HHS Public Access

Author manuscript

J Chem Inf Model. Author manuscript; available in PMC 2023 June 13.

Published in final edited form as:

J Chem Inf Model. 2022 June 13; 62(11): 2696–2712. doi:10.1021/acs.jcim.2c00485.

Delta Machine Learning to Improve Scoring-Ranking-Screening Performances of Protein-Ligand Scoring Functions

Chao Yang[1], **Yingkai Zhang**[1,2]

[1]Department of Chemistry, New York University, New York, NY 10003, United States

[2]NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai 200062, China

Abstract

Protein-ligand scoring functions are widely used in structure-based drug design for fast evaluation of protein-ligand interactions, and there is of strong interest to develop scoring functions with machine learning approaches. In this work, by expanding the training set, developing physically meaningful features, employing our recently developed linear empirical scoring function Lin_F9 (*J. Chem. Inf. Model.* **2021**, *61*, 4630 – 4644) as the baseline, and applying extreme gradient boosting (XGBoost) with Δ-machine learning, we have further improved robustness and applicability of machine-learning scoring functions. Besides the top performances for scoring-ranking-screening power tests of CASF-2016 benchmark, the new scoring function Δ_{Lin_F9}XGB also achieves superior scoring and ranking performances in different structure types that mimic real docking applications. The scoring power of Δ_{Lin_F9}XGB for locally optimized poses, flexible re-docked poses and ensemble docked poses of CASF-2016 core set achieve Pearson's correlation coefficient (R) of 0.853, 0.839 and 0.813, respectively. In addition, large-scale docking-based virtual screening test on LIT-PCBA dataset demonstrates the reliability and robustness of Δ_{Lin_F9}XGB in virtual screening application. The Δ_{Lin_F9}XGB scoring function and its code are freely available on the web at: (https://yzhang.hpc.nyu.edu/Delta_LinF9_XGB)

Graphical Abstract

Corresponding Author: yingkai.zhang@nyu.edu.

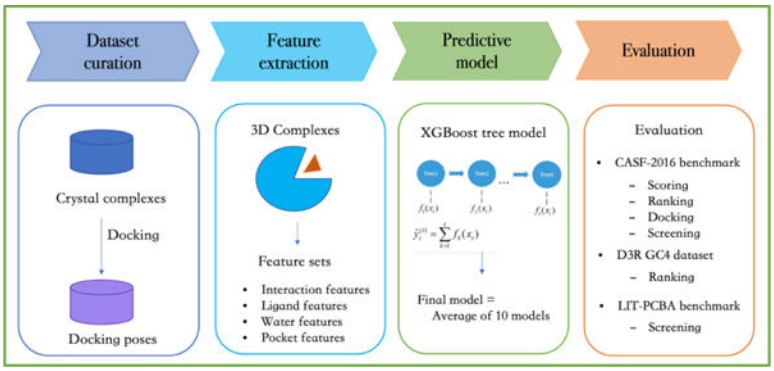
ASSOCIATED CONTENT

Supporting Information

This material is available free of charge via the Internet at http://pubs.acs.org.

Tables S1-S5 (Training and Validation sets, Feature set, Performances of Training and Validation sets, Screening performances on CASF-2016 benchmark, AUC performances on LIT-PCBA benchmark), Figures S1-S9 (Detail of Vina features, Docking performances of extended docking-scoring tests, Pose prediction protocol for BACE1 dataset, Targets with more than one binding site in LIT-PCBA, ROC curves for LIT-PCBA benchmark) (PDF)

The authors declare no competing financial interest.



INTRODUCTION

Molecular docking is one of widely utilized computational tools for structure-based drug discovery that attempts to predict the ligand binding pose and provide an estimate of binding affinity for protein-ligand complex.[1–3] The most critical component of docking is the scoring function, and a robust scoring function should perform well across a variety of applications.[4–11] In recent years, a variety of machine-learning (ML) scoring functions[12–25] have been developed and outperformed classical scoring functions at retrospective benchmarks, and some scoring functions also performed well on prospective structure-based virtual screening (SBVS) application, such as AtomNet[26] and vScreenML.[27, 28] Wallach et al introduced AtomNet, the first CNN-based scoring function incorporating 3D structural information, and they applied AtomNet in several VS campaigns.[26, 29–31] Adeshina et al proposed the vScreenML and used it for prospective SBVS against human acetylcholinesterase (AChE).[27] Besides the generic scoring functions, target-specific ML scoring functions have been developed to focus on certain protein target or family,[20, 21, 32, 33] which can outperform other models on that particular target case. However, the target specific scoring function approach will not be applicable for a novel target with little experimental data available. Thus, it is of significant interest to develop robust ML scoring functions.

Several key metrics[5–9] have been developed to assess performance of scoring function for different tasks, including: (I) scoring, which assesses the linear correlation between predicted and experimental measured binding affinities; (II) ranking, which evaluates the ranking ability of a scoring function to rank the known ligands for a certain target protein by predicted binding affinities; (III) docking, which evaluates the ability of a scoring function to identify native binding site and binding mode of ligand among computer generated decoys; (IV) screening, which assesses the ability to identify true binders for a given target from random molecule libraries. Extensive retrospective and comparative studies[1, 5–9, 34–42] demonstrate that some widely used classical scoring functions, such as GlideScore[43, 44] and Autodock Vina[45], perform relatively well in docking and screening tasks, but their scoring power are less satisfactory. Many ML scoring functions have achieved significantly better scoring power on crystal structures.[17–19, 25, 33, 46–63] However, more extensive evaluations indicate that this enhancement in scoring performance accompany with significant under-performance in docking and screening power tests compared to classical

scoring functions.[64, 65] It remains a challenge for the scoring function development to not only improve scoring power, but also perform well for docking and screening tasks.

To tackle the challenge, recently we have employed the Δ-machine learning approach,[66, 67] in which a correction term to the Vina scoring function is parametrized with machine learning, to develop two successive scoring functions, $\Delta_{vina}RF_{20}$[66] and $\Delta_{Vina}XGB$,[67] achieving top performances for all metrics of CASF-2016 benchmark compared to 33 classical scoring functions. On the other hand, based on a small high-quality training data, we have developed a linear empirical scoring function, Lin_F9, which achieves better scoring and ranking powers than Vina on different structure types, including crystal pose, local optimized pose and docked pose.[68] Lin_F9 has been successfully applied to virtual screening and rational design of SARS-Cov-2 main protease inhibitors.[69]

A major motivation for the current work is to use Lin_F9 as the new baseline scoring function and incorporate Δ-Learning machine learning approach to further enhance scoring and ranking performances on different structure types. Here, Δ-Lin_F9 machine learning strategy via eXreme Gradient Boosting (XGBoost)[70] have been explored. The training set is enlarged to include more experimental measured weak binders. One is crystal structures with weak binding affinities obtained from updated PDBbind database[71, 72]. The other is to use computer-generated decoys with weak binding affinities obtained from BindingDB database[73, 74]. In addition, in order to learn from docked poses, top 1 docked poses from end-to-end (E2E) docking are also included in the training set. More details of E2E docking protocol are described in Methods. For feature exploration, the previous used Vina 58 features in $\Delta_{Vina}XGB$ is replaced by a specialized Vina 48 features, in which polar-polar, polar-nonpolar, nonpolar-nonpolar, hydrogen bond and metal-ligand interactions in different distance ranges are described using a series of gauss functions. The overall feature set consists of 76 protein-ligand features and 16 ligand-specific features.

In this article, we described the development of a new state-of-the-art scoring function $\Delta_{Lin\_F9}XGB$. The overall evaluation indicates that $\Delta_{Lin\_F9}XGB$ can not only perform consistently among the top compared to classical scoring functions in CASF-2016 benchmark, but also achieve superior prediction accuracy on different structure types, including docked poses that mimic real docking applications. In addition, we evaluated the screening performance of $\Delta_{Lin\_F9}XGB$ on LIT-PCBA dataset,[75] which consists of 15 diverse target proteins with large-scale experiment-verified actives/decoys. The $\Delta_{Lin\_F9}XGB$ scoring function and its code are freely available on the web at: (https://yzhang.hpc.nyu.edu/Delta_LinF9_XGB).

METHODS

Data Preparation

Training Set—The main component of our training set is inherited from the previous $\Delta_{Vina}XGB$'s training set, which is based on PDBbind (v2016) database[71] and CSAR decoy set[76, 77]. The details of this component can refer to the paper.[67] We cleaned the data by removing 3 covalent ligands and by removing previous constructed PDBbind decoys. The

afterward cleaned data consists of 6816 PDBbind binders and 6321 CSAR decoys. In terms of binder set, 1556 weak binders ($pK_d < 6$) and 510 strong binders ($pK_d > 9$) were selected from PDBbind (v2018) general set[71] and added to our binder set. It should be noted that the binder set have both with water (receptor-bound waters) and without water structures. All these binders come from PDBbind database and meet the requirements: (I) should be K_d/K_i binding data; (II) local optimized pose with RMSD <=2 Å from crystal pose; (III) Noncovalent ligand. Furthermore, in order to add more strong binder poses, 235 docked poses obtained by flexible re-docking ($pK_d > 9$ and RMSD <= 1 Å from crystal pose) were selected and added to the binder set. As shown in Table S1, the overall binder subset of training set has 9117 complex structures with Lin_F9 local optimized ligand poses used to generate features, and the experimental measured binding affinities (pK_d(exp)) are the labels.

In terms of decoy set 1 (see Table S1), which serves as a negative control of binding pose and binding affinity in the whole training set, a total of 7111 structures with estimated binding affinities (pK_d(est)) are constructed using CSAR[76, 77] decoy set and BindingDB[73, 74] weak binders. For 6321 CSAR decoys inherited from Δ_{Vina}XGB training set[67], the (pK_d(est)) for each decoy is determined by comparing the RMSD between decoy and crystal pose, as well as by comparing the Lin_F9 predicted binding affinity (pK_d(Lin_F9)) and pK_d(exp) of crystal pose: if the RMSD is no larger than 1 Å, which means the decoy is similar as crystal pose, pK_d(est) is assigned as the pK_d(exp) of crystal pose; else, for RMSD larger than 1 Å, pK_d(Lin_F9) is calculated and compared with pK_d(exp): if the pK_d(Lin_F9) is less than the pK_d(exp), the pK_d(est) is assigned as pK_d(Lin_F9); otherwise, pK_d(est) is assigned as the maximum value between pK_d(exp) – 0.5×(RMSD-1) and 0.5×pK_d(exp). So, the pK_d(est) is smaller when RMSD is larger, but not smaller than half of pK_d(exp). In addition, 790 decoys were obtained from top 1 docked poses of very weak binders ($pK_d < 3$) in BindingDB using E2E flexible docking. As these weak binders do not have crystal protein-ligand structures, the above pK_d(est) protocol is not applicable. For these BindingDB decoys, the pK_d(est) is determined by only comparing the pK_d(Lin_F9) and pK_d(exp): if the pK_d(Lin_F9) is less than the pK_d(exp), the pK_d(est) is assigned as pK_d(Lin_F9), otherwise, pK_d(est) is assigned as pK_d(exp).

In addition, in order to learn from flexible docked poses of complexes, we construct a decoy set 2 (see Table S1), in which top 1 docked poses are generated from the above binder set using E2E docking protocol that combines ligand conformer generation and flexible docking: starting with ligand 2D SDF file, a maximum of 10 conformers per ligand are first generated with OpenBabel 2.4.1 version using genetic algorithm, and then are docked to the target protein by flexible ligand docking using Lin_F9 scoring function. The top 1 scored pose is used to construct decoy set 2. Next, we only select top 1 docked pose that is diverse from its crystal pose, but the predicted pK_d(Lin_F9) is not different too much with pK_d(exp). It meets the requirements: (I) RMSD between top 1 docked pose and crystal pose, minus RMSD between its locally optimize pose and crystal pose, should larger than 0.5 Å. (II) The difference between pK_d(exp) and pK_d(Lin_F9) is smaller than 3. The decoy set 2 consists of 5715 E2E top1 docked poses.

Overall, our training set consists of 21,943 complexes, including 9,117 crystal structures (locally optimized near native poses) and 12,826 docked structures, which have no overlap with the following validation and test sets. Based on the UniProt ID, these 21,943 complexes come from 1366 target proteins (with UniProt ID) and 42 antibody proteins (without UniProt ID). There are 7,493 structures with waters and 14,450 structures without waters in the training set, and the latter includes 6,321 docked decoys obtained from CSAR decoy set.

Validation Set—The binder subset of validation set is same as the previous Δ_{Vina}XGB's validation set[67], which included 316 complexes with three different structure types: crystal pose, local optimized pose without water and local optimized pose with receptor-bound waters. In addition, we also construct a decoy set (see Table S1) that consists of E2E top 1 docked poses of these 316 complexes in both dry and water environments. The overall validation set consists of 1578 complexes. This validation set is used to (1) conduct the early stopping in model training to avoid the overfitting of XGBoost on training set; (2) select a model that can perform well on different structure types.

Test Set—CASF-2016 benchmark[9] is used to evaluate the performance of our scoring function. Besides the standard assessment of four different powers (scoring power, ranking power, docking power and screening power) defined in CASF-2016 benchmark, we have also assessed scoring power and ranking power of our scoring function on locally optimized poses (LocalOpt) of CASF-2016 core set, in which ligand crystal poses have been locally optimized with Lin_F9.

Feature Generation

Table S2 summarizes all features employed in our scoring function development. The feature set consists of 28 buried solvent accessible surface area (bSASA) features, 48 Vina features, 3 bridge water features, 2 Beta-cluster features, one ligand efficiency (using Lin_F9 score divided by number of heavy atoms) and 10 ligand descriptors computed using RDKit version 2020.09.4. For bSASA features, same as Δ_{Vina}XGB, a total of 30 bSASA features are computed regarding to three different structures (complex, ligand, and protein). Each structure comprises of one total bSASA term and nine pharmacophore-based bSASA terms where pharmacophore types are characterized based on SYBYL[78] atom types and DOCK[79] neighboring atoms. MSMS[80] program is employed to calculate the atomic SASA with a 1.0 Å probe radius and the $\text{bSASA} = \text{SASA}_{\text{unbound}} - \text{SASA}_{\text{bound}}$. As the halogen atoms are only presented in ligand molecules, we kept only halogen-based bSASA complex term to avoid zero variance of halogen-based bSASA protein term and avoid redundancy with halogen-based bSASA ligand term. This resulted in 28 bSASA terms in our feature set.

Different from Vina 58 features[81] used in Δ_{Vina}XGB,[67] there are 48 Vina features employed in our scoring function. As shown in Table S2 and Figure S1, polar-polar, polar-nonpolar and nonpolar-nonpolar interactions in different distance ranges are described using a series of gauss functions, in which the defined polar and nonpolar atoms are based on X-Score atom types[82] (same as in Vina[45]). Also, anti-hydrogen bond, hydrogen bond and metal-ligand terms in different distances are described using a series of gauss functions as well. The anti-hydrogen bond terms describe polar-polar atoms that can't possibly be

hydrogen bond. The metal-ligand terms describe metal-ligand interactions in protein-ligand complexes. There are 37 gauss functions to describe the above interactions (see Figure S1). In addition, 6 ligand specific terms and 5 interaction terms (1 repulsion and 2 desolvation and 2 electrostatic terms) from the Vina 58 features[81] are employed as well (see Table S2). This resulted in 48 Vina terms in our feature set.

In addition, 3 bridge water features (number of bridge waters, the Lin_F9 score between bridge water and protein, the Lin_F9 score between bridge water and ligand), which are inherited from the previous Δ_{Vina}XGB, are added to our feature set. Co-crystallized waters that involve in protein-ligand interactions are considered as bridge water molecules based on the following criteria: (1) contact with both ligand and protein, the distance between oxygen atom of water and polar atoms of protein-ligand should within the range of 2.0 and 3.5 Å; (2) likely to form hydrogen bond networks, the angles between polar atoms in ligand, oxygen atom of bridge water, and polar atoms in protein are no less than 60 degrees; (3) favorable for protein-ligand binding, Lin_F9 score for bridge water is negative value when using protein or ligand as receptor.

Moreover, our feature set contains two Beta-cluster[83] features (ligand BetaScore and ligand coverage), which compute ligand and Beta-cluster overlaps in order to describe potential ligand-pocket complementarity.[83, 84] Beta cluster is a pseudo-molecular representation of fragment-centric pockets detected by AlphaSpace2.0[83]. It mimics the shape as well as atomic details of potential molecular binders. Ligand BetaScore is obtained by summing of the best Lin_F9 score of each beta-atom overlapping with ligand heavy atoms (atom distance < 1.6 Å means overlapping), this feature describes the occupied pocket ligandability. Ligand coverage is calculated by number of overlapped ligand heavy atoms divided by total number of heavy atoms. This feature describes the percentage of ligand atoms occupying the pocket.

Furthermore, 10 ligand descriptors (shown in Table S2), such as molecular logP and topological polar surface area (TPSA), are computed for each ligand using RDKit version 2020.09.4 and added to our feature set. In our above Vina 48 features, there are 6 ligand-specific terms. Thus, our feature set consists of 16 ligand-specific features and 76 protein-ligand features.

Lin_F9 Scoring Function

Lin_F9[68] is a newly developed linear scoring function that employs 9 empirical terms, including 5 Vina empirical terms (*Gauss*$_1$, *Repulsion*, *Hydrophobic*, *Hydrogen Bond*, *Number of torsions*), as well as 4 new empirical terms (two new Gauss terms, one metal bond term, one new torsion penalty term). The details of each term and Lin_F9 scoring function development can refer to this paper.[68] For the CASF-2016 benchmark scoring test, Lin_F9 performs best among 34 classical scoring functions with Pearson's correlation coefficient (R) of 0.680. We have implemented Lin_F9 in a fork of Smina docking suite as an optional built-in scoring function for protein-ligand docking. Lin_F9 is accessible through: https://yzhang.hpc.nyu.edu/Lin_F9/. Recently, Lin_F9 has been applied to the prospective virtual screening and rational inhibitor design to target SARS-Cov-2 main protease.[69]

Δ-Lin_F9 XGBoost Strategy

Similar as previous Δ-Vina strategy,[66, 67] the difference between Lin_F9 score and experimental binding affinity is used to parameterize a correction term by using XGBoost, and our Δ-Lin_F9 scoring function in term of pK_d has the following formula:

$$pK_d(\Delta_{Lin\_F9}XGB) = pK_d(Lin\_F9) + \Delta pK_d(XGBoost) \quad (1)$$

Given a training sample i with input feature vector $x_i = (x_{i\_1}, x_{i\_2}, \ldots, x_{i\_p})$, K additive trees are parameterized to predict the output y_i, in which each new tree corrects the difference between target and predictions made by all of the previous trees, as the equation shown below:

$$\hat{y}_i^{(K)} = \sum_{k=1}^{K} f_k(x_i) = \hat{y}_i^{(K-1)} + f_K(x_i). \quad (2)$$

Here, $\hat{y}_i^{(K-1)}$ is the prediction from previous K-1 trees and $f_K(x_i)$ is the K-th tree model. The objective function consists of loss function ($l(y_i, \hat{y}_i)$) and regularization term ($\Omega(f_k)$) for tree complexity as follows:

$$\mathcal{L}(f_k) = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{i=1}^{k} \Omega(f_i) = \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(k-1)} + f_k(x_i)\right) + \Omega(f_k) + constant \quad (3)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\omega^2 \quad (4)$$

Here, mean squared error (MSE) is used as our loss function and the regularization term Ω is used to control the model complexity in which T and ω refer the number of leaves and the scores on leaves in f_k respectively. To reduce correlations among trees, only a random subset of features is chosen for splitting in each tree development.

In our development of $\Delta_{Lin\_F9}XGB$, we use the ΔpK_d as the label y and employ the XGBoost package (version 1.20) in Python 3.7 to build the XGBoost model. The input feature vector x has p = 92 features. The hyper-parameters utilized in our model are n_estimators = 800, learning_rate = 0.04, subsample = 0.6, colsample_bytree = 0.8, min_child_weight = 2, max_depth = 12, and loss function (regression type) = “reg: squarederror”. By using the validation set, the early stopping round with 50 steps is applied to reduce overfitting of training data. Considering the high variance of XGBoost models, the final scoring prediction is the average of ten XGBoost models initialized with different random seeds.

Evaluation Methods

CASF 2016 Benchmark—The comparative assessment of scoring functions (CASF) benchmark[5–9] provides four different powers (scoring power, ranking power, docking power and screening power) for evaluation of scoring function's performance. CASF-2016[9] is the latest version of CASF benchmark, which has tested more than 30 prevailing scoring. Pearson's correlation coefficient (R) between predicted binding affinity and experimental measured binding affinity is used to evaluate scoring power. The CASF-2016 benchmark includes 57 targets and 5 known ligands for each target. The Spearman's rank correlation coefficient (ρ) is used as the quantitative indicator of ranking power by averaging over 57 targets. Docking power is evaluated by the success rate of top scored poses having RMSD less than 2 Å in comparison with the crystal pose. Screening power, which refers the ability of a scoring function to identify true binders among a pool of molecules, is evaluated by two quantitative indicators: one is the success rate of identifying the highest-affinity binder among the 1%, 5% and 10% top-ranked ligands over all 57 targets, and the other is the enhancement factor (EF) computed with the following formula:

$$\mathrm{EF}_{\alpha} = \frac{NTB_{\alpha}}{NTB_{total} \cdot \alpha} \quad (5)$$

Here, NTB_{α} is the number of true binders among top α ranked candidates (e.g. $\alpha = 1\%, 5\%, 10\%$) based on predicted binding affinities. NTB_{total} is the total number of true known binders for a given target. The final EF_{α} is also the average over all 57 targets.

Extended Docking-Scoring tests of CASF-2016 core set—Besides standard assessment, we also carried out extended tests of scoring and ranking performances with various structure types of protein-ligand complexes that are generated by several flexible docking protocols as which have first been introduced in tests of Lin_F9. They are summarized in Table 1, including: (i) flexible re-docking, (ii) E2E docking, (iii) ensemble docking using all 5 protein structures and (iv) ensemble docking using 4 non-native protein structures. All these docking experiments are conducted using a fork of Smina docking suite with Lin_F9 scoring function. Both dry environment (protein without water) and water environment (protein with receptor-bound waters) are evaluated for these docking-scoring tests. For each docking-scoring test, after docking, the top 5 docked poses are selected for re-scoring using $\Delta_{\mathrm{Lin_F9}}$XGB, and the best-scored pose from re-scoring is used for scoring, ranking and docking performances evaluation. In term of docking performance evaluation, the symmetry corrected RMSD between the best-scored pose and crystal pose is calculated by open source tool DockRMSD.[85]

D3R GC4 Datasets—Drug Design Data Resource (D3R)[86–89] Grand Challenge 4 (GC4)[86] consist of two sub-challenges: one is BACE1 Sub-challenge for pose prediction of 20 macrocyclic BACE inhibitors and affinity ranking of 154 macrocyclic BACE inhibitors, and the other is CatS Sub-challenge for affinity ranking of 459 CatS inhibitors. Our group has participated in GC4 BACE1 Sub-challenge, and the pose prediction of our submitted model achieved an average RMSD of 1.01 Å. In further exploration after competition, the

average RMSD can be decreased to 0.74 Å using a similarity-based constraint docking method[32]. Similar as in the work of Lin_F9, our latest predicted poses of whole BACE1 dataset are used to test the affinity ranking performances of our scoring function and Vina (as comparison); for the CatS dataset (459 inhibitors), predicted poses obtained from Top submitter's model (Max Totrov group, receipt id: x4svd) in D3R website[86] are used. It should be noted that, before re-scoring, all the poses are local optimized by Vina or Lin_F9.

LIT-PCBA Dataset—LIT-PCBA is an unbiased data set designed for benchmarking virtual screening (VS) and machine-learning, and it can be directly used for the evaluation of screening performance of scoring functions.[75, 90] LIT-PCBA dataset consists of 15 diverse target sets, 8020 true actives and 2,675,399 true inactive compounds. The high imbalance between actives and inactives is intended to mimic the real-life screening tasks, which makes it quite challenging for computational screening methods, thereby offering an opportunity to estimate performance of virtual screening protocols in practical applications.

For each of the 15 targets in LIT-PCBA, several PDB templates are available (no more than 15 crystal complexes) as input receptor files for docking. In order to save the CPU cost, we only used a limit number of PDB templates for targets with more than 100,000 compounds. For each target, only one ligand binding site was considered for docking. For example, there are 2 different structures (HAT domain and BRD domain structures) in KAT2A target, we only selected HAT domain structure as the receptor for docking. RDKit 2020.09.4 version[91, 92] was used to read SMILES string and add hydrogens and generate initial 3D conformer for each compound. The following docking protocol was same as the E2E docking described above. After docking, for each compound, top 5 docked poses were selected for re-scoring using Δ_{Lin_F9}XGB and the best-score from re-scoring was used to rank the compound in the library. The EF at top 1% was used as the quantitative indicator to evaluate the screening power.

RESULTS AND DISCUSSION

CASF-2016 Benchmark Assessment

The scoring-ranking-docking-screening performance of Δ_{Lin_F9}XGB was tested on standard CASF-2016 benchmark and compared with other traditional scoring functions (Figure 1). Meanwhile, scoring power comparison with several recently developed machine learning scoring functions has been carried out, as shown in Figure 2 and Table 2. In addition, based on CASF-2016 benchmark, the Δ_{Lin_F9}XGB's scoring and ranking performances on locally optimized poses, which are obtained by Lin_F9 local optimization of crystal poses, were also tested.

Standard Assessment—In Figure 1, Δ_{Lin_F9}XGB is compared with three scoring functions, $\Delta_{Vina}RF_{20}$, Δ_{Vina}XGB and Lin_F9, previously developed in our group,[66, 67] as well as 33 traditional scoring functions that have been evaluated by Su et al.[9] Our evaluation results show that Δ_{Lin_F9}XGB achieves Top 1 performances on scoring power and ranking power, it also achieves Top 2 performances on both the enhancement factor at top 1% ($EF_{1\%}$) and success rate at top 1% ($SC_{1\%}$) of screening power tasks. For the scoring

power, the Pearson's correlation coefficient (R) and the Root-Mean-Square-Error (RMSE) of Δ_{Lin_F9}XGB are 0.845 and 1.240, which are better than our previous best result (R = 0.796 and RMSE=1.327 for Δ_{Vina}XGB), as shown in Figure 1A. For the ranking power comparison, Figure 1B shows that Δ_{Lin_F9}XGB achieves the best ranking power with the average Spearman's rank correlation coefficient $\rho = 0.704$, which is also much better than our previous best result ($\rho = 0.647$ for Δ_{Vina}XGB). In addition, Kendall correlation coefficient (τ) ranking performance of Δ_{Lin_F9}XGB is evaluated as well, with $\tau = 0.625$. For the docking power assessment, 86.7% of Δ_{Lin_F9}XGB predicted best-scored pose is considered to be successfully docked if 2.0 Å RMSD threshold is used in comparison with the crystal pose. In Figure 1C, the docking power of Δ_{Lin_F9}XGB is ranked at the 7th place among all 37 scoring functions. Enhancement factor and success rate at top 1% level are computed as the indicators for the screening power, which evaluates the ability of a scoring function to identify true binders from random compounds. As shown in Figure 1D and 1E, the screening power is ranked at the 2nd place among all 37 scoring functions. $EF_{1\%}$ and $SC_{1\%}$ of Δ_{Lin_F9}XGB are 12.61 and 40.4%. When computing the success rate at the top 5% level ($SC_{5\%}$) and at the top 10% level ($SC_{10\%}$), Δ_{Lin_F9}XGB achieves 59.6% and 68.4%, which are much better than previous Δ_{Vina}XGB's performance (see Table S3 in Supporting Info).

In addition, in Figure 2, we compared with several advanced ML scoring functions for the scoring power of CASF-2016 benchmark test, since these ML scoring functions are mainly developed for protein-ligand binding affinity prediction. As can be seen, our Δ_{Lin_F9}XGB model ranks at the 6th position among these start-of-the-art models. The top 5 performers are graphDelta[17] (graph-convolutional neural network model, Pearson's R = 0.87), ECIF::LD-GBT[18] (gradient boosting tree model incorporating extended connectivity interaction features and RDKit ligand features, Pearson' R = 0.866), OnionNet-2[15] (convolutional neural network model with inputs based on rotation-free specific contacts between protein and ligand in different shells, Pearson's R = 0.864), TopBP[12] (a consensus model incorporating different ML methods and with inputs based on algebraic topology for characterizing biomolecular complexes, Pearson's R = 0.861), ECIF::GBT[18] (gradient boosting tree model incorporating only extended connectivity interaction features, Pearson's R = 0.857). Other methods, such as persistent spectral based ML models (Mol-PSI[13] and PerSpect ML[14]), algebraic graph theory-based model (AGL-Score[19]) and usage of diverse ligand-based features in previous ML model (RF-Score v3+RDKit[22]), also show very good scoring power in CASF-2016 test. All these methods enrich the methodology for ML scoring function development, and our Δ_{Lin_F9}XGB also achieves state-of-the-art scoring performance among these methods. It should be noted that, although all the ML scoring functions presented in Figure 2 use PDBbind dataset for their training and validation, the differences in PDBbind versions (v2007, v2010, v2014 and v2016) as well as the choice of using PDBbind refined set or general set, will also affect the scoring performance test on CASF-2016 benchmark.

Moreover, we also compared with several ML scoring functions that have been evaluated with at least three different metrics for CASF-2016 benchmark. Many other ML scoring

functions that only presented scoring power in their original paper are not summarized in Table 2. As can be seen, Δ_{Lin_F9}XGB shows best scoring and ranking powers among these scoring functions. AEScore,[93] a deep neural network model, also has very good scoring power (R = 0.830), but its docking power is very low (success rate = 35.8%). This low success rate is also observed with AK-score (ensemble),[94] a CNN-based scoring function, reporting a success rate of 36.0%. A similar scoring function that employed Δ-Learning to retain docking power is Δ-AEScore,[93] which reports a success rate of 85.6%. But its screening power is far less satisfying ($EF_{1\%}$ = 6.16), worse than the Vina ($EF_{1\%}$ = 7.70). Recently, Wegner et al proposed DeepDock,[95] a method based on geometric deep learning to predict the ligand binding poses using distance potential, achieving very good docking power (success rate = 87.0%) and screening power ($EF_{1\%}$ = 16.41). Scoring and ranking powers are not evaluated since DeepDock is not trained to predict binding affinities. Their study inspired us to train a native pose identification model in our future work to further improve the screening power. Standard assessment shows that Δ_{Lin_F9}XGB is already a very robust and competitive protein-ligand scoring function for different tasks.

Locally Optimized Poses Assessment—Based on the crystal structure of CASF-2016 core set, we also evaluated scoring power and ranking power of Δ_{Lin_F9}XGB on its locally optimized (LocalOpt) pose, which is locally optimized from crystal pose using Lin_F9 scoring function. Both LocalOpt pose in dry environment (C_o) and LocalOpt pose in water environment (C_o^{rw}) are evaluated. Figure 3A, B illustrate the scoring power and ranking power on LocalOpt (C_o and C_o^{rw}), together with performance on crystal structure without local optimization (C). Compared with Δ_{Vina}XGB, $\Delta_{Vina}RF_{20}$, Lin_F9 and Vina, our new developed Δ_{Lin_F9}XGB achieves much better scoring and ranking power on both C_o and C_o^{rw}. For the scoring power, the Pearson's R of Δ_{Lin_F9}XGB on C_o and C_o^{rw} are 0.853 and 0.834. The RMSE of Δ_{Lin_F9}XGB on C_o and C_o^{rw} are 1.162 and 1.205. For the ranking power, the Spearman's ρ of Δ_{Lin_F9}XGB on C_o and C_o^{rw} are 0.693 and 0.700, respectively. A scatter plot of experimental pK_d vs predicted pK_d for LocalOpt pose (C_o) is shown in Figure 4B. The results indicate that Δ_{Lin_F9}XGB performs consistently well on the near native poses.

Docking Tests of CASF-2016 Core Set

In order to further test the scoring, ranking and docking performances for real docking application, we enlarged the evaluation category from re-scoring of crystal pose and LocalOpt pose to re-scoring of docking poses. Several docking tests on CASF-2016 core set (illustrated in Table 1) are carried out, and the docked poses are re-scored by Δ_{Lin_F9}XGB to select the best-scored pose for scoring, ranking and docking evaluations.

Flexible re-docking test—In docking preparation of flexible re-docking, both the ligand conformer and protein conformation come from the corresponding crystal protein-ligand complex. After docking, the top 5 docked poses for each complex were selected and then re-scored by Δ_{Lin_F9}XGB. Similar as LocalOpt pose, both docked poses without water (C_{fd})

and with water molecules (C_{fd}^{rw}) are evaluated. In the evaluation process, the best-scored pose from re-scoring is used for assessment, and the performance is compared with its baseline Lin_F9, as well as Δ_{Vina}XGB, Δ_{Vina}RF$_{20}$ and Vina.

Figure 5A, B illustrate the scoring power and ranking power on C_{fd} and C_{fd}^{rw}. The baseline Lin_F9 achieves much better scoring and ranking performances than Vina on both C_{fd} and C_{fd}^{rw}, which has been discussed in our previous paper. Then, Δ_{Lin_F9}XGB is used to re-score the top 5 docked poses, and the best-scored pose from re-scoring further improved the scoring and ranking performances a lot. The Pearson's R of Δ_{Lin_F9}XGB on C_{fd} and C_{fd}^{rw} are 0.839 and 0.826, respectively. The RMSE of Δ_{Lin_F9}XGB on C_{fd} and C_{fd}^{rw} are 1.204 and 1.238. For the ranking power, the Spearman's ρ of Δ_{Lin_F9}XGB on C_{fd} and C_{fd}^{rw} are 0.712 and 0.723, respectively. A scatter plot of experimental pK_d vs predicted pK_d for flexible re-docking pose (C_{fd}) is shown in Figure 4C. As far as we know, the scoring and ranking performances for flexible re-docking of CASF-2016 core set are better than existing ML scoring functions. In addition, docking success rates of Vina, Δ_{Vina}RF$_{20}$, Δ_{Vina}XGB, Lin_F9 and Δ_{Lin_F9}XGB on both C_{fd} and C_{fd}^{rw} are computed to assess the ability of scoring function to identify near-native pose, as shown in Figure 5C. Vina achieves highest docking success rate on both C_{fd} and C_{fd}^{rw}. At a 2 Å RMSD threshold, the docking success rates for Vina, Δ_{Vina}RF$_{20}$, Δ_{Vina}XGB, Lin_F9 and Δ_{Lin_F9}XGB on C_{fd} are 69.1%, 67.0%, 64.9%, 57.9% and 57.9%, respectively. For C_{fd}^{rw}, the docking success rates for Vina, Δ_{Vina}RF$_{20}$, Δ_{Vina}XGB, Lin_F9 and Δ_{Lin_F9}XGB are 84.6%, 83.9%, 80.0%, 80.0% and 79.3%, respectively. All scoring function's performances improved a lot after keeping receptor-bound water, which demonstrates the importance of explicit water molecules for molecular docking. The difference of docking success rates between Vina and Δ_{Lin_F9}XGB decreases from around 11.2% to 5.3% when water molecules are included. It is observed that scoring and ranking powers of Δ_{Lin_F9}XGB are significantly better than Vina on both C_{fd} and C_{fd}^{rw}, while the docking success rate of Vina is higher. This trend is consistent for other docking tests (see Figure 5C).

End-to-End Docking Test—Here we evaluated the flexible docking on CASF-2016 core set in an end-to-end (E2E) protocol, in which ligand conformer generation and flexible docking are combined. For this E2E docking protocol, maximum 10 conformers were generated for each small molecule, and all these conformers were docked to the target protein. After docking, the top 5 docked poses were re-scored using Δ_{Lin_F9}XGB, and the best-scored pose from re-scoring was used to assess the performance of Δ_{Lin_F9}XGB model.

Figure 5A, B illustrate the scoring power and ranking power on E2E docked pose in dry environment (C_{E2E}) and in water environment (C_{E2E}^{rw}). The baseline Lin_F9 achieves much better scoring and ranking performances than Δ_{Vina}RF$_{20}$ and Vina on both C_{E2E} and C_{E2E}^{rw}. Similar as above flexible re-docking results, the best-scored pose from re-scoring

by $\Delta_{Lin\_F9}XGB$ further improved the performance. For the scoring power, the Pearson's R of $\Delta_{Lin\_F9}XGB$ on C_{E2E} and C_{E2E}^{rw} are 0.805 and 0.785, respectively. The RMSE of $\Delta_{Lin\_F9}XGB$ on C_{E2E} and C_{E2E}^{rw} are 1.314 and 1.356. For the ranking power, the Spearman's ρ of $\Delta_{Lin\_F9}XGB$ on C_{E2E} and C_{E2E}^{rw} are 0.647 and 0.618. For the docking power, the docking success rates of best-scored pose for Vina, $\Delta_{Vina}RF_{20}$, Lin_F9 and $\Delta_{Lin\_F9}XGB$ are shown in Figure 5C. Once again, Vina achieves highest docking success rate on both C_{E2E} and C_{E2E}^{rw}. At a 2 Å RMSD threshold, the docking success rates for Vina, $\Delta_{Vina}RF_{20}$, $\Delta_{Vina}XGB$, Lin_F9 and $\Delta_{Lin\_F9}XGB$ on C_{E2E}^{rw} are 69.5%, 68.8%, 68.4%, 63.9% and 59.6%, respectively. Though the docking power of Vina is better, its scoring and ranking performances are less satisfied. Overall, our $\Delta_{Lin\_F9}XGB$ achieves much better scoring and ranking performances for this E2E docking test.

Ensemble (E5 and E4) docking tests—Ensemble docking is a practically useful approach to account for protein flexibility in docking applications by docking a ligand into a selected ensemble of protein structures. The CASF-2016 core set can be used to evaluate the performance of ensemble docking, as it includes 57 targets and 5 protein structures for each target. For each ligand in CASF-2016 core set, it can be docked into 5 protein structures with the E2E docking protocol. After docking, the top 5 docked poses for each protein structure were selected and re-scored by $\Delta_{Lin\_F9}XGB$. The best-scored pose from re-scoring was selected to calculate scoring- ranking-docking performances of our scoring function.

From Figure 5A and 5B, we can see that in comparison with the other four scoring functions, $\Delta_{Lin\_F9}XGB$ achieves much better scoring and ranking powers for this ensemble E5 docking test. For the scoring power, the Pearson's R of $\Delta_{Lin\_F9}XGB$ on C_{E5} and C_{E5}^{rw} are 0.813 and 0.790, respectively. The RMSE of $\Delta_{Lin\_F9}XGB$ on C_{E5} and C_{E5}^{rw} are 1.283 and 1.343. For the ranking power, the Spearman's ρ of $\Delta_{Lin\_F9}XGB$ on C_{E5} and C_{E5}^{rw} are 0.677 and 0.616. Both the scoring power and ranking power of ensemble E5 docking are slightly better than the above E2E docking with native protein structure. This suggests that, for $\Delta_{Lin\_F9}XGB$, ensemble docking can improve the scoring and ranking performances than docking with a single structure. At a 2 Å RMSD threshold, docking success rates for Vina, $\Delta_{Vina}RF_{20}$, $\Delta_{Vina}XGB$, Lin_F9, $\Delta_{Lin\_F9}XGB$ on C_{E5}^{rw} are 56.5%, 56.8%, 56.8%, 52.3% and 51.9%, respectively.

A more stringent test for ensemble docking is to exclude the native protein structure for each ligand from the ensemble. The docked poses without water (C_{E4}, in which E4 represents 4 ensemble protein structures used) and with water molecules (C_{E4}^{rw}) were assessed. As shown in Figure 5, for the scoring power, the Pearson's R of $\Delta_{Lin\_F9}XGB$ on C_{E4} and C_{E4}^{rw} are 0.808 and 0.768, respectively. The RMSE of $\Delta_{Lin\_F9}XGB$ on C_{E4} and C_{E4}^{rw} are 1.309 and 1.418. For the ranking power, the Spearman's ρ of $\Delta_{Lin\_F9}XGB$ on C_{E4} and C_{E4}^{rw} are 0.661 and 0.593. At a 2 Å RMSD threshold for the best-scored pose, docking success rates for Vina,

$\Delta_{Vina}RF_{20}$, $\Delta_{Vina}XGB$, Lin_F9, $\Delta_{Lin\_F9}XGB$ on C^{rw}_{E4} are 36.1%, 36.8%, 37.2%, 37.2% and 37.2%, respectively.

Altogether, in term of (i) flexible-redocking, (ii) E2E docking, (iii) ensemble docking test including native protein structure and (iv) ensemble docking test excluding native protein structure, the scoring and ranking performances of our new developed $\Delta_{Lin\_F9}XGB$ achieves consistently superior prediction accuracy on these real docking tests.

Case Studies of D3R GC4 Datasets

Here we evaluated affinity ranking performances of $\Delta_{Lin\_F9}XGB$ on two D3R GC4 challenge datasets [86] regarding beta secretase 1 (BACE1) and Cathepsin S (CatS) respectively. Both targets are of significant pharmaceutical interests.[96–101] The structure-based ranking protocol depends on the protein-ligand complex structures. However, both BACE1 and CatS datasets are very challenge for pose prediction using traditional docking program (such as Smina) and ligand conformer generation method (such as RDKit).[32, 87, 102–104] For BACE1 dataset, we used a similarity-based constraint docking method to generate the near-native poses. The method uses similar co-crystal macrocycle ring with BACE1 structure as reference in the sampling process and has achieved very good pose prediction performance (see Figure S6). For the CatS dataset, the poses were obtained from Max Totrov group's submitted data (receipt ID: x4svd) on D3R website, since they have achieved top2 pose prediction in previous GC3 CatS competition.[87]

BACE1 macrocyclic inhibitor dataset.—The BACE1 dataset encompasses 154 small molecules inhibitors, in which 151 of 154 ligands have macrocycle rings.[86] The measured binding affinities of the dataset span over five orders of magnitude range of IC_{50} (pIC_{50} range from 4.2 to 9.3). The macrocycle ring size ranges from 14 to 17, and conformation of these macrocycle rings are hard to properly generated due to the limitation of ligand conformational sampling methods.[105, 106] Here, we evaluated the ranking powers of Vina, $\Delta_{Vina}RF_{20}$, $\Delta_{Vina}XGB$, Lin_F9 and $\Delta_{Lin\_F9}XGB$ on whole 154 ligands, in which the poses were predicted using our similarity-based constraint docking method (see Figure S6). Table 3 illustrates the ranking powers of Vina, $\Delta_{Vina}RF_{20}$, $\Delta_{Vina}XGB$, Lin_F9 and $\Delta_{Lin\_F9}XGB$ on LocalOpt pose. For the ranking power, the Spearman's ρ of Vina, $\Delta_{Vina}RF_{20}$, $\Delta_{Vina}XGB$, Lin_F9 and $\Delta_{Lin\_F9}XGB$ are 0.332, 0.299, 0.307, 0.439 and 0.481, respectively. The Kendall's τ of Vina, $\Delta_{Vina}RF_{20}$, $\Delta_{Vina}XGB$, Lin_F9 and $\Delta_{Lin\_F9}XGB$ are 0.222, 0.201, 0.211, 0.311 and 0.349. The baseline Lin_F9 achieves better ranking power when compared with Vina, $\Delta_{Vina}RF_{20}$ and $\Delta_{Vina}XGB$. $\Delta_{Lin\_F9}XGB$ further improved the ranking power, which achieves the Top 3 place when compared with top 20 submissions on D3R website (the best performer in D3R achieves $\rho = 0.54$ and $\tau = 0.39$),[86, 102, 103, 107–109] as can be seen in Figure 6. It should be noted that our $\Delta_{Lin\_F9}XGB$ is a general scoring function evaluated for this target-specific challenging case. The scoring power is consistently improved by $\Delta_{Lin\_F9}XGB$ since the Pearson's R and RMSE of $\Delta_{Lin\_F9}XGB$ are better in general.

CatS dataset—The GC4 CatS dataset is composed of 459 small molecule inhibitors with measured binding affinities spanning over three orders of magnitude range of IC_{50} (pIC_{50} range from 5.0 to 8.2).[86] The challenge for ranking these CatS inhibitors might come from their large size, high flexibility and similar chemical structures. The D3R organizers observed an obvious improvement in participant performance for CaS between GC3 and GC4, which may come from the use of GC3 CatS data or other CatS data from ChemBL[110–112] to develop target-specific machine learning models.[86, 87] Similarly, we evaluated the ranking performance of five general scoring functions: Vina, $\Delta_{Vina}RF_{20}$, $\Delta_{Vina}XGB$, Lin_F9 and $\Delta_{Lin\_F9}XGB$ on GC4 CatS dataset. The poses were obtained from Max Totrov group's submitted data on D3R website.[107] They have participated in both GC3 and GC4 CatS Sub-challenges with available predicted pose structures, and their submitted data performs very well on pose prediction (Top 2 pose prediction in GC3[87]). Based on their predicted poses for 459 CatS inhibitors, Table 4 illustrates the scoring and ranking powers of the five scoring functions on LocalOpt pose. For the ranking power comparison, Spearman's ρ of Vina, $\Delta_{Vina}RF_{20}$, $\Delta_{Vina}XGB$, Lin_F9 and $\Delta_{Lin\_F9}XGB$ are 0.430, 0.430, 0.399, 0.446 and 0.457, respectively. The Kendall's τ of Vina, $\Delta_{Vina}RF_{20}$, $\Delta_{Vina}XGB$, Lin_F9 and $\Delta_{Lin\_F9}XGB$ are 0.293, 0.296, 0.275, 0.304 and 0.309. The improvement of ranking power for CatS is not as significant as the above BACE1. This could be attributed to either limitation of general scoring function for CatS or narrow binding affinities range of the dataset (only three orders of magnitude range for 459 CatS inhibitors). In Figure 7, we also compared with top 20 submissions on D3R website,[86, 107, 108] $\Delta_{Lin\_F9}XGB$ only ranks at the 15th place. The top 2 performers in D3R GC4 CatS used target-specific 3D-QSAR model[107] and target-specific ligand-based deep neural network model (unpublished yet).

Assessment of Screening Power on LIT-PCBA Dataset

Many previous studies evaluated virtual screening methods based on the Directory of Useful Decoys (DUD)[113] and its successor DUD-E[114], in which most of the presumed decoys have not been experimentally verified. To overcome this drawback, recently Tran-Nguyen and co-workers proposed LIT-PCBA,[75] a dataset derived from dose-response assays in the PubChem BioAssay database.[115, 116] All the actives and inactives in LIT-PCBA were taken from the experimental data under homogeneous conditions. Preliminary virtual screening (VS) experiments indicated that LIT-PCBA is very challenging, due to the (I) high imbalance active/inactive compounds to mimic the real screening hit rate, (II) common molecular properties shared between active and inactive compounds, (III) weak potencies of the active compounds. One main limitation of LIT-PCBA dataset is that, more than half of the primary assays (8 of 15 targets) are cell-based phenotypic assays, so many actives are not validated against their putative target. Structure-based virtual screening tests on this benchmark may have some issues, nevertheless, LIT-PCBA still provide valuable clues for evaluation of scoring functions in large-scale VS.

Here, VS experiments on all 15 targets in LIT-PCBA were carried out using the E2E docking protocol as described above in CASF-2016 docking test. It should be noted that, in order to save the CPU cost, the PDB templates of some targets (targets with more than 100,000 compounds) used for docking experiments are less than the original LIT-PCBA

provided. In addition, some targets in LIT-PCBA, such as ALDH1, IDH1 and KAT2A, have more than one ligand binding site in the PDB templates (shown in Figure S7). Based on the assay description and co-crystal ligand type, we only selected one docking site for each target. Also, previous Δ_{Vina}XGB is excluded in this VS evaluation since it needs to calculate the time-consuming ligand stability features using RDKit (need to generate maximum 1000 conformers per ligand).

As shown in Table 5, the $EF_{1\%}$ metric is used as the quantitative indicator to evaluate the screening performances of Vina, $\Delta_{Vina}RF_{20}$, Lin_F9 and Δ_{Lin_F9}XGB on LIT-PCBA dataset. The average $EF_{1\%}$ metric of Δ_{Lin_F9}XGB over all 15 targets is 5.55, which clearly outperforms Vina (average $EF_{1\%}$ = 2.78), $\Delta_{Vina}RF_{20}$(average $EF_{1\%}$ = 3.18) and Lin_F9 (average $EF_{1\%}$ = 2.21). Similar with the previous virtual screening results from Tran-Nguyen et al,[90] the average $EF_{1\%}$ values range from 2 to 6, indicating the challenge of the dataset. In addition, counting number of targets that satisfy the increasing thresholds (2, 5, and 10) of $EF_{1\%}$ values, serves as a comprehensive metric to evaluate the generalization ability of scoring function on diverse targets. As can be seen, Δ_{Lin_F9}XGB achieves the best screening performance among these four scoring functions. At $EF_{1\%} > 2$, $EF_{1\%} > 5$, and $EF_{1\%} > 10$, number of satisfied targets for Δ_{Lin_F9}XGB are 13, 8 and 2, respectively. Vina only have 6 satisfied targets at $EF_{1\%} > 2$ threshold, limiting its applicability in real virtual screening. Two target sets (TP53 and VDR) are really challenging for Δ_{Lin_F9}XGB since it yields $EF_{1\%} < 2$. These two challenge cases were also observed by Tran-Nguyen et al,[90] and they mentioned the main reason for this failure is the weak potencies of the actives. Overall, the results indicate that Δ_{Lin_F9}XGB has the top early hit enrichment ability among these four scoring functions for this challenging LIT-PCBA dataset.

In addition, we compared the ROC curves and AUC values (in the Table S5, Figure S8 and S9 of the Supporting Information) in this evaluation. The average AUC values of Vina, $\Delta_{Vina}RF_{20}$, Lin_F9 and Δ_{Lin_F9}XGB on LIT-PCBA benchmark are 0.571, 0.560, 0.586 and 0.603, respectively. The AUC results are quite different from the $EF_{1\%}$ results, since Δ_{Lin_F9}XGB only shows slightly better performance based on AUC values. It is due to the reason that $EF_{1\%}$ metric focuses on the early hit enrichment. The ROC curves in Figure S9 also show that, compared with Vina, $\Delta_{Vina}RF_{20}$ and Lin_F9, actives in 7 of 15 targets (ADRB2, ALDH1, GBA, IDH1, KAT2A, OPRK1 and PPARG) can be earlier enriched by Δ_{Lin_F9}XGB.

In Table 6, we also collected LIT-PCBA benchmark test results from three other groups. Tran-Nguyen et al[90] evaluated 5 scoring functions (Surflex,[117] Pafnucy,[24]$\Delta_{Vina}RF_{20}$,[66] IFP[118] and GRIM[119]) where the IFP achieved the best performance (average $EF_{1\%}$= 7.46). They used Surflex-Dock to generate top 20 poses for each ligand and other 4 scoring functions were used to re-scoring these poses. For targets with several PDB templates, the highest score value for each compound was used to evaluate the early hit enrichment performance (for IFP and GRIM, the score for each compound is the highest similarity between the templates and the docking poses). Compared with original Surflex-Dock, all re-scoring methods improved the screening performance a lot, and the re-scoring based on

simple interaction fingerprints (IFP) or interaction graphs (GRIM) outperforms ML scoring functions. It should be noted that the performance of IFM and GRIM are highly dependent on the PDB template: choosing one that is not well represented in the dataset can lead to much worse results. Zhou et al[120] reported the test results of their template-based virtual screening models (FINDSITE$^{comb2.0}$ and FRAGSITE), which also showed comparable screening performance (FRAGSITE's average $EF_{1\%}$= 4.78). Sunseri et al[121] assessed the built-in CNN models of GNINA[122] compared to 4 other scoring functions (RFScore-4,[123] RFScore-VS,[49] Vina,[45] Vinardo[124]). Their CNN default (Affinity) also achieved comparable performance (average $EF_{1\%}$= 4.64). It is interesting to note that, the early enrichment performance of Vina (average $EF_{1\%}$= 2.78) in our test is better than their Vina result (average $EF_{1\%}$= 1.71), despite the average AUC values of Vina are the same (average AUC = 0.57 in both our and their tests). We find that the docking protocols are somewhat different: (I) they used one conformer per ligand for docking, while we used maximum 10 conformers per ligand for docking; (II) they used --autobox_add 16 to define docking box, while we used the default --autobox_add 4; (III) they used all the PDB templates for docking, while we used less PDB templates for some targets. One comparable target is ESR1_ago, since all 15 PDB templates are used in our docking. Our Vina test result is 15.38, while their Vina test result is 7.69. This target could contribute most of the Vina performance difference. The results suggest that LIT-PCBA benchmark early hit enrichment performance is not only dependent on the protein-ligand scoring function, but also is influenced by docking protocols and parameters.

Moreover, target-specific scoring functions developed based on LIT-PCBA dataset (split into training set and validation set with ratio 3:1) show better performance over the generic scoring functions summarized in Table 6. Shen et al[125] reported their finding that ligand-based target-specific models (2D fingerprint-based QSAR models, best model performance $EF_{1\%}$ = 14.59) and structure-based target-specific models (descriptor-based XGBoost models, best model performance $EF_{1\%}$ = 8.93) outperform classical scoring function (Glide SP, $EF_{1\%}$ = 3.37) on 7 targets of LIT-PCBA validation set. With the abundant specific target training samples, target-specific scoring functions (both ligand-based and structure-based) can outperform current generic ML scoring functions as a promising alternative. However, it should be noted that the target-specific scoring function approach will not be applicable for a novel target with little experimental data available.

CONCLUSION

In order to develop a robust protein-ligand scoring function that can perform well for a variety of docking tasks, we have explored our previous developed Lin_F9 scoring function as the baseline and via Δ-Learning XGBoost approach to correct Lin_F9 score. The training set is enlarged to include docked poses, and physically meaningful features are explored. Our new scoring function Δ_{Lin_F9}XGB can not only perform consistently among the top compared to traditional scoring functions for scoring-ranking-screening powers of CASF-2016 benchmark, but also achieves superior scoring and ranking powers on real docked poses, including flexible re-docking, E2E docking and ensemble docking. Also, compared with Vina and $\Delta_{Vina}RF_{20}$ and Δ_{Vina}XGB, Δ_{Lin_F9}XGB achieves better ranking

power for target-specific cases (BACE1 and CatS) of D3R GC4. Large-scale docking-based virtual screening test on LIT-PCBA dataset demonstrates the reliability and robustness of Δ_{Lin_F9}XGB in virtual screening application. In summary, although there remains substantial room for virtual screening performance improvement, our extensive test results suggest that Δ_{Lin_F9}XGB has improved both robustness and applicability of machine-learning scoring functions in real docking application, and can serve as a very useful re-scoring tool for structure-based inhibitor design.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENT

This work was supported by the U.S. National Institutes of Health (R35-GM127040). We thank NYU-ITS for providing computational resources.

DATA AND SOFTWARE AVAILABILITY

Δ_{Lin_F9}XGB is accessible through: https://yzhang.hpc.nyu.edu/Delta_LinF9_XGB. RDKit 2020.09.4 version[91, 92] is used to read SMILES string and add hydrogens and generate initial 3D conformer for each compound. OpenBabel 2.4.1 version[126] is used to generate multiple conformations based on RMSD. MGLTools 1.5.4 version[127] is used for preparing PDBQT files of protein and ligand. DockRMSD[85] is used for the calculation of RMSD between docking pose and original crystal pose of the same ligand molecule.

References

1. Forli S; Huey R; Pique ME; Sanner MF; Goodsell DS; Olson AJ, Computational Protein–Ligand Docking and Virtual Drug Screening with the Autodock Suite. Nature protocols 2016, 11, 905–919. [PubMed: 27077332]
2. Irwin JJ; Shoichet BK, Docking Screens for Novel Ligands Conferring New Biology: Miniperspective. J. Med. Chem 2016, 59, 4103–4120. [PubMed: 26913380]
3. Lyu J; Wang S; Balius TE; Singh I; Levit A; Moroz YS; O'Meara MJ; Che T; Algaa E; Tolmachova K, Ultra-Large Library Docking for Discovering New Chemotypes. Nature 2019, 566, 224–229. [PubMed: 30728502]
4. Guedes IA; Pereira FS; Dardenne LE, Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. Front. pharmacol 2018, 9, 1089. [PubMed: 30319422]
5. Cheng T; Li X; Li Y; Liu Z; Wang R, Comparative Assessment of Scoring Functions on a Diverse Test Set. J. Chem. Inf. Model 2009, 49, 1079–1093. [PubMed: 19358517]
6. 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. [PubMed: 24708446]
7. 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. [PubMed: 24716849]
8. 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. Nature protocols 2018, 13, 666–680. [PubMed: 29517771]

9. 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 2018, 59, 895–913. [PubMed: 30481020]
10. Leach AR; Shoichet BK; Peishoff CE, Prediction of Protein− Ligand Interactions. Docking and Scoring: Successes and Gaps. J. Med. Chem 2006, 49, 5851–5855. [PubMed: 17004700]
11. Huang S-Y; Grinter SZ; Zou X, Scoring Functions and Their Evaluation Methods for Protein–Ligand Docking: Recent Advances and Future Directions. Phys. Chem. Chem. Phys 2010, 12, 12899–12908. [PubMed: 20730182]
12. Cang Z; Mu L; Wei G-W, Representability of Algebraic Topology for Biomolecules in Machine Learning Based Scoring and Virtual Screening. PLoS Comp. Biol 2018, 14, e1005929.
13. Jiang P; Chi Y; Li X-S; Liu X; Hua X-S; Xia K, Molecular Persistent Spectral Image (Mol-Psi) Representation for Machine Learning Models in Drug Design. Briefings Bioinf 2022, 23, bbab527.
14. Meng Z; Xia K, Persistent Spectral–Based Machine Learning (Perspect Ml) for Protein-Ligand Binding Affinity Prediction. Sci. Adv 2021, 7, eabc5329. [PubMed: 33962954]
15. 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.
16. Zheng L; Fan J; Mu Y, Onionnet: A Multiple-Layer Intermolecular-Contact-Based Convolutional Neural Network for Protein–Ligand Binding Affinity Prediction. ACS omega 2019, 4, 15956–15965. [PubMed: 31592466]
17. Karlov DS; Sosnin S; Fedorov MV; Popov P, Graphdelta: Mpnn Scoring Function for the Affinity Prediction of Protein–Ligand Complexes. ACS omega 2020, 5, 5150–5159. [PubMed: 32201802]
18. Sánchez-Cruz N; Medina-Franco JL; Mestres J; Barril X, Extended Connectivity Interaction Features: Improving Binding Affinity Prediction through Chemical Description. Bioinformatics 2020.
19. Nguyen DD; Wei G-W, Agl-Score: Algebraic Graph Learning Score for Protein–Ligand Binding Scoring, Ranking, Docking, and Screening. J. Chem. Inf. Model 2019, 59, 3291–3304. [PubMed: 31257871]
20. Xiong G-L; Ye W-L; Shen C; Lu A-P; Hou T-J; Cao D-S, Improving Structure-Based Virtual Screening Performance Via Learning from Scoring Function Components. Briefings Bioinf 2021, 22, bbaa094.
21. Fresnais L; Ballester PJ, The Impact of Compound Library Size on the Performance of Scoring Functions for Structure-Based Virtual Screening. Briefings Bioinf 2021, 22, bbaa095.
22. Boyles F; Deane CM; Morris GM, Learning from the Ligand: Using Ligand-Based Features to Improve Binding Affinity Prediction. Bioinformatics 2020, 36, 758–764. [PubMed: 31598630]
23. Jiménez J; Skalic M; Martinez-Rosell G; De Fabritiis G, K Deep: Protein–Ligand Absolute Binding Affinity Prediction Via 3d-Convolutional Neural Networks. J. Chem. Inf. Model 2018, 58, 287–296. [PubMed: 29309725]
24. Stepniewska-Dziubinska MM; Zielenkiewicz P; Siedlecki P, Development and Evaluation of a Deep Learning Model for Protein–Ligand Binding Affinity Prediction. Bioinformatics 2018, 34, 3666–3674. [PubMed: 29757353]
25. Li J; Fu A; Zhang L, An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. Interdiscip. Sci.: Comput. Life Sci 2019, 11, 320–328.
26. Wallach I; Dzamba M; Heifets A, Atomnet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-Based Drug Discovery. arXiv preprint arXiv:1510.02855 2015.
27. Adeshina YO; Deeds EJ; Karanicolas J, Machine Learning Classification Can Reduce False Positives in Structure-Based Virtual Screening. PNAS 2020, 117, 18477–18488. [PubMed: 32669436]
28. Ghislat G; Rahman T; Ballester PJ, Recent Progress on the Prospective Application of Machine Learning to Structure-Based Virtual Screening. Curr. Opin. Chem. Biol 2021, 65, 28–34. [PubMed: 34052776]
29. Stecula A; Hussain MS; Viola RE, Discovery of Novel Inhibitors of a Critical Brain Enzyme Using a Homology Model and a Deep Convolutional Neural Network. J. Med. Chem 2020, 63, 8867–8875. [PubMed: 32787146]

30. Hsieh C-H; Li L; Vanhauwaert R; Nguyen KT; Davis MD; Bu G; Wszolek ZK; Wang X, Miro1 Marks Parkinson's Disease Subset and Miro1 Reducer Rescues Neuron Loss in Parkinson's Models. Cell Metab 2019, 30, 1131–1140. e7. [PubMed: 31564441]
31. Huang C; Bernard D; Zhu J; Dash RC; Chu A; Knupp A; Hakey A; Hadden MK; Garmendia A; Tang Y, Small Molecules Block the Interaction between Porcine Reproductive and Respiratory Syndrome Virus and Cd163 Receptor and the Infection of Pig Cells. Virol. J 2020, 17, 116. [PubMed: 32727587]
32. Yang Y; Lu J; Yang C; Zhang Y, Exploring Fragment-Based Target-Specific Ranking Protocol with Machine Learning on Cathepsin S. J. Comput. Aided Mol. Des 2019, 33, 1095–1105. [PubMed: 31729618]
33. Wang D; Ding X; Cui C; Xiong Z; Zheng M; Luo X; Jiang H; Chen K, Improving the Virtual Screening Ability of Target-Specific Scoring Functions Using Deep Learning Methods. Front. pharmacol 2019, 10, 924. [PubMed: 31507420]
34. Liu J; Wang R, Classification of Current Scoring Functions. J. Chem. Inf. Model 2015, 55, 475–482. [PubMed: 25647463]
35. Ferrara P; Gohlke H; Price DJ; Klebe G; Brooks CL, Assessing Scoring Functions for Protein−Ligand Interactions. J. Med. Chem 2004, 47, 3032–3047. [PubMed: 15163185]
36. Halperin I; Ma B; Wolfson H; Nussinov R, Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions. Proteins: Struct. Funct. Bioinform 2002, 47, 409–443.
37. Kim R; Skolnick J, Assessment of Programs for Ligand Binding Affinity Prediction. J. Comput. Chem 2008, 29, 1316–1331. [PubMed: 18172838]
38. Perola E; Walters WP; Charifson PS, A Detailed Comparison of Current Docking and Scoring Methods on Systems of Pharmaceutical Relevance. Proteins: Struct. Funct. Bioinform 2004, 56, 235–249.
39. Plewczynski D; Łaźniewski M; Augustyniak R; Ginalski K, Can We Trust Docking Results? Evaluation of Seven Commonly Used Programs on Pdbbind Database. J. Comput. Chem 2011, 32, 742–755. [PubMed: 20812323]
40. Warren GL; Andrews CW; Capelli A-M; Clarke B; LaLonde J; Lambert MH; Lindvall M; Nevins N; Semus SF; Senger S, A Critical Assessment of Docking Programs and Scoring Functions. J. Med. Chem 2006, 49, 5912–5931. [PubMed: 17004707]
41. Wang Z; Sun H; Yao X; Li D; Xu L; Li Y; Tian S; Hou T, Comprehensive Evaluation of Ten Docking Programs on a Diverse Set of Protein–Ligand Complexes: The Prediction Accuracy of Sampling Power and Scoring Power. Phys. Chem. Chem. Phys 2016, 18, 12964–12975. [PubMed: 27108770]
42. Gaillard T, Evaluation of Autodock and Autodock Vina on the Casf-2013 Benchmark. J. Chem. Inf. Model 2018, 58, 1697–1706. [PubMed: 29989806]
43. Friesner RA; Banks JL; Murphy RB; Halgren TA; Klicic JJ; Mainz DT; Repasky MP; Knoll EH; Shelley M; Perry JK, Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. J. Med. Chem 2004, 47, 1739–1749. [PubMed: 15027865]
44. Halgren TA; Murphy RB; Friesner RA; Beard HS; Frye LL; Pollard WT; Banks JL, Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. J. Med. Chem 2004, 47, 1750–1759. [PubMed: 15027866]
45. Trott O; Olson AJ, Autodock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. J. Comput. Chem 2010, 31, 455–461. [PubMed: 19499576]
46. Zilian D; Sotriffer CA, Sfcscore Rf: A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein–Ligand Complexes. J. Chem. Inf. Model 2013, 53, 1923–1933. [PubMed: 23705795]
47. Ragoza M; Hochuli J; Idrobo E; Sunseri J; Koes DR, Protein–Ligand Scoring with Convolutional Neural Networks. J. Chem. Inf. Model 2017, 57, 942–957. [PubMed: 28368587]

48. Ain QU; Aleksandrova A; Roessler FD; Ballester PJ, Machine-Learning Scoring Functions to Improve Structure-Based Binding Affinity Prediction and Virtual Screening. Wiley Interdiscip. Rev. Comput. Mol. Sci 2015, 5, 405–424. [PubMed: 27110292]
49. Wójcikowski M; Ballester PJ; Siedlecki P, Performance of Machine-Learning Scoring Functions in Structure-Based Virtual Screening. Sci. Rep 2017, 7, 46710. [PubMed: 28440302]
50. Ballester PJ; Mitchell JB, A Machine Learning Approach to Predicting Protein–Ligand Binding Affinity with Applications to Molecular Docking. Bioinformatics 2010, 26, 1169–1175. [PubMed: 20236947]
51. Li H; Leung K-S; Wong M-H; Ballester PJ, Substituting Random Forest for Multiple Linear Regression Improves Binding Affinity Prediction of Scoring Functions: Cyscore as a Case Study. BMC Bioinf 2014, 15, 291.
52. 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. [PubMed: 20845954]
53. Durrant JD; McCammon JA, Nnscore 2.0: A Neural-Network Receptor–Ligand Scoring Function. J. Chem. Inf. Model 2011, 51, 2897–2903. [PubMed: 22017367]
54. Lavecchia A, Machine-Learning Approaches in Drug Discovery: Methods and Applications. Drug Discov. Today 2015, 20, 318–331. [PubMed: 25448759]
55. Li H; Leung KS; Wong MH; Ballester PJ, 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. [PubMed: 27490034]
56. Sunseri J; Ragoza M; Collins J; Koes DR, A D3r Prospective Evaluation of Machine Learning for Protein-Ligand Scoring. J. Comput. Aided Mol. Des 2016, 30, 761–771. [PubMed: 27592011]
57. Sunseri J; King JE; Francoeur PG; Koes DR, Convolutional Neural Network Scoring and Minimization in the D3r 2017 Community Challenge. J. Comput. Aided Mol. Des 2019, 33, 19–34. [PubMed: 29992528]
58. Nguyen DD; Cang Z; Wu K; Wang M; Cao Y; Wei G-W, Mathematical Deep Learning for Pose and Binding Affinity Prediction and Ranking in D3r Grand Challenges. J. Comput. Aided Mol. Des 2019, 33, 71–82. [PubMed: 30116918]
59. Gomes J; Ramsundar B; Feinberg EN; Pande VS, Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity. arXiv preprint arXiv:1703.10603 2017.
60. 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, e1429.
61. Su M; Feng G; Liu Z; Li Y; Wang R, Tapping on the Black Box: How Is the Scoring Power of a Machine-Learning Scoring Function Dependent on the Training Set? J. Chem. Inf. Model 2020, 60, 1122–1136. [PubMed: 32085675]
62. Li H; Sze KH; Lu G; Ballester PJ, Machine-Learning Scoring Functions for Structure-Based Drug Lead Optimization. Wiley Interdiscip. Rev. Comput. Mol. Sci 2020, 10, e1465.
63. Li H; Sze KH; Lu G; Ballester PJ, Machine-Learning Scoring Functions for Structure-Based Virtual Screening. Wiley Interdiscip. Rev. Comput. Mol. Sci 2021, 11, e1478.
64. Gabel J; Desaphy J; Rognan D, Beware of Machine Learning-Based Scoring Functions on the Danger of Developing Black Boxes. J. Chem. Inf. Model 2014, 54, 2807–2815. [PubMed: 25207678]
65. 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. [PubMed: 24528282]
66. Wang C; Zhang Y, Improving Scoring-Docking-Screening Powers of Protein–Ligand Scoring Functions Using Random Forest. J. Comput. Chem 2017, 38, 169–177. [PubMed: 27859414]
67. Lu J; Hou X; Wang C; Zhang Y, Incorporating Explicit Water Molecules and Ligand Conformation Stability in Machine-Learning Scoring Functions. J. Chem. Inf. Model 2019, 59, 4540–4549. [PubMed: 31638801]
68. Yang C; Zhang Y, Lin_F9: A Linear Empirical Scoring Function for Protein–Ligand Docking. J. Chem. Inf. Model 2021, 61, 4630–4644. [PubMed: 34469692]

69. Fischer C; Vepřek NA; Peitsinis Z; Rühmann K-P; Yang C; Spradlin JN; Dovala D; Nomura DK; Zhang Y; Trauner D, De Novo Design of Sars-Cov-2 Main Protease Inhibitors. Synlett 2022, 33, 458–463.
70. Chen T; Guestrin C Xgboost: A Scalable Tree Boosting System In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016; pp 785–794.
71. Liu Z; Su M; Han L; Liu J; Yang Q; Li Y; Wang R, Forging the Basis for Developing Protein–Ligand Interaction Scoring Functions. Acc. Chem. Res 2017, 50, 302–309. [PubMed: 28182403]
72. Liu Z; Li Y; Han L; Li J; Liu J; Zhao Z; Nie W; Liu Y; Wang R, Pdb-Wide Collection of Binding Data: Current Status of the Pdbbind Database. Bioinformatics 2015, 31, 405–412. [PubMed: 25301850]
73. Gilson MK; Liu T; Baitaluk M; Nicola G; Hwang L; Chong J, Bindingdb in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. Nucleic Acids Res 2016, 44, D1045–D1053. [PubMed: 26481362]
74. Liu T; Lin Y; Wen X; Jorissen RN; Gilson MK, Bindingdb: A Web-Accessible Database of Experimentally Determined Protein–Ligand Binding Affinities. Nucleic Acids Res 2007, 35, D198–D201. [PubMed: 17145705]
75. Tran-Nguyen V-K; Jacquemard C; Rognan D, Lit-Pcba: An Unbiased Data Set for Machine Learning and Virtual Screening. J. Chem. Inf. Model 2020, 60, 4263–4273. [PubMed: 32282202]
76. Dunbar JB Jr; Smith RD; Yang C-Y; Ung PM-U; Lexa KW; Khazanov NA; Stuckey JA; Wang S; Carlson HA, Csar Benchmark Exercise of 2010: Selection of the Protein–Ligand Complexes. J. Chem. Inf. Model 2011, 51, 2036–2046. [PubMed: 21728306]
77. Huang S-Y; Zou X, Scoring and Lessons Learned with the Csar Benchmark Using an Improved Iterative Knowledge-Based Scoring Function. J. Chem. Inf. Model 2011, 51, 2097–2106. [PubMed: 21830787]
78. Clark M; Cramer III RD; Van Opdenbosch N, Validation of the General Purpose Tripos 5.2 Force Field. J. Comput. Chem 1989, 10, 982–1012.
79. Jiang L; Rizzo RC, Pharmacophore-Based Similarity Scoring for Dock. J. Phys. Chem. B 2015, 119, 1083–1102. [PubMed: 25229837]
80. Sanner MF; Olson AJ; Spehner JC, Reduced Surface: An Efficient Way to Compute Molecular Surfaces. Biopolymers 1996, 38, 305–320. [PubMed: 8906967]
81. Koes DR; Baumgartner MP; Camacho CJ, Lessons Learned in Empirical Scoring with Smina from the Csar 2011 Benchmarking Exercise. J. Chem. Inf. Model 2013, 53, 1893–1904. [PubMed: 23379370]
82. Wang R; Lai L; Wang S, Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. J. Comput. Aided Mol. Des 2002, 16, 11–26. [PubMed: 12197663]
83. Katigbak J; Li H; Rooklin D; Zhang Y, Alphaspace 2.0: Representing Concave Biomolecular Surfaces Using B-Clusters. J. Chem. Inf. Model 2020, 60, 1494–1508. [PubMed: 31995373]
84. Rooklin D; Wang C; Katigbak J; Arora PS; Zhang Y, Alphaspace: Fragment-Centric Topographical Mapping to Target Protein–Protein Interaction Interfaces. J. Chem. Inf. Model 2015, 55, 1585–1599. [PubMed: 26225450]
85. Bell EW; Zhang Y, Dockrmsd: An Open-Source Tool for Atom Mapping and Rmsd Calculation of Symmetric Molecules through Graph Isomorphism. J. Cheminformatics 2019, 11, 1–9.
86. Parks CD; Gaieb Z; Chiu M; Yang H; Shao C; Walters WP; Jansen JM; McGaughey G; Lewis RA; Bembenek SD, D3r Grand Challenge 4: Blind Prediction of Protein–Ligand Poses, Affinity Rankings, and Relative Binding Free Energies. J. Comput. Aided Mol. Des 2020, 34, 99–119. [PubMed: 31974851]
87. Gaieb Z; Parks CD; Chiu M; Yang H; Shao C; Walters WP; Lambert MH; Nevins N; Bembenek SD; Ameriks MK, D3r Grand Challenge 3: Blind Prediction of Protein–Ligand Poses and Affinity Rankings. J. Comput. Aided Mol. Des 2019, 33, 1–18. [PubMed: 30632055]
88. Gaieb Z; Liu S; Gathiaka S; Chiu M; Yang H; Shao C; Feher VA; Walters WP; Kuhn B; Rudolph MG, D3r Grand Challenge 2: Blind Prediction of Protein–Ligand Poses, Affinity Rankings, and Relative Binding Free Energies. J. Comput. Aided Mol. Des 2018, 32, 1–20. [PubMed: 29204945]

89. Gathiaka S; Liu S; Chiu M; Yang H; Stuckey JA; Kang YN; Delproposto J; Kubish G; Dunbar JB; Carlson HA, D3r Grand Challenge 2015: Evaluation of Protein–Ligand Pose and Affinity Predictions. J. Comput. Aided Mol. Des 2016, 30, 651–668. [PubMed: 27696240]

90. Tran-Nguyen V-K; Bret G; Rognan D, True Accuracy of Fast Scoring Functions to Predict High-Throughput Screening Data from Docking Poses: The Simpler the Better. J. Chem. Inf. Model 2021.

91. Tosco P; Stiefl N; Landrum G, Bringing the Mmff Force Field to the Rdkit: Implementation and Validation. J. Cheminformatics 2014, 6, 1–4.

92. Bento AP; Hersey A; Félix E; Landrum G; Gaulton A; Atkinson F; Bellis LJ; De Veij M; Leach AR, An Open Source Chemical Structure Curation Pipeline Using Rdkit. J. Cheminformatics 2020, 12, 1–16.

93. Meli R; Anighoro A; Bodkin MJ; Morris GM; Biggin PC, Learning Protein-Ligand Binding Affinity with Atomic Environment Vectors. J. Cheminformatics 2021, 13, 1–19.

94. Kwon Y; Shin W-H; Ko J; Lee J, Ak-Score: Accurate Protein-Ligand Binding Affinity Prediction Using an Ensemble of 3d-Convolutional Neural Networks. Int. J. Mol. Sci 2020, 21, 8424.

95. Méndez-Lucio O; Ahmad M; del Rio-Chanona EA; Wegner JK, A Geometric Deep Learning Approach to Predict Binding Conformations of Bioactive Molecules. Nat. Mach. Intell 2021, 3, 1033–1039.

96. Vassar R, Bace1 Inhibitor Drugs in Clinical Trials for Alzheimer's Disease. Alzheimer's Res. Ther 2014, 6, 89. [PubMed: 25621019]

97. Hsiao C-C; Rombouts F; Gijsen HJ, New Evolutions in the Bace1 Inhibitor Field from 2014 to 2018. Bioorg. Med. Chem. Lett 2019, 29, 761–777. [PubMed: 30709653]

98. Moussa-Pacha NM; Abdin SM; Omar HA; Alniss H; Al-Tel TH, Bace1 Inhibitors: Current Status and Future Directions in Treating Alzheimer's Disease. Med. Res. Rev 2020, 40, 339–384. [PubMed: 31347728]

99. Jeppsson F; Eketjäll S; Janson J; Karlström S; Gustavsson S; Olsson L-L; Radesäter A-C; Ploeger B; Cebers G; Kolmodin K, Discovery of Azd3839, a Potent and Selective Bace1 Inhibitor Clinical Candidate for the Treatment of Alzheimer Disease. J. Biol. Chem 2012, 287, 41245–41257. [PubMed: 23048024]

100. Wilkinson RD; Williams R; Scott CJ; Burden RE, Cathepsin S: Therapeutic, Diagnostic, and Prognostic Potential. Biol. Chem 2015, 396, 867–882. [PubMed: 25872877]

101. Ahmad S; Bhagwati S; Kumar S; Banerjee D; Siddiqi MI, Molecular Modeling Assisted Identification and Biological Evaluation of Potent Cathepsin S Inhibitors. J. Mol. Graphics Model 2020, 96, 107512.

102. Wang B; Ng H-L, Deep Neural Network Affinity Model for Bace Inhibitors in D3r Grand Challenge 4. J. Comput. Aided Mol. Des 2020, 34, 201–217. [PubMed: 31916049]

103. Basciu A; Koukos PI; Malloci G; Bonvin AM; Vargiu AV, Coupling Enhanced Sampling of the Apo-Receptor with Template-Based Ligand Conformers Selection: Performance in Pose Prediction in the D3r Grand Challenge 4. J. Comput. Aided Mol. Des 2020, 34, 149–162. [PubMed: 31720895]

104. Elisée E; Gapsys V; Mele N; Chaput L; Selwa E; de Groot BL; Iorga BI, Performance Evaluation of Molecular Docking and Free Energy Calculations Protocols Using the D3r Grand Challenge 4 Dataset. J. Comput. Aided Mol. Des 2019, 33, 1031–1043. [PubMed: 31677003]

105. Martin SJ; Chen I-J; Chan AE; Foloppe N, Modelling the Binding Mode of Macrocycles: Docking and Conformational Sampling. Biorg. Med. Chem 2020, 28, 115143.

106. Kotelnikov S; Alekseenko A; Liu C; Ignatov M; Padhorny D; Brini E; Lukin M; Coutsias E; Dill KA; Kozakov D, Sampling and Refinement Protocols for Template-Based Macrocycle Docking: 2018 D3r Grand Challenge 4. J. Comput. Aided Mol. Des 2020, 34, 179–189. [PubMed: 31879831]

107. Lam PC-H; Abagyan R; Totrov M, Macrocycle Modeling in Icm: Benchmarking and Evaluation in D3r Grand Challenge 4. J. Comput. Aided Mol. Des 2019, 33, 1057–1069. [PubMed: 31598897]

108. Nguyen DD; Gao K; Wang M; Wei G-W, Mathdl: Mathematical Deep Learning for D3r Grand Challenge 4. J. Comput. Aided Mol. Des 2020, 34, 131–147. [PubMed: 31734815]

109. El Khoury L; Santos-Martins D; Sasmal S; Eberhardt J; Bianco G; Ambrosio FA; Solis-Vasquez L; Koch A; Forli S; Mobley DL, Comparison of Affinity Ranking Using Autodock-Gpu and Mm-Gbsa Scores for Bace-1 Inhibitors in the D3r Grand Challenge 4. J. Comput. Aided Mol. Des 2019, 33, 1011–1020. [PubMed: 31691919]
110. Mendez D; Gaulton A; Bento AP; Chambers J; De Veij M; Félix E; Magariños MP; Mosquera JF; Mutowo P; Nowotka M, Chembl: Towards Direct Deposition of Bioassay Data. Nucleic Acids Res 2019, 47, D930–D940. [PubMed: 30398643]
111. Gaulton A; Hersey A; Nowotka M; Bento AP; Chambers J; Mendez D; Mutowo P; Atkinson F; Bellis LJ; Cibrián-Uhalte E, The Chembl Database in 2017. Nucleic Acids Res 2017, 45, D945–D954. [PubMed: 27899562]
112. Gaulton A; Bellis LJ; Bento AP; Chambers J; Davies M; Hersey A; Light Y; McGlinchey S; Michalovich D; Al-Lazikani B, Chembl: A Large-Scale Bioactivity Database for Drug Discovery. Nucleic Acids Res 2012, 40, D1100–D1107. [PubMed: 21948594]
113. Huang N; Shoichet BK; Irwin JJ, Benchmarking Sets for Molecular Docking. J. Med. Chem 2006, 49, 6789–6801. [PubMed: 17154509]
114. Mysinger MM; Carchia M; Irwin JJ; Shoichet BK, Directory of Useful Decoys, Enhanced (Dud-E): Better Ligands and Decoys for Better Benchmarking. J. Med. Chem 2012, 55, 6582–6594. [PubMed: 22716043]
115. Wang Y; Xiao J; Suzek TO; Zhang J; Wang J; Zhou Z; Han L; Karapetyan K; Dracheva S; Shoemaker BA, Pubchem's Bioassay Database. Nucleic Acids Res 2012, 40, D400–D412. [PubMed: 22140110]
116. Wang Y; Bolton E; Dracheva S; Karapetyan K; Shoemaker BA; Suzek TO; Wang J; Xiao J; Zhang J; Bryant SH, An Overview of the Pubchem Bioassay Resource. Nucleic Acids Res 2010, 38, D255–D266. [PubMed: 19933261]
117. Jain AN, Surflex-Dock 2.1: Robust Performance from Ligand Energetic Modeling, Ring Flexibility, and Knowledge-Based Search. J. Comput. Aided Mol. Des 2007, 21, 281–306. [PubMed: 17387436]
118. Marcou G; Rognan D, Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. J. Chem. Inf. Model 2007, 47, 195–207. [PubMed: 17238265]
119. Desaphy J; Raimbaud E; Ducrot P; Rognan D, Encoding Protein–Ligand Interaction Patterns in Fingerprints and Graphs. J. Chem. Inf. Model 2013, 53, 623–637. [PubMed: 23432543]
120. Zhou H; Cao H; Skolnick J, Fragsite: A Fragment-Based Approach for Virtual Ligand Screening. J. Chem. Inf. Model 2021, 61, 2074–2089. [PubMed: 33724022]
121. Sunseri J; Koes DR, Virtual Screening with Gnina 1.0. Molecules 2021, 26, 7369. [PubMed: 34885952]
122. McNutt AT; Francoeur P; Aggarwal R; Masuda T; Meli R; Ragoza M; Sunseri J; Koes DR, Gnina 1.0: Molecular Docking with Deep Learning. J. Cheminformatics 2021, 13, 1–20.
123. Li H; Leung K-S; Wong M-H; Ballester PJ, Correcting the Impact of Docking Pose Generation Error on Binding Affinity Prediction. BMC Bioinf 2016, 17, 13–25.
124. Quiroga R; Villarreal MA, Vinardo: A Scoring Function Based on Autodock Vina Improves Scoring, Docking, and Virtual Screening. PloS one 2016, 11, e0155183. [PubMed: 27171006]
125. Shen C; Weng G; Zhang X; Leung EL-H; Yao X; Pang J; Chai X; Li D; Wang E; Cao D, Accuracy or Novelty: What Can We Gain from Target-Specific Machine-Learning-Based Scoring Functions in Virtual Screening? Briefings Bioinf 2021, 22, bbaa410.
126. O'Boyle NM; Banck M; James CA; Morley C; Vandermeersch T; Hutchison GR, Open Babel: An Open Chemical Toolbox. J. Cheminformatics 2011, 3, 33.
127. Rizvi SMD; Shakil S; Haneef M, A Simple Click by Click Protocol to Perform Docking: Autodock 4.2 Made Easy for Non-Bioinformaticians. EXCLI J 2013, 12, 831. [PubMed: 26648810]

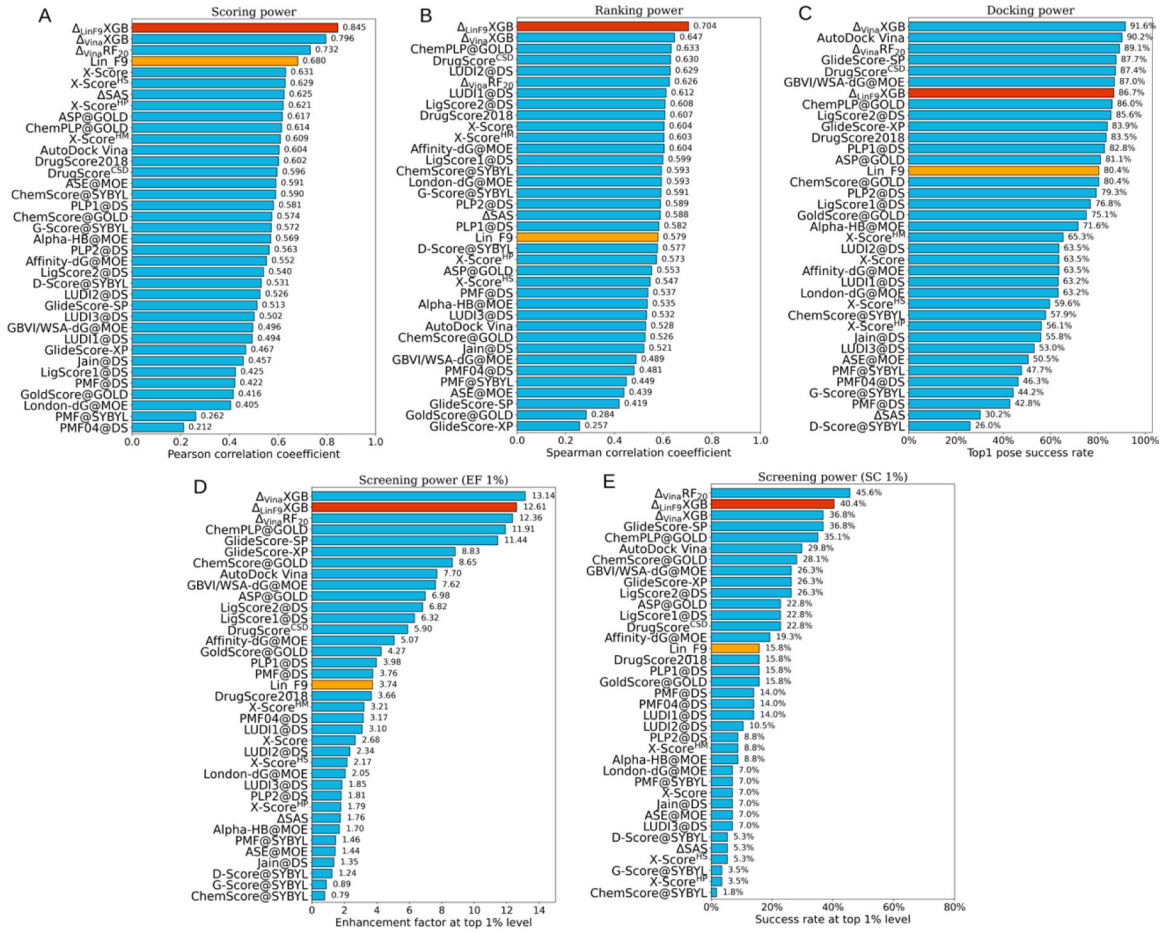


Figure 1.

Performances of scoring functions on CASF-2016 benchmark. (A) Scoring power evaluated by Pearson correlation coefficient, (B) ranking power measured by Spearman correlation coefficient, (C) docking power calculated by success rate for top1 poses (include crystal structures), screening power measured by (D) enhancement factor and (E) success rate at top 1% level. Performances of $\Delta_{Lin\_F9}XGB$ are colored red, performances of Lin_F9 are colored orange and all other scoring functions are colored cyan. All scoring functions are ranked in a descending order.

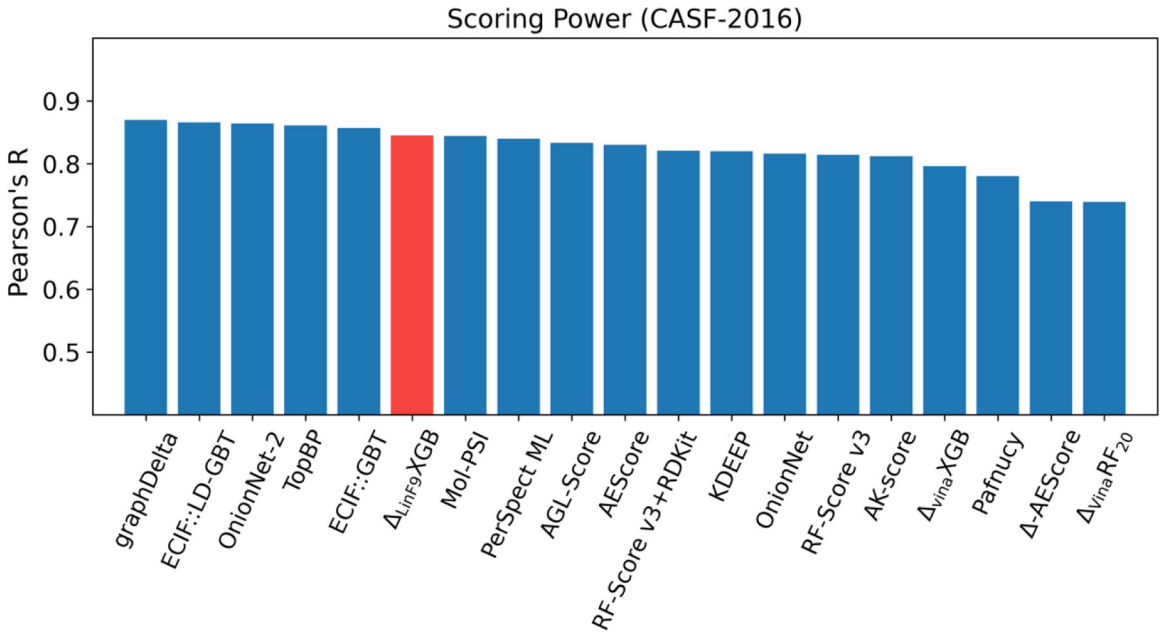


Figure 2.

Scoring power comparison of several state-of-the-art ML scoring functions on CASF-2016 benchmark. The Pearson correlation coefficients of other ML scoring functions are taken from refs 12–19, 22–24, 66–67, 95–96.

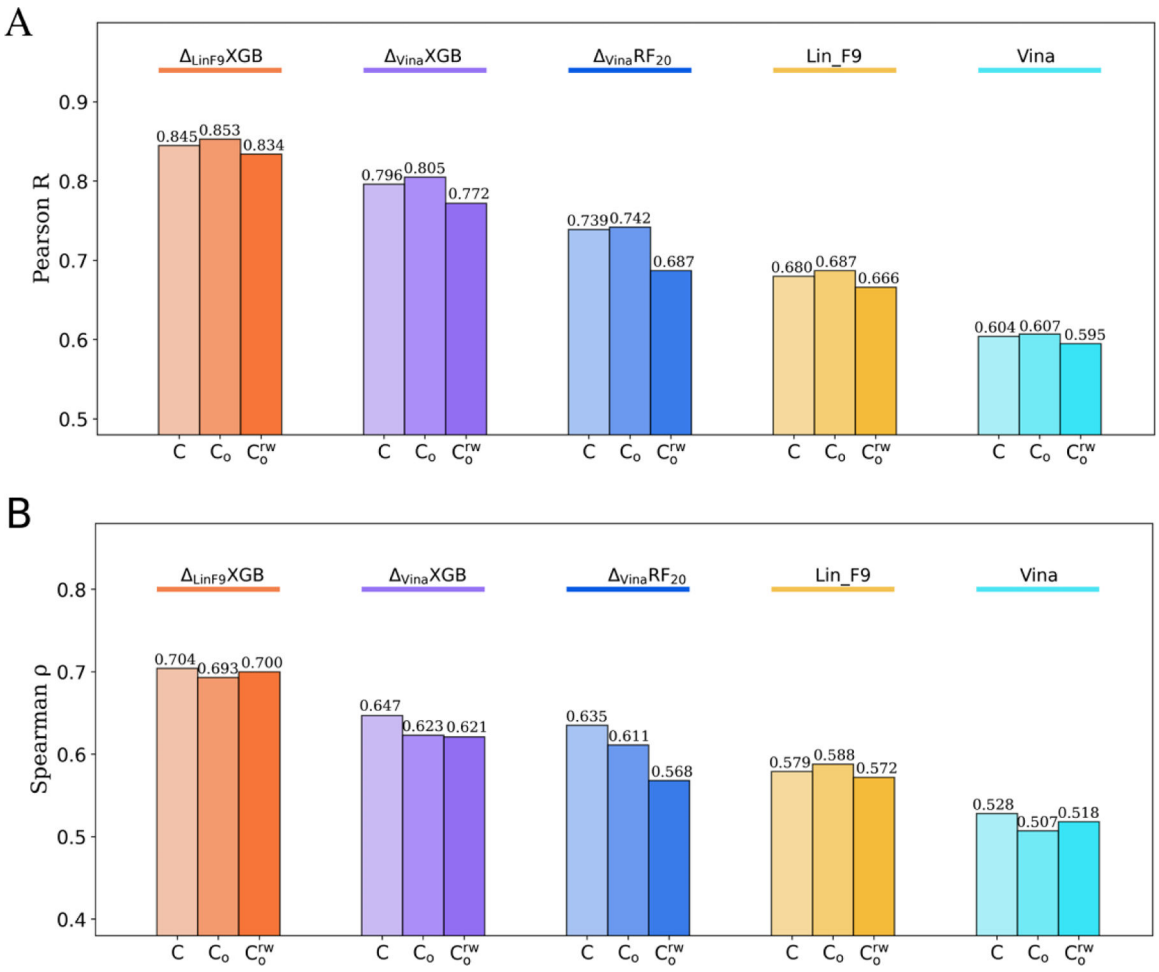


Figure 3.

Scoring and ranking performances of $\Delta_{Lin\_F9}XGB$, $\Delta_{Vina}XGB$, $\Delta_{Vina}RF_{20}$, Lin_F9 and Vina on LocalOpt pose, as well as crystal pose. (A) Pearson correlation coefficient used to measure scoring power. (B) Spearman correlation coefficient used for ranking power. Performances of $\Delta_{Lin\_F9}XGB$, $\Delta_{Vina}XGB$, $\Delta_{Vina}RF_{20}$, Lin_F9 and Vina are colored red, purple, blue, orange and cyan, respectively. For each scoring function, performance on crystal pose (C), LocalOpt pose (C_o), and local optimized pose with receptor-bound water molecules (C_o^{rw}) are displayed from left to right with gradually changed color.

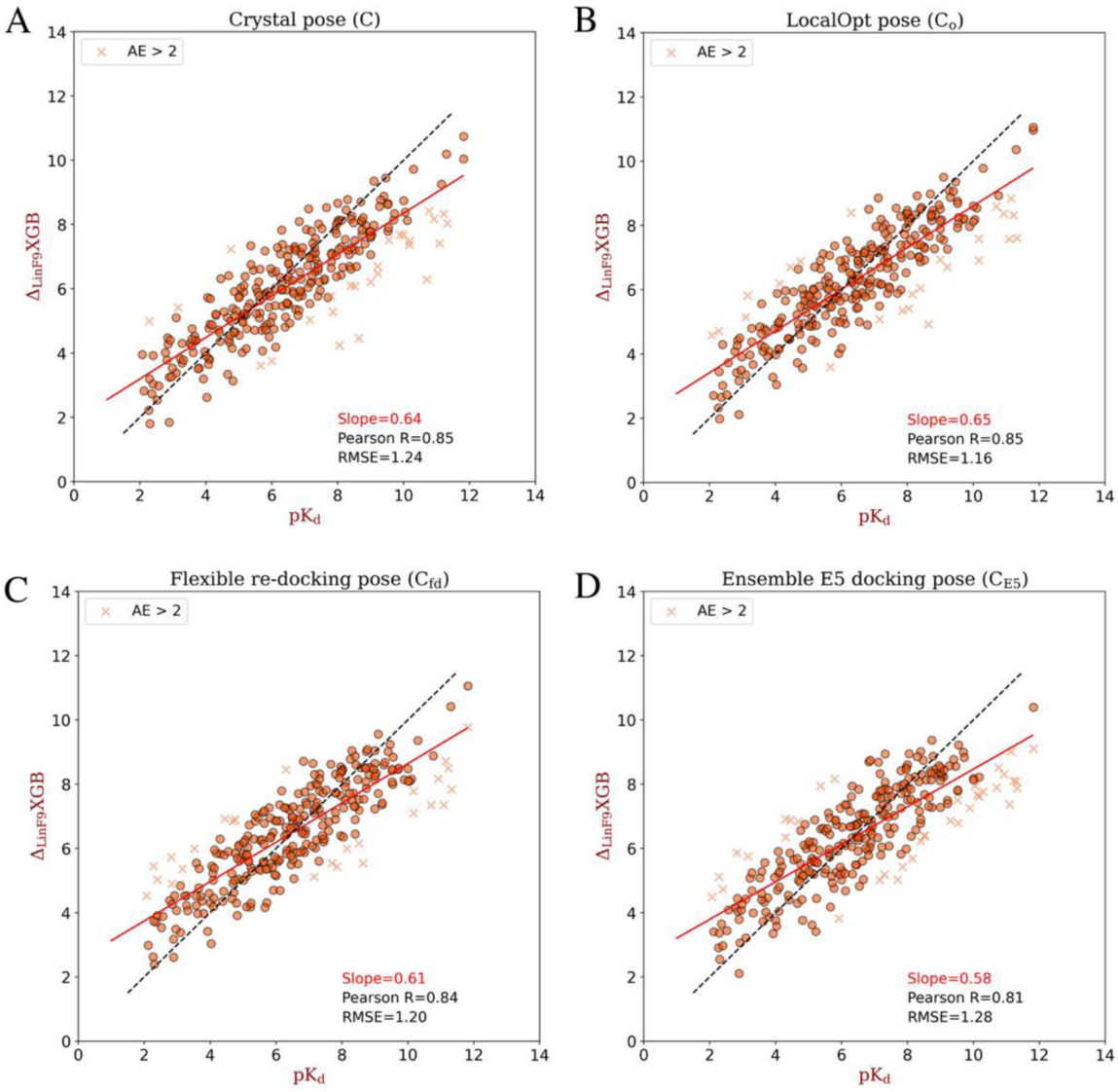


Figure 4.

Scatter plots between experimental pK_d and predicted pK_d of different poses re-scored by $\Delta_{Lin\_F9}XGB$. (A) Crystal pose, (B) locally optimized pose, (C) flexible re-docking pose and (D) ensemble E5 docking pose. The absolute error (AE) in pK_d larger than 2 are plotted with marker "x", and others are plotted with marker "o". Pearson correlation coefficient (R) and root-mean-square error (RMSE) between predicted pK_d and experimental pK_d are shown for each plot. The solid red line for each plot corresponds to the linear fit between predicted pK_d and experimental pK_d, the slope value for this linear fit is shown in the plot.

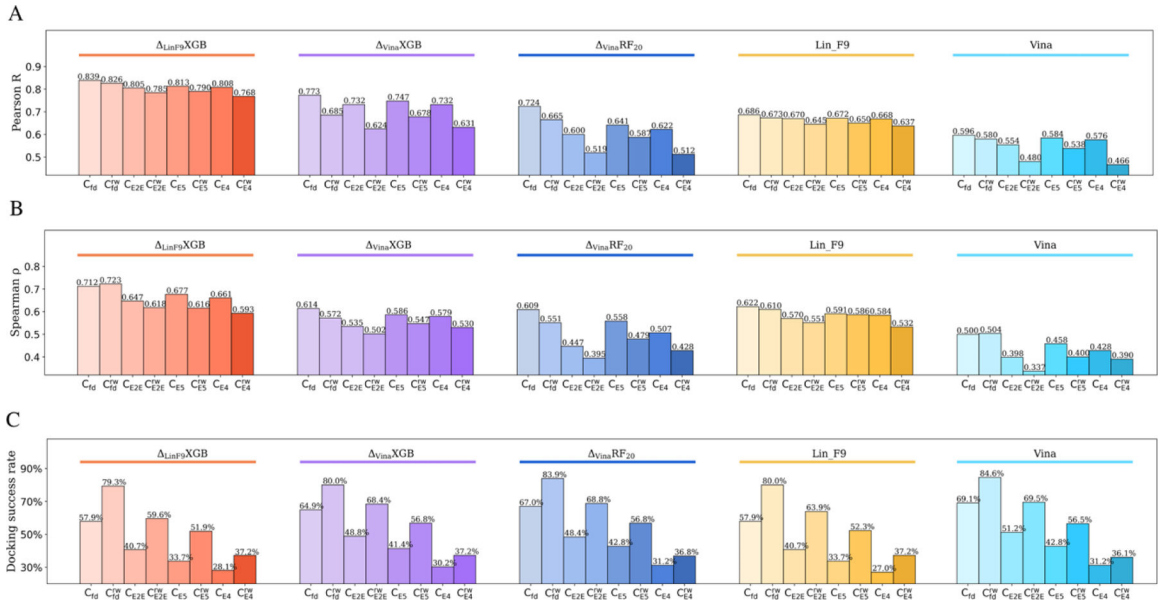


Figure 5.

Scoring, ranking and docking powers of $\Delta_{Lin\_F9}XGB$, $\Delta_{Vina}XGB$, $\Delta_{Vina}RF_{20}$, Lin_F9 and Vina for different docking tests on CASF-2016 core set. (A) Pearson correlation coefficient used to measure the scoring power. (B) Spearman correlation coefficient for ranking power. (C) Docking power measured by success rate of best-scored pose (RMSD < 2 Å). Performances of $\Delta_{Lin\_F9}XGB$, $\Delta_{Vina}XGB$, $\Delta_{Vina}RF_{20}$, Lin_F9 and Vina are colored red, purple, blue, orange and cyan, respectively. For each scoring function, performances on flexible re-docking poses (C_{fd} and C_{fd}^{rw}), E2E docking poses (C_{E2E} and C_{E2E}^{rw}), ensemble (E5) docking poses (C_{E5} and C_{E5}^{rw}), ensemble (E4) docking poses (C_{E4} and C_{E4}^{rw}) are displayed from left to right with gradually changed color.

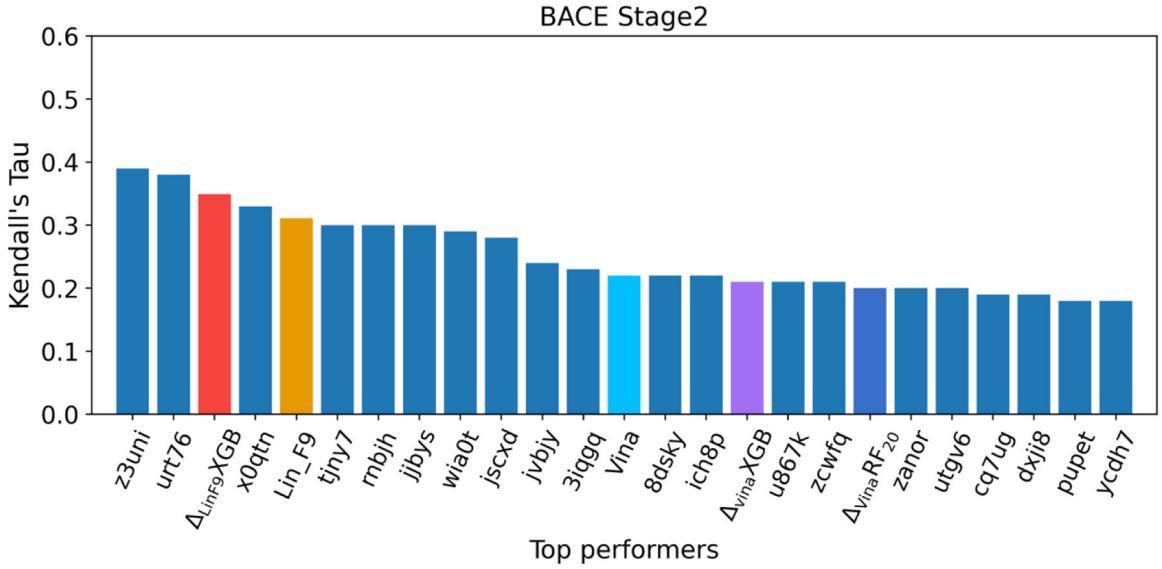


Figure 6.

Affinity ranking performances of top 20 performers in D3R GC4 as well as five scoring functions (Vina, $\Delta_{Vina}RF_{20}$, $\Delta_{Vina}XGB$, Lin_F9,$\Delta_{Lin\_F9}XGB$) for the BACE1 Stage 2. Ranking power is evaluated by Kendall rank correlation coefficient.

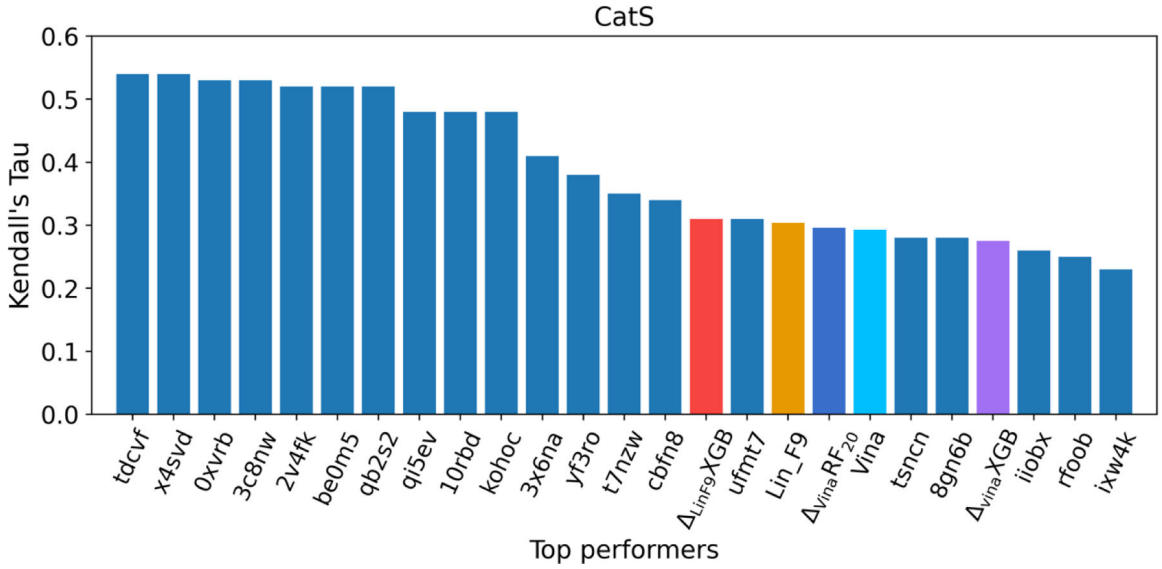


Figure 7.

Affinity ranking performances of top 20 performers in D3R GC4 as well as five scoring functions (Vina, $\Delta_{Vina}RF_{20}$, $\Delta_{Vina}XGB$, Lin_F9,$\Delta_{Lin\_F9}XGB$) for the CatS dataset. Ranking power is evaluated by Kendall rank correlation coefficient.

Table 1.

Different docking-scoring tests of CASF-2016 benchmark. The details are described in Ref. 68.

Name	Ligand conformation for each ligand	Protein conformation for each ligand	Docking method
flexible re-docking	native ligand pose	native protein structure	flexible ligand docking
E2E docking	Computer-generated maximum 10 conformers[a]	native protein structure	flexible ligand docking
ensemble (E5) docking	Computer-generated maximum 10 conformers	5 protein structures (include native protein structure)	flexible ligand docking
ensemble (E4) docking	Computer-generated maximum 10 conformers	4 non-native protein structures	flexible ligand docking

[a]Computer-generated maximum 10 conformers per ligand using OpenBabel.

Table 2.

CASF-2016 benchmark test results for several ML scoring functions. (The highest values of each column are shown in bold)

Model	CASF-2016 Metrics				
	Scoring	Ranking	Docking	Screening	
	Pearson R	Spearman ρ	Success Rate at top1 pose	EF at top 1%	Success Rate at top 1%
$\Delta_{Vina}RF_{20}$ [66]	0.739	0.635	89.1%	12.36	**45.6%**
$\Delta_{Vina}XGB$ [67]	0.796	0.647	**91.6%**	13.14	36.8%
$\Delta_{Lin\_F9}XGB$	**0.845**	**0.704**	86.7%	12.61	40.4%
$\Delta - AEScore$[93]	0.740	0.590	85.6%	6.16	19.3%
AEScore[93]	0.830	0.640	35.8%	–	–
AK-score (ensemble)[94]	0.812	0.670	36.0%	–	–
DeepDock[95]	–	–	87.0%	**16.41**	43.9%

Table 3.

Scoring and ranking performances of Vina, $\Delta_{Vina}RF_{20}$, $\Delta_{Vina}XGB$, Lin_F9, $\Delta_{Lin\_F9}XGB$ on BACE1 dataset.

Scoring functions	Pearson R	Spearman ρ	Kendall τ	RMSE
Vina	0.334	0.332	0.222	1.842
$\Delta_{Vina}RF_{20}$	0.293	0.299	0.201	2.441
$\Delta_{Vina}XGB$	0.345	0.307	0.211	1.790
Lin_F9	0.481	0.439	0.311	1.950
$\Delta_{Lin\_F9}XGB$	**0.517**	**0.481**	**0.349**	**1.518**

Table 4.

Scoring and ranking performances of Vina, $\Delta_{Vina}RF_{20}$, $\Delta_{Vina}XGB$, Lin_F9, $\Delta_{Lin\_F9}XGB$ on CatS dataset.

Scoring functions	Pearson R	Spearman ρ	Kendall τ	RMSE
Vina	0.427	0.430	0.293	0.841
$\Delta_{Vina}RF_{20}$	0.455	0.430	0.296	0.621
$\Delta_{Vina}XGB$	0.441	0.399	0.275	0.657
Lin_F9	0.451	0.446	0.304	0.680
$\Delta_{Lin\_F9}XGB$	**0.464**	**0.457**	**0.309**	**0.611**

Table 5.

Screening performance comparison of Vina, $\Delta_{Vina}RF_{20}$, Lin_F9 and $\Delta_{Lin\_F9}XGB$ on LIT-PCBA dataset.[a]

Target set	Scoring Function				PDB Templates	Number of Actives	Number of Inactives
	Vina	$\Delta_{Vina}RF_{20}$	Lin_F9	$\Delta_{Lin\_F9}XGB$			
ADRB2*	0	0	0	11.76	4	17	312,433
ALDH1	1.49	1.66	1.58	6.46	2	7167	137,822
ESR1-ago*	15.38	15.38	0	7.69	**15**	13	5,582
ESR1-ant*	3.92	2.94	2.94	3.92	**15**	102	4,947
FEN1	0.54	0.81	1.90	2.17	**1**	369	355,323
GBA	4.82	6.63	7.23	9.64	3	166	294,202
IDH1	0	0	2.56	5.13	10	39	361,691
KAT2A	0.52	0.52	2.06	7.73	1	194	348,257
MAPK1*	2.92	1.95	1.62	2.60	**15**	308	62,522
MTORC1*	2.06	3.09	2.06	2.06	**11**	97	32,972
OPRK1*	0	0	4.17	12.5	**1**	24	269,776
PKM2	1.65	2.93	0.73	2.56	2	546	245,485
PPARG*	7.41	11.11	3.70	7.41	**15**	27	5,210
TP53*	0	0	2.53	1.27	**6**	79	4,168
VDR	1.02	0.68	0.11	0.34	1	882	355,094
Average	**2.78**	**3.18**	**2.21**	**5.55**			
EF1% > 2	6	6	8	13			
EF1% > 5	2	3	1	8			
EF1% > 10	1	2	0	2			

[a]Enrichment factor at top 1% (EF1%) is used as the quantitative indicator to evaluate the screening performance for each target set. The average EF1% over all 15 targets are highlighted in bold. PDB templates same as the original benchmark used are highlighted in green color. The 8 targets using cell-based phenotypic assays are marked with *.

Table 6.

Collected LIT-PCBA benchmark test results from different groups.

Model	Average EF1%	Number of Targets (EF1% > 2)	Number of Targets (EF1% > 5)	Number of Targets (EF1% > 10)	References
FINDSITE$^{comb2.0}$	3.04	5	4	1	Zhou et al[120]
FRAGSITE	4.78	11	5	1	
RFScore-4	1.67	4	1	0	Sunseri et al[121]
RFScore-VS	1.75	5	2	0	
Vina	1.71	6	1	0	
Vinardo	1.70	4	2	0	
CNN Default (Affinity)	4.64	6	6	2	
Surflex	2.51	6	3	0	Tran-Nguyen et al[90]
Pafnucy	5.32	9	7	3	
$\Delta_{Vina}RF_{20}$	5.38	10	7	3	
IFP	**7.46**	11	**9**	4	
GRIM	6.87	12	8	**5**	
Vina	2.78	6	2	1	Our test
$\Delta_{Vina}RF_{20}$	3.18	6	3	2	
Lin_F9	2.21	8	1	0	
$\Delta_{Lin\_F9}XGB$	5.55	**13**	8	2	