JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

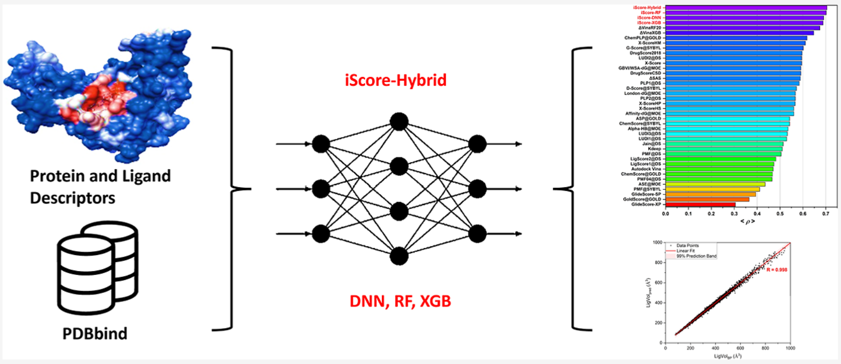
iScore: A ML-Based Scoring Function for De Novo Drug Discovery

Sayyed Jalil Mahdizadeh and Leif A. Eriksson*

Cite This: J. Chem. Inf. Model. 2025, 65, 2759–2772

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: In the quest for accelerating de novo drug discovery, the development of efficient and accurate scoring functions represents a fundamental challenge. This study introduces iScore, a novel machine learning (ML)-based scoring function designed to predict the binding affinity of protein–ligand complexes with remarkable speed and precision. Uniquely, iScore circumvents the conventional reliance on explicit knowledge of protein–ligand interactions and a full picture of atomic contacts, instead leveraging a set of ligand and binding pocket descriptors to directly evaluate binding affinity. This approach enables skipping the inefficient and slow conformational sampling stage, thereby enabling the rapid screening of ultrahuge molecular libraries, a crucial advancement given the practically infinite dimensions of chemical space. iScore was rigorously trained and validated using the PDBbind 2020 refined set, CASF 2016, CSAR NRC-HiQ Set1/2, DUD-E, and target fishing data sets, employing three distinct ML methodologies: Deep neural network (iScore-DNN), random forest (iScore-RF), and eXtreme gradient boosting (iScore-XGB). A hybrid model, iScore-Hybrid, was subsequently developed to incorporate the strengths of these individual base learners. The hybrid model demonstrated a Pearson correlation coefficient (R) of 0.78 and a root-mean-square error (RMSE) of 1.23 in cross-validation, outperforming the individual base learners and establishing new benchmarks for scoring power ($R = 0.814$, RMSE = 1.34), ranking power ($\rho = 0.705$), and screening power (success rate at top 10% = 73.7%). Moreover, iScore-Hybrid demonstrated great performance in the target fishing benchmarking study.

1. INTRODUCTION

Molecular docking is undoubtedly the most widely used technique in structure-based computer-aided drug discovery that aims to predict the binding mode and binding affinity of small organic molecules toward a target protein.[1] The performance (speed and accuracy) of a molecular docking program strongly depends on its two main components, sampling and scoring.[2] Sampling refers to a search algorithm that evaluates a finite number of ligand conformations within and around the binding site of a target protein to elucidate the ligand binding mode. Scoring refers to a class of computational methods, called scoring functions, that are formulated to predict the binding affinity of each ligand conformation within the protein binding site.[3] The performance of a scoring function can be determined by three evaluation metrics:[4] "scoring power" that indicates the degree of correlation in the predicted versus experimentally determined binding affinity values, "ranking power" that is the capability of scoring function for accurately ranking a given set of active ligands with respect to their predicted binding affinity values, toward a particular protein target, and "screening power" that refers to the ability of scoring function to identify the true ligand with the highest affinity against a given protein target among a set of random decoy molecules. While the scoring part of a typical molecular docking calculation is relatively fast, the sampling part is time-consuming, computationally expensive, and inefficient.[5] Therefore, in the traditional molecular docking and virtual screening, the calculation time and cost scale exponentially with an increasing number and degree of

Received: November 26, 2024
Revised: February 25, 2025
Accepted: February 25, 2025
Published: March 4, 2025

https://doi.org/10.1021/acs.jcim.4c02192

freedom of molecules under evaluation. On the other hand, the size of available molecular databases for virtual screening remains extremely limited, ranging from several million to a few billion molecules. This represents only a tiny fraction of the actual chemical space, which is estimated to contain up to $\sim 10^{60}$ feasible drug-like molecules.[6]

Traditional scoring functions can be categorized in three main classes based on the way they are formulated: force-field based, empirical, and knowledge-based scoring functions.[7] Despite significant improvements in the past decade, several recent studies clearly show that the performance of traditional scoring functions is quite limited in both scoring power and ranking power aspects.[8] On the other hand, the most successful scoring approaches, such as free-energy perturbation (FEP) techniques,[9] are very sensitive to the force field selection and ligand parametrization. Moreover, the widespread application of FEP methods has been significantly constrained by their high computational demands, even for small molecular libraries. However, recent breakthroughs in machine learning (ML) algorithms and big data mining, coupled with the exponential growth of computing power, have paved the way for promising applications of ML-driven approaches in the development of novel scoring functions.[10] ML-based scoring functions have demonstrated remarkable performance in various benchmarking studies.[11] They are also typically several orders of magnitude faster than traditional scoring functions. Stepniewska-Dziubinska et al.[12] developed a deep neural network (DNN), Pafnucy, to estimate the binding affinity of ligand−receptor complexes. The model represents the complex using a 3D grid and applies 3D convolution to generate a feature map from this representation, treating atoms from both protein and ligand equivalently. Their model was tested on the CASF-2013 and CASF-2016 "scoring power" benchmark showing good performance with a Pearson correlation coefficient/root mean squared error (RMSE) of 0.70/1.62 and 0.78/1.42, respectively. Zheng et al.[13] employed a deep convolutional neural network model called OnionNet for protein−ligand binding affinity prediction. This model utilizes rotation-free, element-pair-specific contact features between ligand and protein atoms, categorized into different distance ranges to capture both local and nonlocal interaction information. The predictive performance of OnionNet was evaluated using the CASF-2013 and CASF-2016 benchmarks, achieving strong results with Pearson correlation coefficients/RMSE of 0.78/1.50 and 0.81/1.28, respectively. Using 3D convolutional neural network (3D-CNN) architecture, Li et al.[14] developed a model called DeepAtom. This model automatically extracts binding-related atomic interaction patterns from the voxelized structure of the complex and was used for the binding affinity prediction. DeepAtom achieved a Pearson correlation coefficient and RMSE values of 0.81 and 1.32 on the CASF-2016 benchmark, respectively. Wang et al.[15] developed a deep learning approach, named DeepDTAF, to predict the protein−ligand binding affinity. DeepDTAF was constructed by integrating local and global contextual features of the protein and ligand encoded by one-hot encoding and integer encoding, respectively. The prediction performance of DeepDTAF was evaluated on the CASF-2016 benchmark showing Pearson correlation coefficient and RMSE values of 0.75 and 1.44, respectively. Using DNN and atom−atom pairwise interactions, Moon et al.[16] developed a model, PIGNet, for protein−ligand binding affinity predictions. The model generalization of PIGNet was further improved by augmenting the training data with a broader range of binding poses and ligands. The evaluation of the model on the CASF-2016 test set demonstrated a Pearson correlation coefficient of 0.76. Li et al.[17] developed a multiobjective neural network (MONN) to predict both noncovalent interactions and binding affinities between compounds and proteins. They compiled a benchmark data set containing noncovalent intermolecular interactions for more than 10,000 compound−protein pairs and systematically evaluated the interpretability of neural attentions in existing models. MONN was evaluated on the BindingDB data set[18] achieving excellent Pearson correlation coefficient and RMSE values of 0.86 and 0.76, respectively. For further details on studies employing various ML techniques for protein−ligand binding affinity prediction, we refer to the review paper by Zhang et al.[19]

Regardless of whether they are classified as classical or modern, all scoring functions developed for receptor−ligand binding affinity prediction face a common challenge: the need for a clear and explicit understanding of protein−ligand interactions (e.g., hydrogen bonds, polar and hydrophobic interactions, and van der Waals contacts). This necessitates a slow and computationally expensive sampling process before scoring calculations can be performed. Consequently, while modern ML-based scoring functions are significantly faster and more accurate, their integration into molecular docking pipelines still struggles to overcome these limitations due to the inherent "bottleneck" in the sampling stage.

In this study, we introduce a novel ML-based scoring function (iScore) that quickly and precisely predicts the binding affinity of protein−ligand complexes without the need for knowledge of explicit intermolecular interactions. Instead, iScore predicts the protein−ligand binding affinity based on a combination set composed of the ligand and binding pocket descriptors. Therefore, the sampling stage can be avoided, which leads to massive savings in time and resources. On the other hand, since iScore architecture is independent of explicit intermolecular interactions, it can be employed to score and rank huge libraries of de novo small molecules against a protein target of interest, which greatly assists researchers in evaluating "unseen" regions of chemical space. iScore has been trained on the PDBbind 2020[20] refined set using three different ML approaches: DNN (iScore-DNN), Random Forest (RF) (iScore-RF), and eXtreme gradient boosting (iScore-XGB). Furthermore, a hybrid scoring function (iScore-Hybrid) has been developed by combining and taking advantage of these three base learners. The scoring power, ranking power, screening power, and target fishing performances of iScore have been extensively tested and compared with other traditional and ML-based scoring functions using various test sets: PDBbind 2016 core set (Comparative Assessment of Scoring Functions, CASF-2016),[4] two data sets from the Community Structure−Activity Resource (CSAR NRC-HiQ Set1 and CSAR NRC-HiQ Set2),[21] Database of Useful Decoys-Enhanced (DUD-E),[22] and target fishing data set.[23] It is important to emphasize that iScore is not a traditional scoring function designed to assess binding poses or ligand conformations. Instead, it relies on the complementarity between molecular descriptors of ligands and binding pocket features, directly predicting binding affinity without the need for a 3D conformational search or explicit calculations of binding energy components. We believe that iScore paves the way for a new era in de novo drug discovery and pharmaceutical research.

2. MATERIALS AND METHODS

2.1. Data Set Preparation. The iScore models have been trained using the PDBbind 2020 refined set as a training data set.[20] The PDBbind 2020 refined set (consisting of 5316 protein–ligand complexes along with the associated experimental affinity data) is a cherry-picked subset of the PDBbind 2020 general set (over 23,496 complexes) by selecting the complexes with no obvious structural issues or steric clashes, crystal resolution <2.5 Å, R-factor <0.25, noncovalently bound ligand, and affinity data reported in either K_d or K_i form in the range of 10 mM to 1 pM. A full description of the criteria used for defining the PDBbind refined set can be found in the original paper.[4] The PDBbind 2016 core set, the first test set in our study and in the CASF-2016 benchmarking, was selected from the PDBbind refined set by applying even strict criteria as follows: (1) the PDBbind refined set was subjected to a sequence similarity clustering with a similarity cutoff of 90% and only the clusters containing more than 5 members were considered, (2) five representative complexes were selected for each remaining cluster based on their affinity data, those with the highest and lowest affinities with at least 100-fold difference, and three additional (intermediate) complexes, (3) the ligands should not be identical or stereoisomers throughout the PDBbind core set, and (4) the electron density map and the ligand binding pose in each complex should be of high quality. This resulted in 285 protein–ligand complexes clustered into 57 clusters in the PDBbind 2016 core set. The two other test data sets used in this study are CSAR NRC-HiQ Set1 and CSAR NRC-HiQ Set2 containing 176 and 167 high-quality protein–ligand complexes, respectively.

Prior to database preparation, the overlapping complexes between the PDBbind 2020 refined set and the PDBbind 2016 core set were removed from the training set. In addition, the overlapping complexes between the PDBbind 2020 refined set and CSAR NRC-HiQ Set1/Set2 were removed from these test sets. The crystal structures were subsequently prepared using the PrepWizard in the Schrödinger 2023-2 program package (https://www.schrodinger.com/). Hydrogen atoms were incorporated, and missing side chain atoms were added by using Prime. After fixing the potential structural defects, water molecules were removed from the complexes and the protonation states of ionizable residues were determined at pH = 7.0 using PROPKA.[24] The correct protonation states of the ligand molecules were also determined at pH = 7.0 using Epik.[25] The prepared complexes were further refined using the OPLS4 force field[26] in a restrained minimization procedure with an RMSD threshold of 0.3 Å for all heavy atoms. The complexes that failed during the preparation stage were discarded. The final prepared data sets contain 4898 (PDBbind 2020 refined set), 285 (PDBbind 2016 core set), 68 (CSAR NRC-HiQ Set1), and 75 (CSAR NRC-HiQ Set2) complexes. The PDB codes for each protein in the data sets are listed in Table S1.

The DUD-E[22] and target fishing[23] data sets were prepared following the same protocol as described above. DUD-E consists of 102 target proteins with an average of 224 active compounds per target and 50 decoys for each active, resulting in 22,886 actives and 1,144,300 decoys, respectively. For each target, the decoys share similar 1D physicochemical properties with known actives, such as molecular weight and LogP, but differ in 2D topology descriptors. The target fishing data set consists of 122 drugs with 6348 known target proteins from 1860 unique PDB-ids compiled from BindingDB[18] (http://www.bindingdb.org) and DrugBank[27] (http://www.drugbank.ca).

2.2. Descriptor Calculations. *2.2.1. Ligand Descriptors.* The 3D structures of the ligand molecules were converted to the corresponding canonical simplified molecular-input line-entry system (SMILES)[28] strings and subsequently a series of 81 1D/2D molecular descriptors were calculated using the RDKit library (https://www.rdkit.org) in Python such as logarithm of the partition coefficient (MolLogP), molecular refractivity (MolMR), exact molecular weight (ExactMolWt), number of heavy atoms (HeavyAtomCount), number of hydrogen-bond acceptors (NumHAcceptors), number of hydrogen-bond donors (NumHDonors), and number of rotatable bonds (NumRotatableBonds). A full list of the molecular descriptors used in this study is presented in Table S2. Figure S1 shows the histogram distribution of some molecular descriptors of the ligands in the training set along with the logarithmic form of the experimental binding affinity values (pK_d).

2.2.2. Binding Pocket Descriptors. The FPocket[29] tool was employed to calculate 41 descriptors of the protein binding pocket such as the pocket volume (pock_vol), number of alpha spheres (nb_AS), mean alpha sphere radius (mean_as_ray), mean alpha sphere solvent accessibility (mean_as_solv_acc), polarity score (polarity_score), hydrophobicity score (hydrophobicity_score), charge score (charge_score), volume score (volume_score), and amino acid composition. The protein binding pocket was explicitly defined by all atoms situated at a certain cutoff distance from the ligand molecule (3–7 Å). The initial assessments showed that a cutoff distance of 5 Å resulted in the best training and binding affinity prediction performance. A full list of the binding pocket descriptors is presented in Table S2. Figure S2 shows the histogram distribution of some descriptors of the binding pocket in the training data set. Furthermore, FPocket suggests an intuitive estimation for the volume of potential ligands ($\mathrm{LigVol_{BP}}$) which was used in this study as a descriptor in training of the scoring models and as a key feature in training of the ultra-fast screening (UFS) model which was used to improve the screening power performance by further filtering false positives (Section 2.4.3).

2.3. ML Algorithms. iScore has been trained using three different ML approaches: DNN,[30] RF,[31] and XGB.[32] In this study, the hyperparameters of iScore-RF and iScore-XGB models were automatically tuned with the Bayesian optimization[33] technique implemented in the Scikit-learn[34] version 1.0.2, while Keras version 2.4.3 (https://keras.io)[35] was employed for hyperparameter optimization of the iScore-DNN model. A 3-fold cross-validation was used to evaluate various hyperparameter combinations, and RMSE was utilized as the object function. The maximum iteration was set to 200.

2.3.1. Deep Neural Network. The Keras package version 2.4.3 in Python 3 was employed to build the iScore-DNN model. The DNN model consists of six layers: an input layer (81 + 41 neural nodes), four hidden layers with 350, 250, 150, and 50 neural nodes, and an output single node layer. The RELU[36] activation function was used for all layers except the output layer where a LINEAR activation function was employed. The loss function and evaluation metric were set to mean-absolute error and mean-squared error, respectively. An inverse-time-decay scheduler with an initial learning rate of 0.001, a decay rate of 0.3, and a decay step of 8000 was used to properly lower the learning rate during the training process

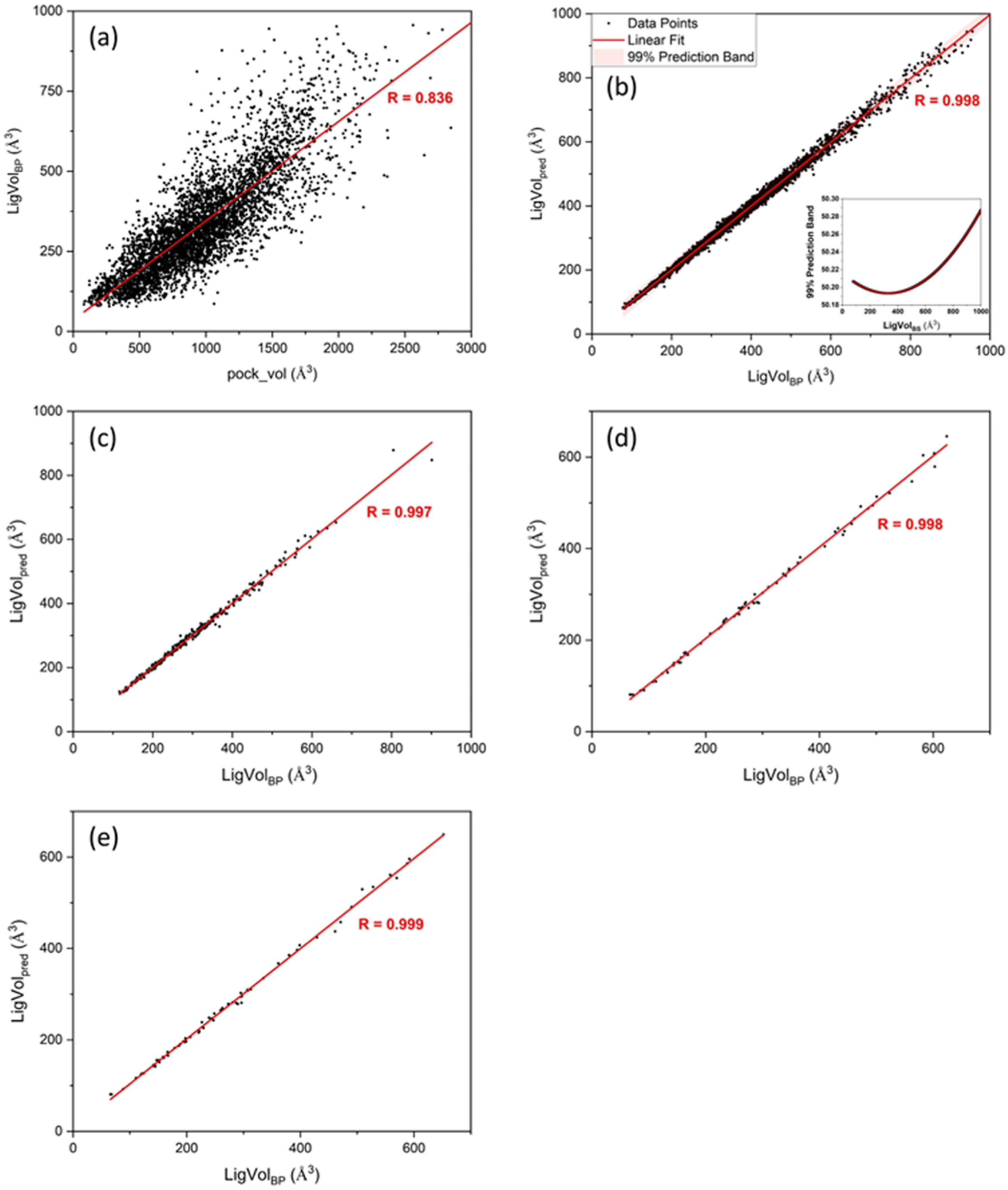


Figure 1. (a) Strong correlation between the LigVol$_{BP}$ and pock_vol. (b) Correlation between LigVol$_{pred}$ estimated from 2D molecular descriptors and LigVol$_{BP}$, with the 99% prediction band region calculated upon 3 × 10-fold XV on the training set. The performance of the RF volume predictor on the (c) PDBbind 2016 core set, (d) CSAR NRC-HiQ Set1, and (e) CSAR NRC-HiQ Set1. The Pearson correlation coefficients (R) are shown in each graph.

with the Adam optimizer and 100 epochs. iScore-DNN was trained thorough 10 × 10-fold cross validation (XV) with random data shuffling in each XV loop. The final output was an average value of over 100 DNN XV models.

2.3.2. Random Forest. The Scikit-learn package version 1.0.2 in Python 3 was used to build the iScore-RF model. The RF consisted of 200 decision trees (n_estimators) with min_samples_split (minimum number of samples required to split an internal node) = 2, min_samples_leaf (minimum number of samples required to be at a leaf node) = 1, and max_features (number of features to consider when looking for the best split) = "auto". The criterion was set to mean-squared error and the estimators were allowed to expand until all leaves were pure. iScore-RF was trained thorough 10 × 10-fold XV with a random data shuffling in each XV loop. The final output was an average value over 100 RF XV models.

2.3.3. eXtreme Gradient Boosting. The XGBoost package version 1.5.2 in Python 3 was used to build the iScore-XGB model. The hyperparameters of the XBG trainer are n_estimators (number of estimators) = 1000, learning_rate

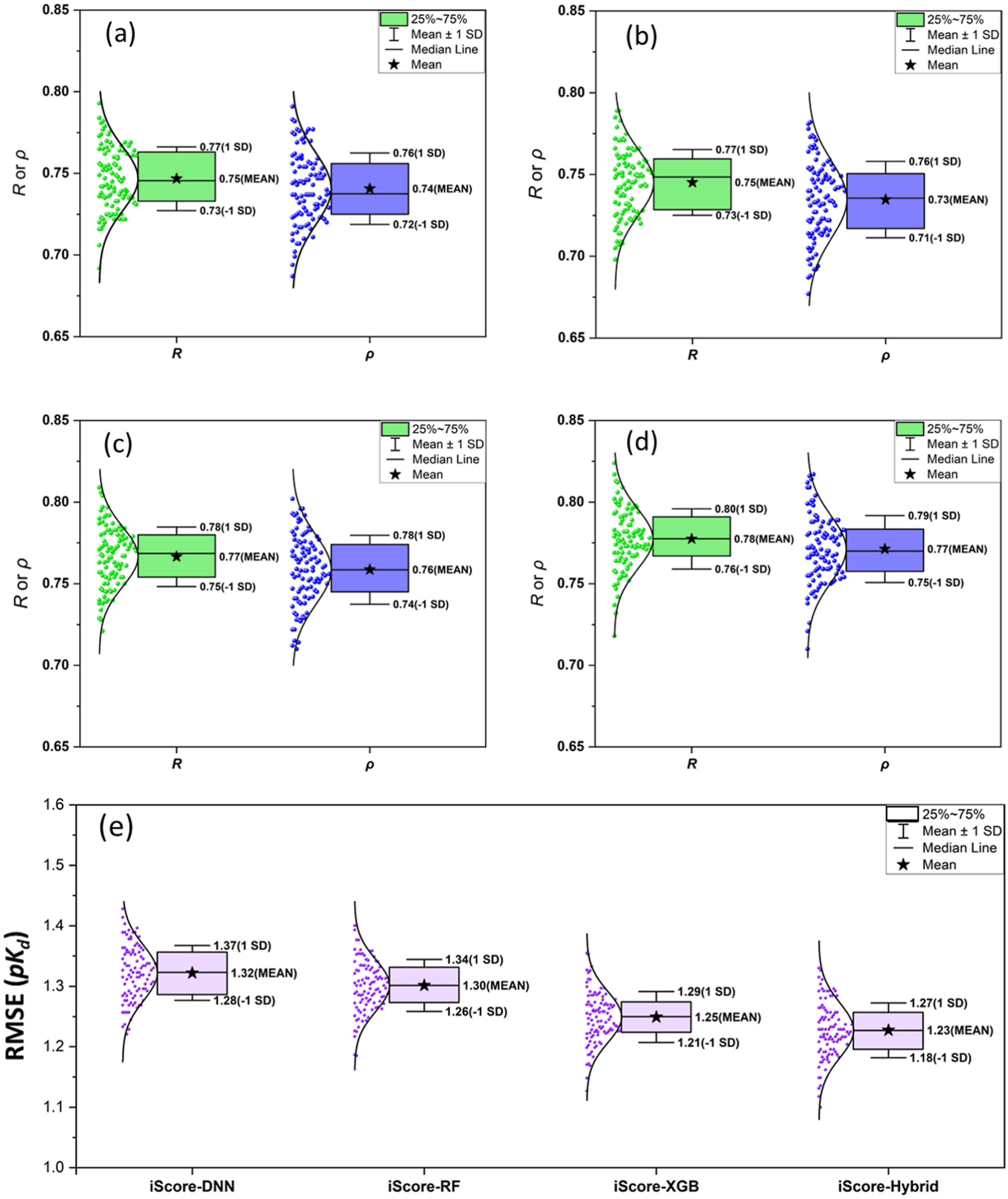


Figure 2. Boxplot presentation and distribution of Pearson (R) and Spearman (ρ) correlation confections along with the mean, median, and SD values for (a) iScore-DNN, (b) iScore-RF, (c) iScore-XGB, and (d) iScore-Hybrid upon 10 × 10-fold XV training campaign. (e) RMSE statistics for the base learners along with the hybrid model.

= 0.01, subsample (Subsample ratio of the training instances prior to growing estimators) = 0.7, colsample_bytree (subsample ratio of columns when constructing each tree) = 1.0, max_depth (maximum depth of an estimator) = 8, and objective (regression type) = "reg/squarederror". iScore-XGB was trained thorough 10 × 10-fold XV with a random data shuffling in each XV loop. The final output was an average value over 100 XGB XV models.

2.3.4. Hybrid Model.

A hybrid scoring function (iScore-Hybrid) was subsequently developed by combining the iScore-DNN, iScore-RF, and iScore-XGB models. For this purpose, the average predicted affinity values over 100 XV of each model (iScore-DNN, iScore-RF, and iScore-XGB) along with experimental affinity data were fed into a DNN trainer. The iScore-Hybrid model consists of six layers: an input layer, four hidden layers with 100, 50, 10, and 50 neural nodes, and an output single node layer. The RELU activation function was

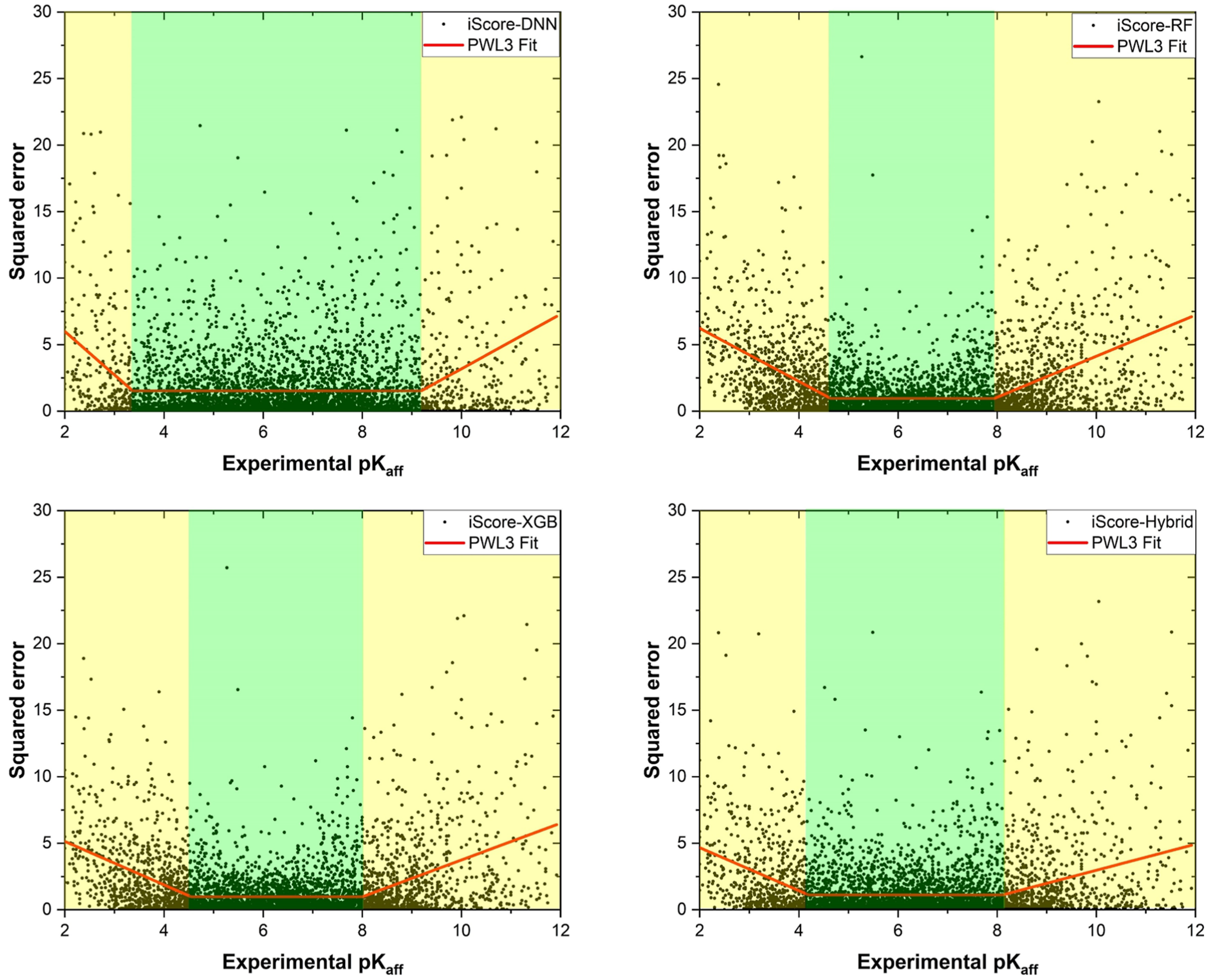


Figure 3. Squared error versus experimental pK_d obtained from each model upon 10 × 10-fold XV training campaign. The piecewise linear function with three segments (PWL3) fitted into the data is shown in red. The green and yellow dashed areas illustrate the trust zone and nonzero-slope regions, respectively.

used for all layers except the output layer, where a LINEAR activation function was employed. The loss function and evaluation metric were set to mean-absolute error and mean-squared error, respectively. An inverse-time-decay scheduler with an initial learning rate of 0.001, decay rate of 0.3, and decay steps of 8000 was used to properly lower the learning rate during the training process with the Adam optimizer and 100 epochs. iScore-Hybrid was trained through 10 × 10-fold XV with a random data shuffling in each XV loop. The final output was an average value over 100 DNN XV models.

2.4. Evaluation Metrics. *2.4.1. Scoring Power.* "Scoring power" indicates the degree of correlation between the predicted and experimentally determined binding affinity values. Hence, the Pearson correlation coefficient (R) was computed as a quantitative indicator of scoring power (eq 1).[4] The RMSE of the regression was considered as an additional indicator (eq 2).[4]

$$R = \frac{\sum_i^n (x_i - \overline{x})}{\sqrt{\sum_i^n (x_i - \overline{x})^2 \sum_i^n (y_i - \overline{y})^2}} \quad (1)$$

$$\mathrm{RMSE} = \sqrt{\frac{(x_i - y_i)^2}{n}} \quad (2)$$

where x_i, y_i, $\overline{x}$, and $\overline{y}$ are the estimated and experimental binding affinity of the ith complex and the corresponding average values, respectively. The summation upper limit (n) is the total number of complexes, i.e., 4898 (PDBbind 2020 refined set), 285 (PDBbind 2016 core set), 68 (CSAR NRC-HiQ Set1), and 75 (CSAR NRC-HiQ Set2).

2.4.2. Ranking Power. "Ranking power" refers to the capability of the scoring function in ranking a given set of active ligands, with respect to their predicted binding affinity values, toward a particular protein target. The PDBbind 2016 core set contains 285 protein−ligand complexes clustered into 57 clusters. Each cluster contains a particular target receptor and 5 different active binders where the difference between binding affinities of the strongest and weakest binders is at least 100-fold. Figure S3 shows a boxplot of experimental binding affinity values for each of the 57 clusters in the PDBbind 2016 core set. The Spearman ranking correlation (ρ, eq 3)[4] was used as an indicator of the ranking power (as CASF-2016 benchmarking) since in contrast to scoring power, ranking power does not request a linear correlation between experimental and predicted binding affinity values.

$$\rho = \sum_i^n (rx_i - \overline{rx})(\frac{(ry_i - \overline{ry})}{\sqrt{\sum_i^n (rx_i - \overline{rx})^2 \sum_i^n (ry_i - \overline{ry})^2}} \quad (3)$$

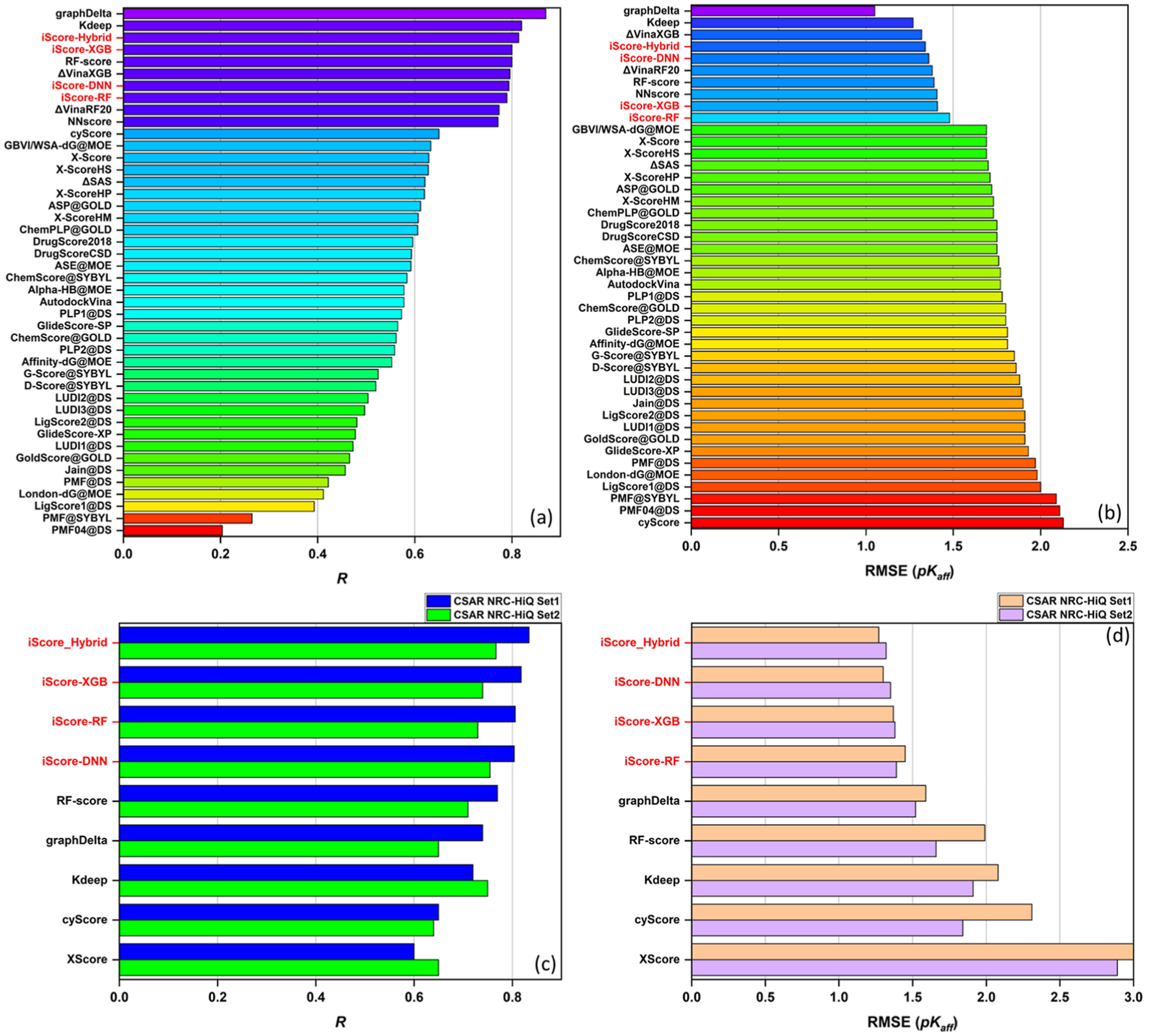


Figure 4. Scoring power performance of iScore models compared to other scoring functions tested on the (a,b) PDBbind 2016 core set and (c,d) CSAR NRC-HiQ Set1 and Set2. Scoring functions are ranked by the Pearson correlation coefficients in descending order.

where rx_i, ry_i, $\overline{rx}$, and $\overline{ry}$ are the rank of estimated and experimental binding affinity of the ith complex and the corresponding average values, respectively. The summation upper limit (n) is the total number of samples in each cluster, i.e., five in this case. The average Spearman ranking correlation, $\langle\rho\rangle$, was subsequently calculated over all 57 target proteins in the PDBbind 2016 core set.

2.4.3. Screening Power. "Screening power" indicates the ability of a scoring function to identify the true binder with the highest affinity for a given protein target among a set of random decoy molecules. The first quantitative reference metric of screening power is the success rate of identifying the ligand with highest affinity against each of the 57 target receptors in the PDBbind 2016 core set, among the 1%, 5%, and 10% top candidates. The second indicator is the success rate of identifying all binders with experimental binding affinity values less than 10 mM ($pK_d \geq 2$), 10 μM ($pK_d \geq 5$), 1 μM ($pK_d \geq 6$), 0.1 μM ($pK_d \geq 7$), 0.01 μM ($pK_d \geq 8$), and 1 nM ($pK_d \geq 9$), among the 1%, 2%, 3%, 5%, and 10% top candidates over all 285 complexes. There are 285, 213, 167, 117, 75, and 39 binders with experimental binding affinity values less than 10 mM, 10 μM, 1 μM, 0.1 μM, 0.01 μM, and 1 nM in the PDBbind 2016 core set, respectively.

The screening power performance of iScore models was further improved by an UFS stage prior to the binding affinity prediction. In the UFS stage, the ligands that volumetrically do not match with a given receptor's binding pocket (too big or too small) will be filtered out. One of the features that the FPocket tool predicts, after receptor binding pocket evaluation, is an intuitive estimation of the volume of potential binders ($\text{LigVol}_{\text{BP}}$) that strongly correlates with the receptor binding pocket volume (pock_vol) (Figure 1a). Hence, an RF-based regression model was trained to calculate $\text{LigVol}_{\text{BP}}$ based on 81 1D/2D molecular descriptors ($\text{LigVol}_{\text{pred}}$) listed in Table S2. From the correlation graph, the 99% prediction band (Figure 1b) was calculated upon 3 × 10-fold XV and used in the UFS

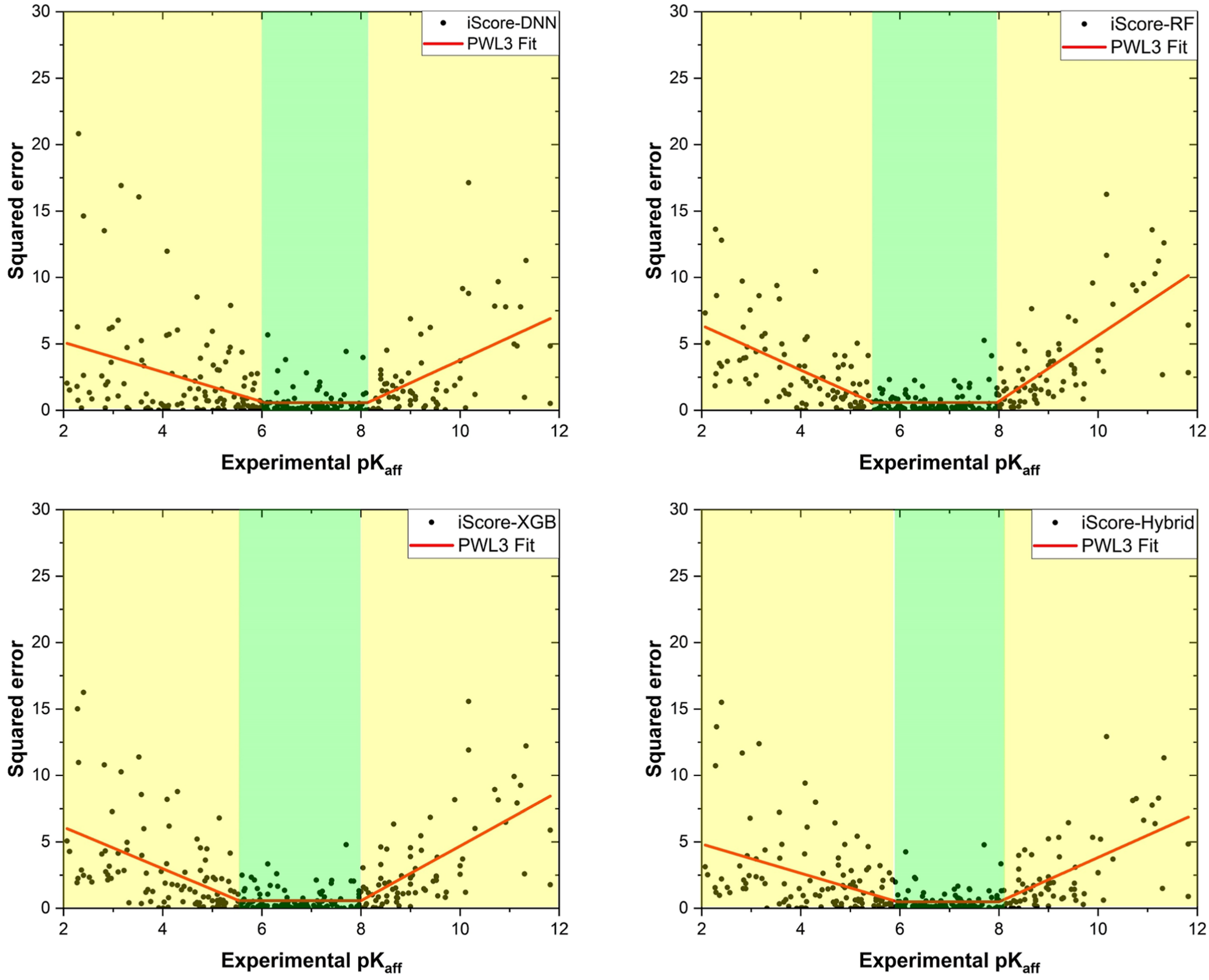


Figure 5. Squared error versus the experimental pK_d obtained from each model upon testing campaign on the PDBbind 2016 core set. The piecewise linear function with three segments (PWL3) fitted into the data is shown in red. The green and yellow dashed areas illustrate the trust zone and nonzero-slope regions, respectively.

stage, so that only the ligands with predicted volumes ($LigVol_{pred}$) within the 99% prediction band from the $LigVol_{BP}$ value were allowed to pass to the scoring stage. The trained RF volume predictor has been tested on three test sets used in this study (PDBbind 2016 core set, CSAR NRC-HiQ Set1 and Set2) and the results showed very close to perfect correlation (Figure 1c−e).

3. RESULTS

3.1. Training. Figure 2a−d shows the 25%−75% boxplot presentation and distribution of Pearson (R) and Spearman (ρ) correlation confections along with the mean, median, and standard deviation (SD) values for the models trained with different ML algorithms (base learners) and the hybrid model, upon 10 × 10-fold XV training campaign. The mean Pearson coefficients are 0.75, 0.75, 0.77, and 0.78 for the iScore-DNN, iScore-RF, iScore-XGB, and iScore-Hybrid models, respectively. The SD profile of the Pearson coefficient is similar for all models and lies in the range of 0.01−0.02. As these figures indicate, the mean Spearman coefficients (ρ) are slightly lower than the mean Pearson coefficients, but the same SD values have been observed. Figure 2e shows the RMSE statistics for three base learners along with the hybrid model. The mean RMSE values are 1.32, 1.30, 1.25, and 1.23 for the iScore-DNN, iScore-RF, iScore-XGB, and iScore-Hybrid models, respectively. The SD profiles of the RMSE metric are similar and cover the range of 0.04−0.05. The results clearly show that iScore-Hybrid outperforms the base learners with higher mean Pearson and Spearman correlation coefficients and lower mean RMSE values in the cross-validation training campaign. The scatter plots comparing the predicted to experimental values for the internal 10-fold cross validation are shown in Figure S4a−d. The 99% prediction band (pink area), Pearson correlation coefficient (R), and RMSE values are shown for each correlation.

To further understand the better performance of the iScore-Hybrid over all base learners, the plots of the squared error (squared difference between the experimental and the predicted pK_d) versus the experimental pK_d (Figure 3) associated with each model have been deeply explored. As Figure 3 illustrates, there are three distinct regions that were quantitatively distinguished after fitting the data into a

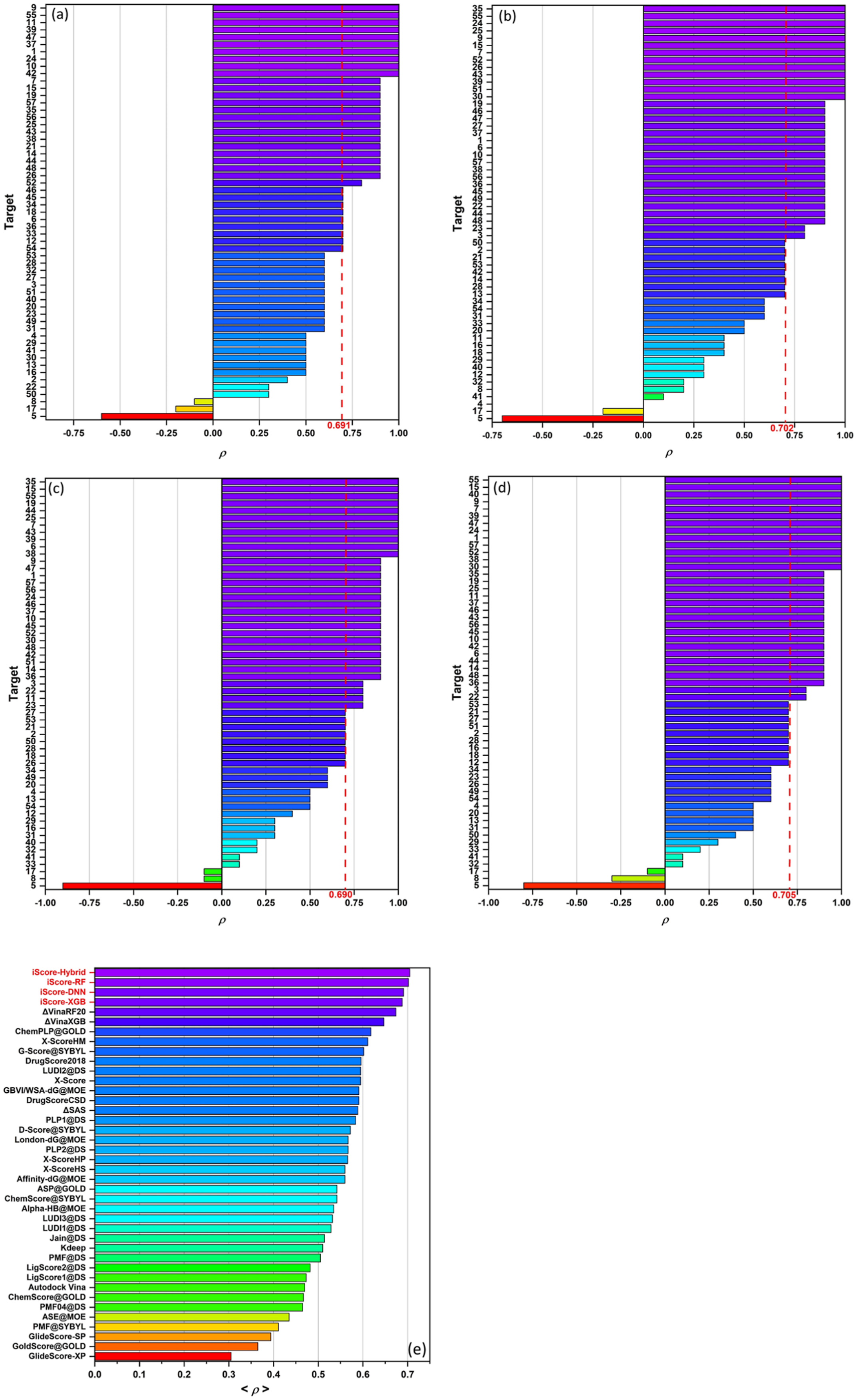


Figure 6. (a) Ranking power performance of (a) iScore-DNN, (b) iScore-RF, (c)-iScore-XGB, and (d) iScore-Hybrid on the CASF-2016 test set based on the Spearman correlation coefficient of individual targets. (e) Comparison between the ranking power performance of iScore models and other scoring functions based on the average Spearman correlation coefficients (vertical red dashed lines in (a−d)) over all targets.

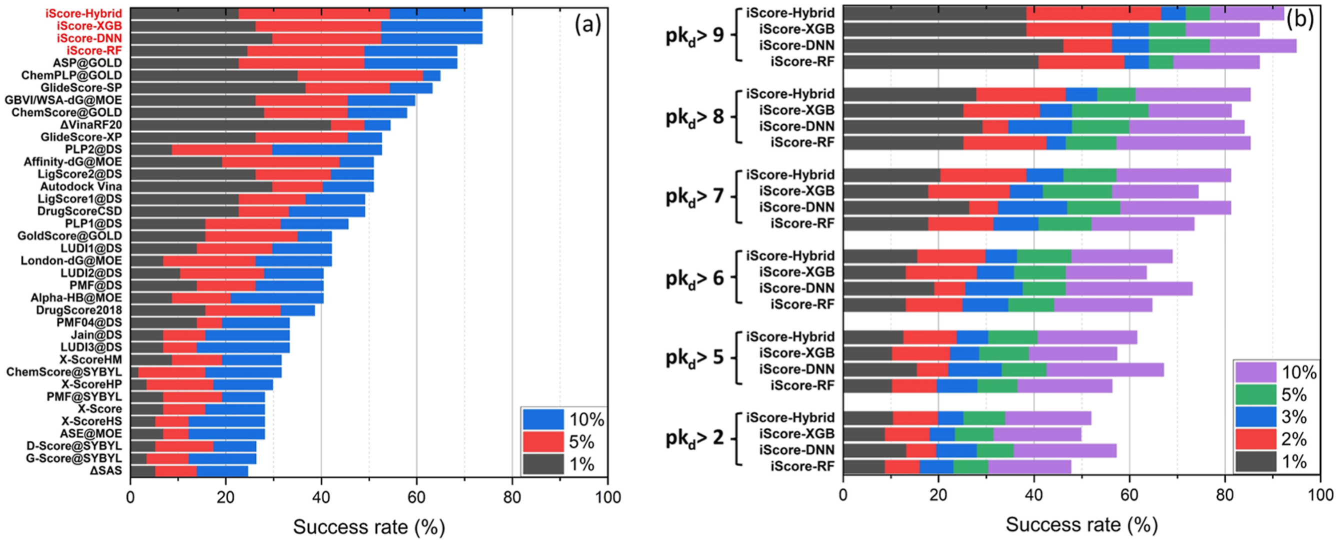


Figure 7. (a) Comparison between screening power performance of iScore and other scoring functions in the success rate of identifying the highest affinity ligand of each 57 target receptors, in the PDBbind 2016 core set, among the 1%, 5%, and 10% top candidates. (b) Success rate of identifying all binders with the experimental binding affinity values less than 10 mM ($pK_d \geq 2$), 10 μM ($pK_d \geq 5$), 1 μM ($pK_d \geq 6$), 0.1 μM ($pK_d \geq 7$), 0.01 μM ($pK_d \geq 8$), and 1 nM ($pK_d \geq 9$), among the 1%, 2%, 3%, 5%, and 10% top candidates over all 285 complexes in the PDBbind core set.

piecewise linear function with three segments (PWL3). The first region (green dashed area) is the trust zone in the midrange pK_d spectrum where the PWL3 function forms a horizontal line indicating the most reliable range of pK_d that the model can predict at the maximum accuracy (minimum error). The other two regions are at the respective ends of the experimental pK_d (yellow dashed areas) where the PWL3 function forms nonzero-slope lines. One can elucidate the overall performance of the models by comparing three determinative factors: the trust zone's length (the bigger the better) and height (the lower the better) and the absolute slope of the lines in the nonzero-slope regions (the lower the better). The maximum trust-zone length is 5.82 [3.37, 9.19] which belongs to the iScore-DNN, while the corresponding values are 3.30 [4.64, 7.94], 3.45 [4.55, 8.00], and 3.91 [4.20, 8.11] for iScore-RF, iScore-XGB, and iScore-Hybrid, respectively. On the other hand, iScore-DNN has the maximum trust-zone height of 1.53 followed by 1.12 (iScore-Hybrid), 0.99 (iScore-XGB), and 0.97 (iScore-RF). Moreover, the minimum absolute slopes of the lines in the nonzero-slope regions belong to iScore-Hybrid which are 1.60 and 0.98 at the low- and high-affinity limits, respectively. The corresponding values are (1.63 and 1.38), (2.00 and 1.54), and (3.27 and 2.05), for iScore-XGB, iScore-RF, and iScore-DNN, respectively. Therefore, in the context of the trust-zone length, iScore-Hybrid showed a better performance than the two base learners iScore-RF and iScore-GXB. In the context of the trust-zone height, iScore-Hybrid outperforms the iScore-DNN by a considerable margin. Interestingly, iScore-Hybrid furthermore showed the best performance at both low- and high-affinity limits, where the base learners suffer from the lower prediction accuracy.

3.2. Benchmarks. The scoring, ranking, and screening power performances of the iScore models have been extensively tested and compared to other traditional and ML-based scoring functions on three different test sets: the PDBbind 2016 core set (scoring, ranking, and screening power), CSAR NRC-HiQ Set1 and Set2 (scoring performance), and DUD-E (screening power). Moreover, the target fishing capability of the models was evaluated and compared to various scoring functions.

3.2.1. Scoring Power. Scatter plots comparing the predicted to the experimental values on the PDBbind-2016 core set are shown in Figure S4e−h. The 99% prediction band (pink area), Pearson correlation coefficient (R), and RMSE values are shown for each correlation. Figure 4a,b illustrates the scoring power performance of the iScore models versus 40 traditional and modern ML-based scoring functions in terms of the Pearson correlation coefficient and RMSE metrics tested on the PDBbind 2016 core set, respectively. As these figures show, iScore-Hybrid outperforms the base learners in the context of scoring power metrics (R = 0.814 and RMSE = 1.30) and ranks among the top scoring functions. Two major competitors are graphDelta[37] and K$_{\text{deep}}$.[38] graphDelta is a ML graph-based scoring function that employs a message-passing neural network for modeling protein−ligand interactions and yielded the best scoring power metrics on the PDBbind 2016 core set with R = 0.87 and RMSE = 1.05. K_{deep} is also an ML-based scoring function that uses a 3D-convolutional neural network for predicting the ligand binding affinities. K_{deep} demonstrated a very good scoring power performance on the PDBbind 2016 core set with R = 0.82 and RMSE = 1.27. Nonetheless, these scoring functions, like any other scoring functions published up to now, require a full picture of the protein−ligand interactions, which imposes some critical limitations on their speed and applicability as discussed in the Introduction. Figure 4c,d compares the Pearson correlation coefficient and RMSE metrics of the iScore models against modern ML-based scoring functions, tested on CSAR NRC-HiQ Set1and Set2, respectively. iScore-Hybrid is among the top 3 best performing scoring functions on the PDBbind 2016 core set and is at the top of the list once tested on CSAR NRC-HiQ Set1 (R = 0.834 and RMSE = 1.27) and CSAR NRC-HiQ Set2 (R = 0.767 and RMSE = 1.32). The iScore base learners also outperform the other scoring functions including graphDelta (R = 0.74, 0.65 and RMSE = 1.59, 1.52) and K_{deep} (R = 0.72, 0.75 and RMSE = 2.08, 1.91) in these sets. As the test results indicate (Figure 4), iScore-XGB and iScore-DNN are the best performing base

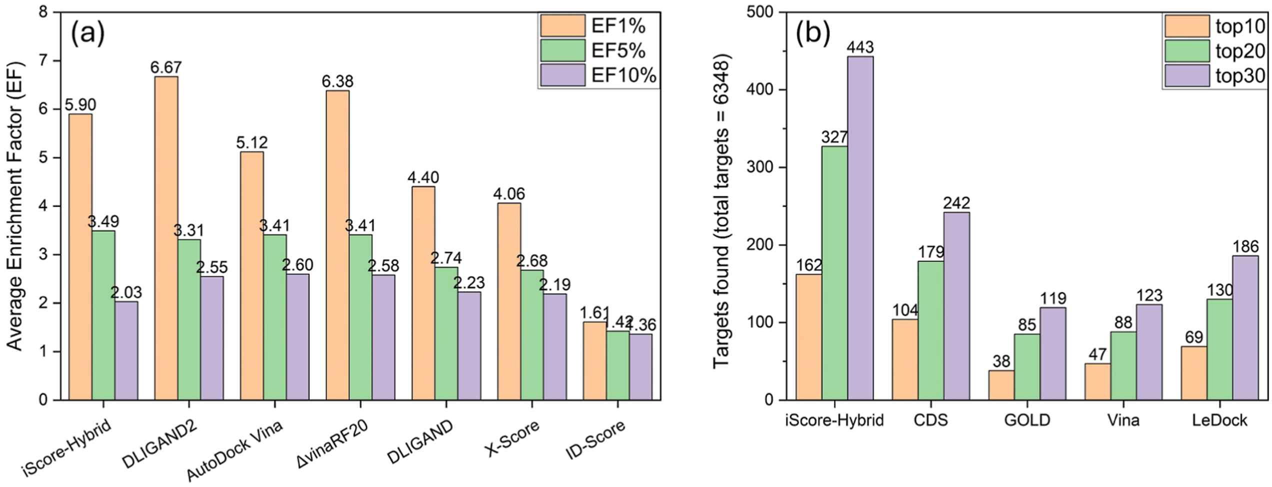


Figure 8. (a) Comparison of iScore-Hybrid's screening power performance with six other scoring functions, evaluated using the average EFs at the top 1% (EF1%), top 5% (EF5%), and top 10% (EF10%) on the DUD-E benchmark. (b) Comparison of iScore-Hybrid's target fishing performance against four different scoring functions, assessed based on the total number of targets identified within the top 10%, 20%, and 30%.

learners in terms of Pearson correlation coefficient and RMSE metrics on all three test sets, respectively.

Figure 5 illustrates the squared error (squared difference between experimental and predicted pK_d) versus the experimental pK_d associated with each iScore model obtained on the PDBbind 2016 core set. The trust zone's lengths of iScore-Hybrid and iScore-DNN are similar (~2.1 [6.0, 8.1]), while iScore-RF and iScore-XGB show a higher value (~2.5 [5.5, 8.0]). iScore-Hybrid has the lowest trust zone's height, 0.48, followed by 0.56 (iScore-DNN), 0.59 (iScore-RF), and 0.60 (iScore-XGB). Furthermore, iScore-Hybrid shows the best performance at the low- and high-affinity limits where the absolute slope of the lines in these regions is 1.10 and 1.67, respectively. The corresponding values are (1.12 and 1.72), (1.68 and 2.48), and (1.56 and 2.06), for iScore-DNN, iScore-RF, and iScore-XGB, respectively. Hence, except for the trust zone's length, iScore-Hybrid outperforms the base learners.

To demonstrate that the models effectively capture protein−ligand interactions, an ablation study was conducted by training two separate models for each base learner—DNN, RF, and XGB. One model utilized only ligand descriptors, while the other relied solely on binding pocket descriptors. The scoring power of these ablation models was then evaluated on the CASF-2016 data set. As shown in Figure S5, the results clearly indicate that using information from only one source (either the ligand or binding pocket) is insufficient for training robust models.

3.2.2. Ranking Power. Figure 6a−d shows the per-target and average (vertical red dashed lines) ranking Spearman correlation coefficients (ρ) over 57 targets in the PDBbind 2016 core set evaluated by the iScore base learners and the hybrid model. Figure 6e illustrates the ranking power performance of the iScore models (based on the average Spearman correlation coefficient) and compares those with several other scoring functions. As this figure shows, iScore-Hybrid ($<\rho> = 0.705$) outperforms not only the base learners but all other scoring functions in the ranking campaign. As Figure 6e indicates, iScore-Hybrid is followed by iScore-RF ($<\rho> = 0.702$), iScore-DNN ($<\rho> = 0.691$), and iScore-XGB ($<\rho> = 0.690$), respectively. It is worth to mention that the ranking power performances of iScore models are significantly better than K_{deep} ($<\rho> = 0.51$) which was one of the major competitors when comparing scoring power.

3.2.3. Screening Power. Figure 7a shows the screening power performance of iScore in terms of the success rate of identifying the highest affinity ligand of each 57 target receptors, in the PDBbind 2016 core set, among the 1%, 5%, and 10% top candidates. Figure 7a demonstrates that the screening performance of iScore (73.7% for iScore-Hybrid, iScore-XGB, and iScore-DNN and 68.4% for iScore-RF) is considerably better than that of all other scoring functions in the screening power campaign. Figure 7b illustrates the success rate of identifying all binders with experimental binding affinity values less than 10 mM ($pK_d \geq 2$), 10 μM ($pK_d \geq 5$), 1 μM ($pK_d \geq 6$), 0.1 μM ($pK_d \geq 7$), 0.01 μM ($pK_d \geq 8$), and 1 nM ($pK_d \geq 9$), among the 1%, 2%, 3%, 5%, and 10% top candidates over all 285 complexes in the PDBbind core set. As this figure shows, the success rate of iScore increases as pK_d of the binder increases. For instance, the success rate of iScore-Hybrid is 51.9, 61.5, 68.9, 81.2, 85.3, and 92.3% for identifying all binders with the experimental binding affinity values less than 10 mM, 10 μM, 1 μM, 0.1 μM, 0.01 μM, and 1 nM, among the 10% top candidates.

For a more robust evaluation, the screening power of iScore was further benchmarked against the DUD-E data set.[22] Due to the hidden bias observed in this data set,[39−41] it is not recommended for use as a training set in ML classification tasks, as it may misleadingly yield strong internal validation results but poor generalization to unseen data. Nonetheless, DUD-E remains a valid benchmark for evaluating the screening power of iScore. As mentioned earlier, since the actives and decoys in this data set share similar 1D descriptors, iScore's performance in this case must rely solely on differences in 2D descriptors.

The DUD-E data set was downloaded and prepared following the approach described in Materials and Methods. The prepared structures of protein−ligand complexes, and active and decoy ligands in SMILES format, are freely available in the Zenodo repository at (DOI: 10.5281/zenodo.14865257). The ligand descriptors and protein binding pocket features were calculated by using RDKit and FPocket, respectively. Subsequently, the binding affinity of each target protein against its active and decoys was predicted by the iScore models and sorted. The Enrichment Factor (EF) at the top 1% (EF1%), 5% (EF5%), and 10% (EF10%) was calculated using the following equation

$$\mathrm{EF}x\ \% = \frac{N_{\mathrm{act}}^{x\,\%}/N^{x\,\%}}{N_{\mathrm{act}}^{\mathrm{total}}/N^{\mathrm{total}}} \tag{5}$$

where $N_{\mathrm{act}}^{x\,\%}$, $N^{x\,\%}$, $N_{\mathrm{act}}^{\mathrm{total}}$, and N^{total} are the number of actives found in the top x %, number of compounds (actives and decoys) in the top x %, total number of actives per target protein, and total number of compounds (actives and decoys) per target protein, respectively. Figure S6 shows EF1%, EF5%, and EF10% per target protein in the DUD-E data set, with the average EFs presented in Figure 8a and compared to six scoring functions reported by Chen et al.[42] The results indicate that iScore ranked third in terms of EF1% (5.90), following DLIGAND2 (6.67) and ΔvinaRF20 (6.38). However, iScore outperforms the other scoring functions in terms of EF5% (3.49) and is the second lowest EF10% (2.03) after ID-Score (1.36). While these benchmarking results clearly demonstrate that 2D molecular descriptors can be used to distinguish actives from decoys in the DUD-E data set, the importance of 1D molecular descriptors for the iScore model in making an effective classification remains undeniable.

3.2.4. Target Fishing. To verify that the iScore models genuinely utilize binding pocket information in affinity predictions—rather than making spurious correlations with features specific to the current training and test sets—a target fishing (reverse docking) study was conducted, as outlined by Lee and Kim.[23] The objective of reverse docking is to identify true targets among a diverse set of clinically relevant protein targets.

The drug molecules and target proteins were retrieved from DrugBank (http://www.drugbank.ca) and PDB (https://www.rcsb.org/), respectively, and prepared following the approach described in Material and Methods. The prepared structures of protein−ligand complexes and prepared drug molecules in SDF format are freely available in the Zenodo repository at DOI: 10.5281/zenodo.14865257. Table S3 lists the drug names, their associated PubChem- and DrugBank-id's, PDB-id, and number of active target proteins per drug molecule. The cocrystallized ligands were used to define the protein binding pocket for its descriptor calculations. The molecular descriptors for each drug were calculated using the RDKit tool. Subsequently, the binding affinity for each drug against all 1860 unique target proteins was predicted by the iScore models and sorted. Figures S7 and 8b present the per-drug and total number of targets identified within the top 10%, 20%, and 30%, respectively. As illustrated in Figure 8b, iScore-Hybrid significantly outperforms the other scoring functions. The total number of active targets identified by iScore-Hybrid at the top 10%, 20%, and 30% thresholds is 162, 327, and 443, respectively, whereas the second-best scoring function "Consensus Docking Score", identified only 104, 179, and 242 active targets at the same thresholds.

3.3. Speed Performance. The iScore models were trained using a single compute node on the Alvis supercomputer allocated by the C3SE supercomputing facility, with one NVIDIA Tesla A100 HGX GPU (40GB RAM), 32 core Intel(R) Xeon(R) Gold 6338 CPU 2 GHz, and 256GB DDR4 RAM. iScore is capable of screening >8000 compound/s (~700 million screenings a day) on a single compute node: 32 cores AMD Ryzen 9 7950X CPU, NVIDIA RTX A4000 GPU (16GB RAM), and 32GB DDR5 RAM.

4. CONCLUSIONS

This work introduces iScore, a cutting-edge ML-based scoring function designed to predict the binding affinity of protein−ligand complexes with unprecedented precision and speed. Unlike traditional scoring functions that rely heavily on the explicit knowledge of intermolecular interactions obtained from explicit binding poses, iScore leverages a novel approach. It utilizes a combination of ligand and binding pocket descriptors to directly predict binding affinities, thereby bypassing the need for extensive conformational sampling. This methodological innovation not only saves significant computational time and resources but also provides the applicability to evaluate vast molecular libraries, offering a leap toward exploring the chemical space more efficiently. The benchmarking of iScore across multiple data sets highlights its robustness and superior performance over traditional and advanced scoring functions. Notably, the development of the hybrid iScore model (iScore-Hybrid), which integrates the strengths of individual base learners, sets new benchmarks in scoring, ranking, screening, and target fishing capabilities, which are essential for drug discovery processes. The innovation of iScore is further underscored by its practical implications. The ability to screen over 8000 compounds per second on a single GPU translates to the screening of 700 million compounds daily, illustrating the scalability and efficiency of iScore in handling ultralarge molecular libraries. This capability is critical in accelerating the drug discovery process, from initial screening to identification of lead compounds. The promising results of iScore not only benchmark a new standard in scoring function development but also open a new era in the utilization of ML technologies for drug discovery.

■ ASSOCIATED CONTENT

Data Availability Statement

The training data sets, pretrained models, and instruction for the retraining is available at https://github.com/i-TripleD/iScore. The prepared DUD-E and target fishing data sets are freely available in the Zenodo repository at zeonodo.org, with DOI: 10.5281/zenodo.14865257.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c02192.

Additional information regarding the pdb codes used for training and benchmarking, database analysis, molecular and binding pocket descriptors, training and test scatter plots, per-target EFs for DUD-E benchmark, and per-drug active targets found for target fishing benchmark (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Leif A. Eriksson − *Department of Chemistry and Molecular Biology, University of Gothenburg, Göteborg 405 30, Sweden;* orcid.org/0000-0001-5654-3109; Email: leif.eriksson@chem.gu.se

Author

Sayyed Jalil Mahdizadeh − *Department of Chemistry and Molecular Biology, University of Gothenburg, Göteborg 405 30, Sweden*; orcid.org/0000-0002-4844-6234

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c02192

Author Contributions

Both authors conceived the study. SJM developed the methodology and trained and benchmarked the models. Both authors wrote and revised the text.

Notes

The authors declare the following competing financial interest(s): Both authors are co-founders of ANYO Labs AB. The authors report no conflicting interests.

■ ACKNOWLEDGMENTS

The Faculty of Science at the University of Gothenburg, the Swedish Science Research Council (VR; grant no. 2019-3684), and the Swedish Cancer Foundation (CF; grant no. 21-1447-Pj) (L.A. Eriksson) are also gratefully acknowledged for funding. The authors thank the National Academic Infrastructure for Supercomputing in Sweden (NAISS) for generous allocations of computing time at supercomputing centers C3SE and NSC in part funded by the Swedish Research Council through grant agreement 2022-06725. Authors greatly appreciate ANYO Labs AB and particularly André Stadelmann for professional parallelization of the tool to maximize the ultimate performance.

■ REFERENCES

(1) Morris, G. M.; Lim-Wilby, M. Molecular Docking. In *Molecular Modeling of Proteins*; Humana Press, 2008; Vol. *443*; pp 365−382..

(2) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46* (12), 2287−2303.

(3) Muegge, I.; Rarey, M. Small molecule docking and scoring. *Rev. Comput. Chem.* **2001**, *17*, 1−60.

(4) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative assessment of scoring functions: the CASF-2016 update. *J. Chem. Inf. Model.* **2019**, *59* (2), 895−913.

(5) Coleman, R. G.; Carchia, M.; Sterling, T.; Irwin, J. J.; Shoichet, B. K. Ligand pose and orientational sampling in molecular docking. *PLoS One* **2013**, *8* (10), No. e75992.

(6) Reymond, J.-L. The chemical space project. *Acc. Chem. Res.* **2015**, *48* (3), 722−730.

(7) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; et al. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49* (20), 5912−5931.

(8) Zheng, L.; Meng, J.; Jiang, K.; Lan, H.; Wang, Z.; Lin, M.; Li, W.; Guo, H.; Wei, Y.; Mu, Y. Improving protein−ligand docking and screening accuracies by incorporating a scoring function correction term. *Brief. Bioinform.* **2022**, *23* (3), bbac051.

(9) Wang, L.; Chambers, J.; Abel, R. Protein−ligand binding free energy calculations with FEP+. In *Biomolecular Simulations: Methods and Protocols*; Humana Press: New York, NY, 2019; Vol. *2022*; pp 201−232..

(10) 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* (1), No. e1429.

(11) Adeshina, Y. O.; Deeds, E. J.; Karanicolas, J. Machine learning classification can reduce false positives in structure-based virtual screening. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117* (31), 18477−18488.

(12) Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P. Development and evaluation of a deep learning model for protein−ligand binding affinity prediction. *Bioinformatics* **2018**, *34* (21), 3666−3674.

(13) 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* (14), 15956−15965.

(14) Li, Y.; Rezaei, M. A.; Li, C.; Li, X. DeepAtom: A framework for protein-ligand binding affinity prediction. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*; IEEE, 2019; pp 303−310.

(15) Wang, K.; Zhou, R.; Li, Y.; Li, M. DeepDTAF: a deep learning method to predict protein−ligand binding affinity. *Brief. Bioinform.* **2021**, *22* (5), bbab072.

(16) Moon, S.; Zhung, W.; Yang, S.; Lim, J.; Kim, W. Y. PIGNet: a physics-informed deep learning model toward generalized drug−target interaction predictions. *Chem. Sci.* **2022**, *13* (13), 3661−3673.

(17) Li, S.; Wan, F.; Shu, H.; Jiang, T.; Zhao, D.; Zeng, J. MONN: a multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Syst.* **2020**, *10* (4), 308−322.

(18) Gilson, M. K.; 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* (D1), D1045−D1053.

(19) Zhang, Y.; Li, S.; Meng, K.; Sun, S. Machine Learning for Sequence and Structure-Based Protein−Ligand Interaction Prediction. *J. Chem. Inf. Model.* **2024**, *64* (5), 1456−1472.

(20) 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* (2), 302−309.

(21) Dunbar, J. B., Jr.; Smith, R. D.; Damm-Ganamet, K. L.; Ahmed, A.; Esposito, E. X.; Delproposto, J.; Chinnaswamy, K.; Kang, Y.-N.; Kubish, G.; Gestwicki, J. E.; et al. CSAR data set release 2012: ligands, affinities, complexes, and docking decoys. *J. Chem. Inf. Model.* **2013**, *53* (8), 1842−1852.

(22) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* **2012**, *55* (14), 6582−6594.

(23) Lee, A.; Kim, D. CRDS: consensus reverse docking system for target fishing. *Bioinformatics* **2020**, *36* (3), 959−960.

(24) Olsson, M. H.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: consistent treatment of internal and surface residues in empirical p K a predictions. *J. Chem. Theory Comput.* **2011**, *7* (2), 525−537.

(25) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: a software program for pK a prediction and protonation state generation for drug-like molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 681−691.

(26) Lu, C.; Wu, C.; Ghoreishi, D.; Chen, W.; Wang, L.; Damm, W.; Ross, G. A.; Dahlgren, M. K.; Russell, E.; Von Bargen, C. D.; et al. OPLS4: Improving force field accuracy on challenging regimes of chemical space. *J. Chem. Theory Comput.* **2021**, *17* (7), 4291−4300.

(27) Knox, C.; Wilson, M.; Klinger, C. M.; Franklin, M.; Oler, E.; Wilson, A.; Pon, A.; Cox, J.; Chin, N. E.; Strawbridge, S. A.; et al. DrugBank 6.0: the DrugBank knowledgebase for 2024. *Nucleic Acids Res.* **2024**, *52* (D1), D1265−D1275.

(28) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31−36.

(29) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinf.* **2009**, *10*, 168.

(30) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521* (7553), 436−444.

(31) Biau, G.; Scornet, E. A random forest guided tour. *Test* **2016**, *25*, 197−227.

(32) Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T. *Xgboost: Extreme Gradient Boosting*, version 0.4−2, 2015.

(33) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; De Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE* **2016**, *104* (1), 148−175.

(34) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(35) Pon, M. Z. A.; Kk, K. P. Hyperparameter tuning of deep learning models in keras. *Sparklinglight Trans. Artificial Intelligence Quantum Computing (STAIQC)* **2021**, *1* (1), 36−40.

(36) Agarap, A. F. Deep learning using rectified linear units (relu). **2018**, arXiv:1803.08375. arXiv preprint.

(37) Karlov, D. S.; Sosnin, S.; Fedorov, M. V.; Popov, P. graphDelta: MPNN scoring function for the affinity prediction of protein−ligand complexes. *ACS Omega* **2020**, *5* (10), 5150−5159.

(38) 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* (2), 287−296.

(39) Xia, J.; Tilahun, E. L.; Reid, T.-E.; Zhang, L.; Wang, X. S. Benchmarking methods and data sets for ligand enrichment assessment in virtual screening. *Methods* **2015**, *71*, 146−157.

(40) Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C. J.; Duca, J. S.; Hornak, V.; Koes, D. R.; Kurtzman, T. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS One* **2019**, *14* (8), No. e0220113.

(41) Sieg, J.; Flachsenberg, F.; Rarey, M. In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *J. Chem. Inf. Model.* **2019**, *59* (3), 947−961.

(42) Chen, P.; Ke, Y.; Lu, Y.; Du, Y.; Li, J.; Yan, H.; Zhao, H.; Zhou, Y.; Yang, Y. DLIGAND2: an improved knowledge-based energy function for protein−ligand interactions using the distance-scaled, finite, ideal-gas reference state. *J. Cheminf.* **2019**, *11*, 52.