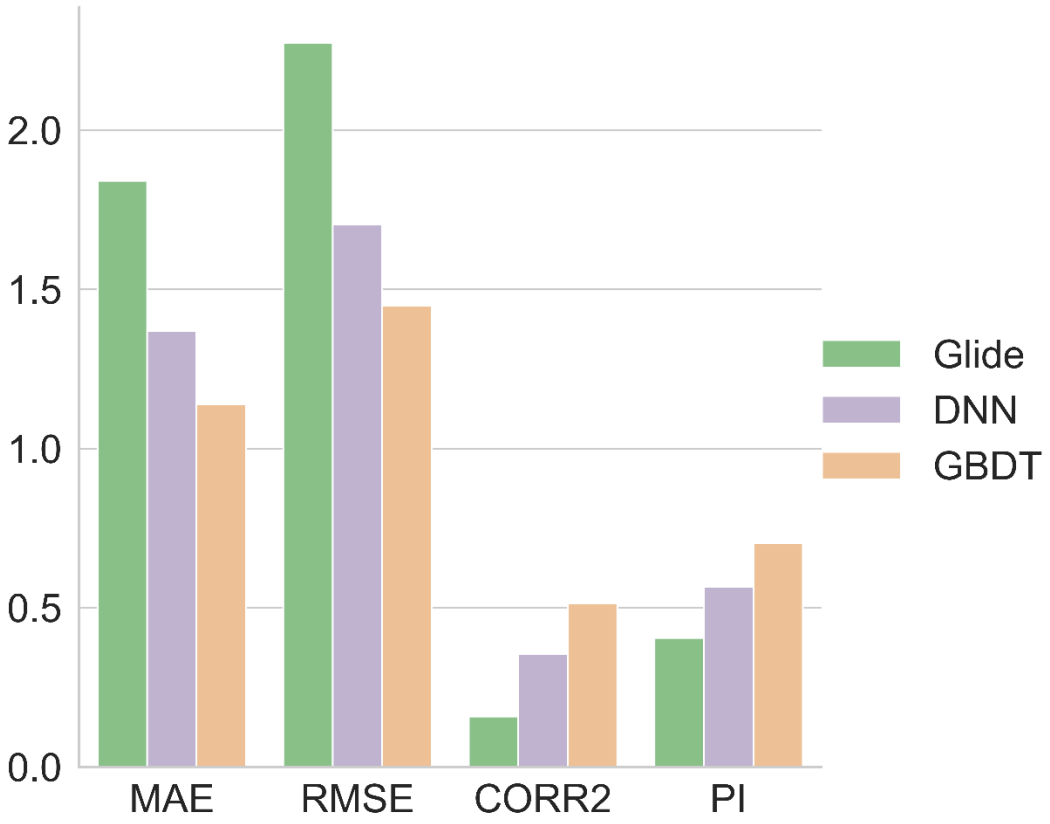

Figure 6. The comparison of performance of SFs by using Glide docking, DNN, and GBDT methods.

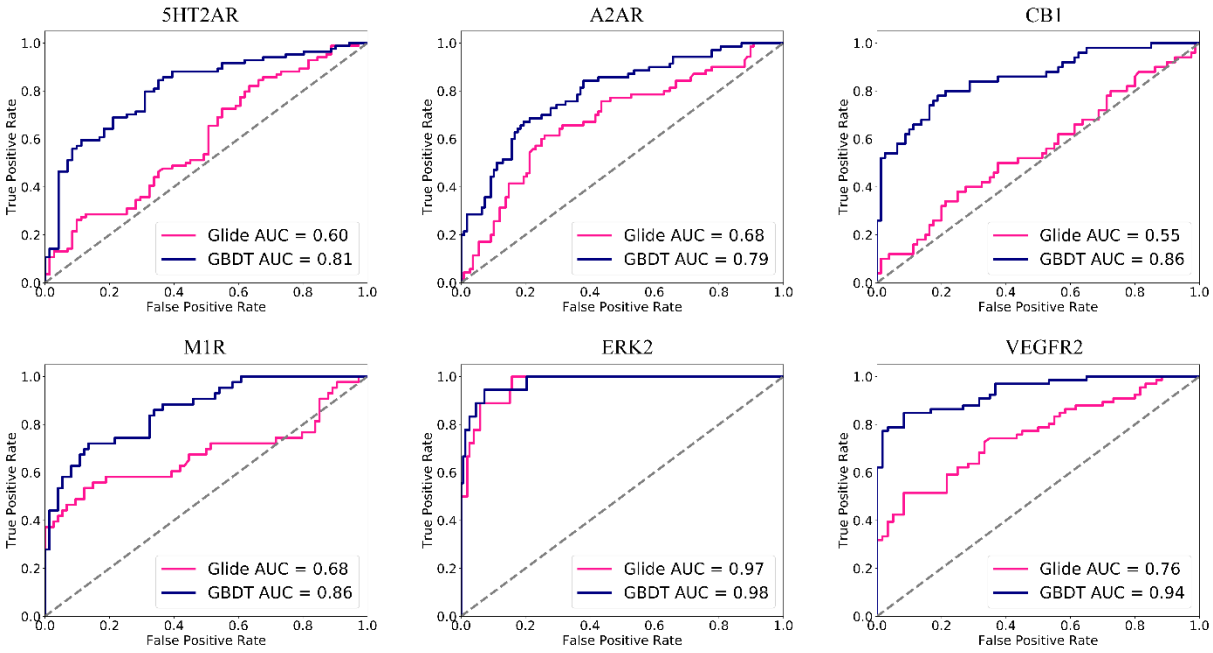

Figure 7. ROC curves of using Glide docking and GBDT method in conjunction with the MIN+GB structure processing protocol on six targets.