SMPLIP-Score: Predicting the Ligand Binding Affinity from Simple and Interpretable On-The-Fly Interaction Fingerprint Pattern Descriptors

Surendra Kumar
Gachon University College of Pharmacy https://orcid.org/0000-0002-8065-5183
Mi-hyun Kim (kmh0515@gachon.ac.kr)
Gachon University College of Pharmacy https://orcid.org/0000-0002-2718-5637

Research article

Keywords: Protein-ligand binding affinity, Interaction fingerprint pattern, Substructural molecular fragments, Random forest, Neural Network, Featurization.

Posted Date: September 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-74202/v1

Version of Record: A version of this preprint was published on March 25th, 2021. See the published version at https://doi.org/10.1186/s13321-021-00507-1.

SMPLIP-Score: Predicting the Ligand Binding Affinity from Simple and Interpretable On-The-Fly Interaction Fingerprint Pattern Descriptors

Surendra Kumar, Mi-hyun Kim*

Gachon Institute of Pharmaceutical Science & Department of Pharmacy, College of Pharmacy, Gachon University, 191 Hambakmoeiro, Yeonsu-gu, Incheon, Republic of Korea

*Author for correspondence

E-mail: kmh0515@gachon.ac.kr

Abstract:

In drug discovery, rapid and accurate prediction of protein-ligand binding affinities is a pivotal task for lead optimization with acceptable on-target potency as well as pharmacological efficacy. Furthermore, researchers hope high correlation between a docking score and a pose with key interactive residues, though scoring functions as a free energy surrogate of a protein-ligand complex have failed to provide the collinearity. Recently, various machine learning or deep learning methods have been proposed to overcome the drawback of scoring functions. Despite their high accuracy, their featurization process is complex and requires high cost for its interpretation (less compatible for human recognition). Here, we propose SMPLIP-Score (Substructural Molecular and Protein-Ligand Interaction Pattern Score), a simple interpretable predictor of the absolute binding affinity. Our simple featurization embedded the interaction fingerprint pattern on the ligand-binding site environment and molecular fragments of ligands into an input vectorized matrix for learning layers (random forest or deep neural network). Despite lower complexity than state-of-the-art models, SMPLIP-Score achieved comparable performance, a Pearson's correlation coefficient up to 0.80 and a RMSE up to 1.18 in p*K* units on several benchmark datasets (PDBbind v.2015, Astex Diverse Set, CSAR NRC HiQ, FEP, PDBbind NMR, and CASF-2016). For this model, generality, predictive power, ranking power, and robustness also were examined with direct interpretation of feature matrices for specific targets.

Keywords: Protein-ligand binding affinity; Interaction fingerprint pattern; Substructural molecular fragments; Random forest; Neural Network; Featurization.

1. Introduction:

The protein-ligand binding in the living organism is a biological phenomenon that involves comprehensive processes such as molecular recognition and changes in protein conformations[1]. During the drug development, any new molecules are evaluated experimentally by measuring its binding strength of a ligand to a protein target *in vivo* or *in vitro*. On the other hands, ligand-based and target-based approaches are being used computationally to predict the binding strengths of ligand[2–4]. In recent years, with the advancement in computational power, the FEP (free-energy perturbation) methods[5,6], MM-GBSA/MM-PBSA (Molecular-Mechanics based Generalized-Born Surface Area/Poisson-Boltzmann Surface Area) approaches[7–9], and molecular docking methods[10–13] are widely used to accurately or relatively predict the ligand binding pose and binding strength with varying computational cost. Notably, for these prediction 1) physics-based, 2) empirical, 3) knowledge-based, and 4) descriptor-based scoring functions have been used[14,15]. These scoring functions are predetermined additive functional form and are implemented in popular molecular docking programs such as AutoDock Vina[13], Glide-Score[10], and Surflex-Dock Score[16]. Though these scoring functions were conveniently and widely used, sometimes they failed to discriminate the binders with non-binders. Furthermore, though scoring functions as a free energy surrogate of a protein-ligand complex have failed to provide the collinearity, researchers hope high correlation between a docking score and a pose with key interactive residues. Thus, the additional functions also have been included in the scoring functions of docking programs[17–19].

In recent years, machine learning, and deep learning methods have achieved remarkable success in image and speech recognition, medical diagnosis, learning associations, classification, and regression analysis[20,21]. Machine learning methods also have now been used to predict the ligand binding strength by replacing linear scoring functions. These methods can be characterized by explicit and implicit features derived from either protein, ligands, or protein-ligand pairs[22]. First of all, the ligand binding strength depends on the vector summation of intermolecular interaction features such as Hydrophobic, H-bond, π-π, cation-π, and charge interaction. Thus, several methods have been developed for extracting these features, all in different ways in the featurization process[23–25]. These features are either derived from an atom-centered or grid-based approach. Gomes at al.[23] represented the structure of protein and ligands as a combination of neighbor list and atom types for featurization in their atom-centered

approach for their deep learning. Wallach et al.[26] and Ragoza et al.[27] represented the protein-ligand complex in a 3D-grid box to extract the various interactions for the classification task. AtomNet[26], Pafnucy[28], K_{DEEP}[24], RosENet[29] are some recent examples using the atom-based or grid-based approach to extract the features to build CNN model. Although their state-of-art deep learning predictors showed robust performance with statistical significance on their tested protein-ligand databases, to interpret deep learning models is a challenge and a problem hampering their further progress. Furthermore, the complexity of the featurization process more impedes researchers' intuitional understanding on predicted results (especially, the relationship of a drug structure or its pose with its binding affinity) during their decision making. Therefore, a simple and interpretable featurization is required to well explain an effective binding mode together with its predictive model, which has a reliable predictive power.

Secondly, diverse representation of protein-ligand interactions has been performed. For examples, there are algebraic graph theory (AGL-Score[30]), multiple layer of specific element pairs (Onion-net[31]), which show the local and non-local interaction (distance-dependent), protein-ligand extended connectivity fingerprint (PLEC-NN[32]), docking features (in ΔvinaRF_{20}[33]), and predefined PL-interaction (ID-Score[14]). Furthermore, molecular fingerprints, popular features in ligand-based virtual screening, have applied to encoding protein-ligand interactions. The fingerprint pattern can help to annotate the protein families into their bound ligands. Recently, versatile tools, which captures the protein-ligand binding interaction information as fingerprint pattern, with a binary string of 1 (if an interaction is present) or 0 (if an interaction is absent), have been developed such as PLIP (Protein-Ligand Interaction Profiler)[34], IFP (Interaction Fingerprint Pattern)[35], SIFt (Structural Interaction Fingerprints)[36] and APIF (Atom-pair based interaction fingerprint)[37]. Among these tools, IFP has gained considerable popularity and suitability in drug discovery experiments such as i) post-processing the docking result[38], ii) prioritizing the scaffold pose[39], iii) predicting the ligand pose[40], iv) selecting the virtual hits[41], v) binding site comparison[42] and v) to design target-oriented libraries[43]. The notable merit of IFP is the on-the-fly calculation of the interactions based on a certain set of rules (atom-types) and geometric relationship (distances, angles) between the interacting atoms from proteins and ligands [35]. Based on the IFP, Chupakhin et al. built a neural network model to predict the ligand-binding mode for chosen three targets (CDK2, p38-α, HSP90-α).[42] Unfortunately, their model was limited to the three

target proteins.

Thirdly, in addition to protein-ligand interaction features, some scoring functions additionally uses some features from ligand structures (eg. Autodock,[12] Autodock Vina,[13] and NNScore 2.0[44]). Lin et al. reported that ligand features can show effective polypharmacological relationship between target proteins.[45] Boyles, et al. predicted the ligand-binding affinity using combined ligand features (derived from RDKit) and different scoring functions (RF-Score, NNScore, Vina) features.[3] Notably, various shape of ligands, from linear to multiple ring systems, can exhibit different binding affinity strength even in a homologous protein class. This study suggested the importance of ligand features in binding affinity prediction model[46,47]. Thus, taking into account, the combination of ligand features along with interaction-based features can further improve the performance of the scoring function.

Based on these reported characteristics and drawbacks, we were motivated in the simplicity of usage, more interpretable features to directly explain protein-ligand binding, and ligand features as capturing polypharmacology. For this purpose, the protein-ligand interaction-based fingerprint and ligand features were generated using IChem and SMF (Substructural Molecular Fragment) tools respectively. From these features, our best prediction model was realized in SMPLIP-Score (Substructural Molecular and Protein-Ligand Interaction Pattern-Score) as shown in **Figure 1**. This work aimed at addressing three points: a) how much reliable predictive models can be built from IFP features of protein-ligand complex and SMF of ligand, b) how much efficient this featurization method is through the performance comparison of SMPLIP-Score with reported state-of-art models, c) how much robust or effective our models are by comparing the predictive performance between simulated docking poses and experimental crystal poses.

[Insert Figure 1 here]

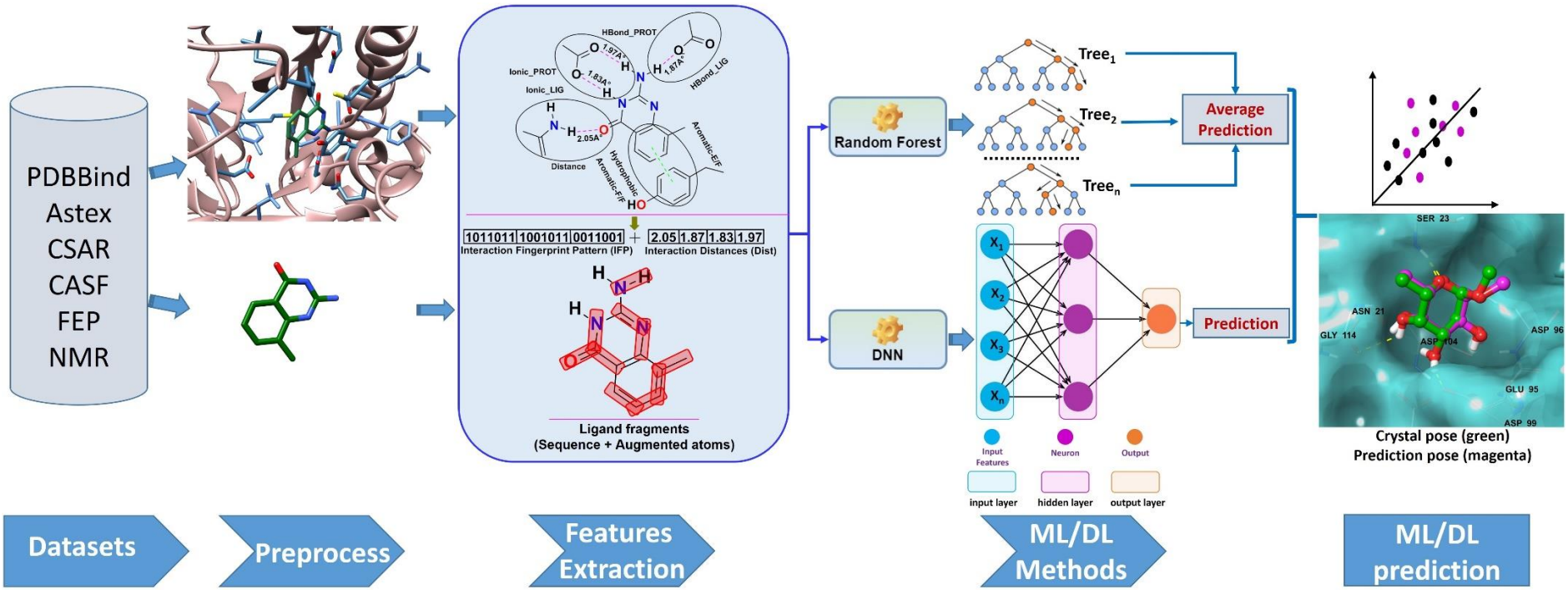

Figure 1: SMPLIP-Score workflow for binding affinity prediction. Publically available protein-ligand binding datasets were used to extract the information encoding the Interaction Fingerprint Pattern (IFP), Interaction Distances (Int-Dist), and ligand fragments to which machine learning or deep learning method may be applied for affinity prediction.

2. Materials and Methods:

2.1. Dataset for learning: The protein-ligand database was downloaded from www.pdbbind-cn.org (PDBbind version 2015) and that includes proteins and ligands in *.pdb and .mol2/.sdf file format respectively assigned with PDB ID. The PDBbind database is a standard dataset that has previously been used to develop the ligand-binding affinity prediction model[24,25,29,31]. This dataset was categorized into three overlapping sets i.e. General, refined, and core sets with a total number of compounds for each set comprised of 11908, 3706, and 195 proteins and ligands. These compounds were resolved by either X-ray or NMR methods with resolution ranges from 0.75 Å to 4.60 Å except NMR solutions. The binding strength of each ligand to proteins was measured in IC_{50}, K_d, and K_i and reported in mM, μM, and nM units respectively. In the present work, we used the refined and core set with the binding strength in only K_d or K_i. The overlapping complexes between the refined set and core set were removed from the refined set. The refined set was randomly partitioned with a ratio of 80:20 into a train and valid set (A total of 6 subsets of the train and valid sets was created with a different random seed). The training and validation were performed on a refined set and core set used to test the prediction performance of the developed models.

2.2. Dataset preprocessing: The PDBbind dataset was cleaned and processed using KNIME analytic platform[48]. The silent features of the KNIME analytic platform that, without any previous background in knowing programming languages, any user can perform several programming tasks using several nodes. We have created datasets cleaning KNIME workflow that includes nodes from the Community and Schrödinger suite[49], which took the protein path as input, iteratively read input PDB structure, add the H-atoms, correct the bond order, remove the water molecules from protein files and convert them into *.mol2 files. During the preprocessing, the protein file with a resolution of < 2.5 Å was retained, so that the well-resolved protein structure should be used for feature construction (**Figure S1**). After preprocessing steps, a total of 3481 and 180 were retained for featurization from refined and core set respectively. **Figure S2** shows the characterization of the input PDBbind dataset after preprocessing.

2.3. Features construction: We used two types of features, which represent the ligand's active-binding site environment; one is based on interaction pattern observed between a ligand and protein's binding site amino acid residues and the second is based on ligand fragments based on atoms and neighboring atoms. The IFP between each protein-ligand complex was calculated

using the IChem tool, which is based on OEChem TK[35]. The Interaction Fingerprint pattern write seven bit-string information for ligand's binding site amino acid residues. These seven bit-string information corresponds to Hydrophobic, aromatic face-to-face, aromatic edge-to-face, H-bond accepted by ligand, hydrogen bond donated by ligand, ionic bond with ligand negatively charged, ionic bond with ligand positively charged under standard geometric rules. So, if any amino acid residues within the binding pocket have formed any interactions with the ligand atoms, then the respective interactions were assigned 1 otherwise 0. Since we have 20 types of standard amino acid in the biological system and considering the 7 interaction information, thus a matrix of 20 x 7 = 140 was constructed. Notable to mention, the favorable interactions are only formed when two interacting atoms from proteins and ligands are in close proximity to each other. These interactions are distance depended and governed by spatial and geometric rules. Thus the interaction distance information from the protein-ligand interaction pair was also extracted and combined with an interaction fingerprint pattern that equals a total of 280 lengths. The Refined and core set of PDBbind database (release 2015) contains more than three thousand co-crystal ligands, which are structurally diverse and varied in shapes and sizes with ligand length up to 47Å. Thus considering the different shapes and sizes of ligands, we used the SMF program to calculate the substructural fragment descriptors for the ligands[50]. Herein two types of substructural fragment descriptors were calculated: the sequence of atoms with a path length up to 6 atoms and atoms with their neighbors and both contribute to a total of 2282 substructural fragment descriptors. The feature construction method used in this work is shown in **Figure 2**.

[Insert Figure 2 here]

2.4. Machine learning and deep learning methods: In recent years, the various branches of artificial intelligence i.e. machine learning and deep learning have gained wide applicability in drug design and discovery that includes, predicting the numerous properties of a set of ligands or predicting affinity of the bound ligand in the protein binding pocket. Based on successful prediction performance, in this work, we used the random forest (RF) as an ensemble learning method and deep neural network (DNN) as a deep learning method.

2.4.1. Ensemble learning: The ensemble learning-based methods combine several models, which were built individually to improve the prediction performance. The ensemble learning method can be divided into two categories: bagging and boosting. The bagging is also called bootstrap aggregation, where multiples sample sets are produced and these sets are trained by

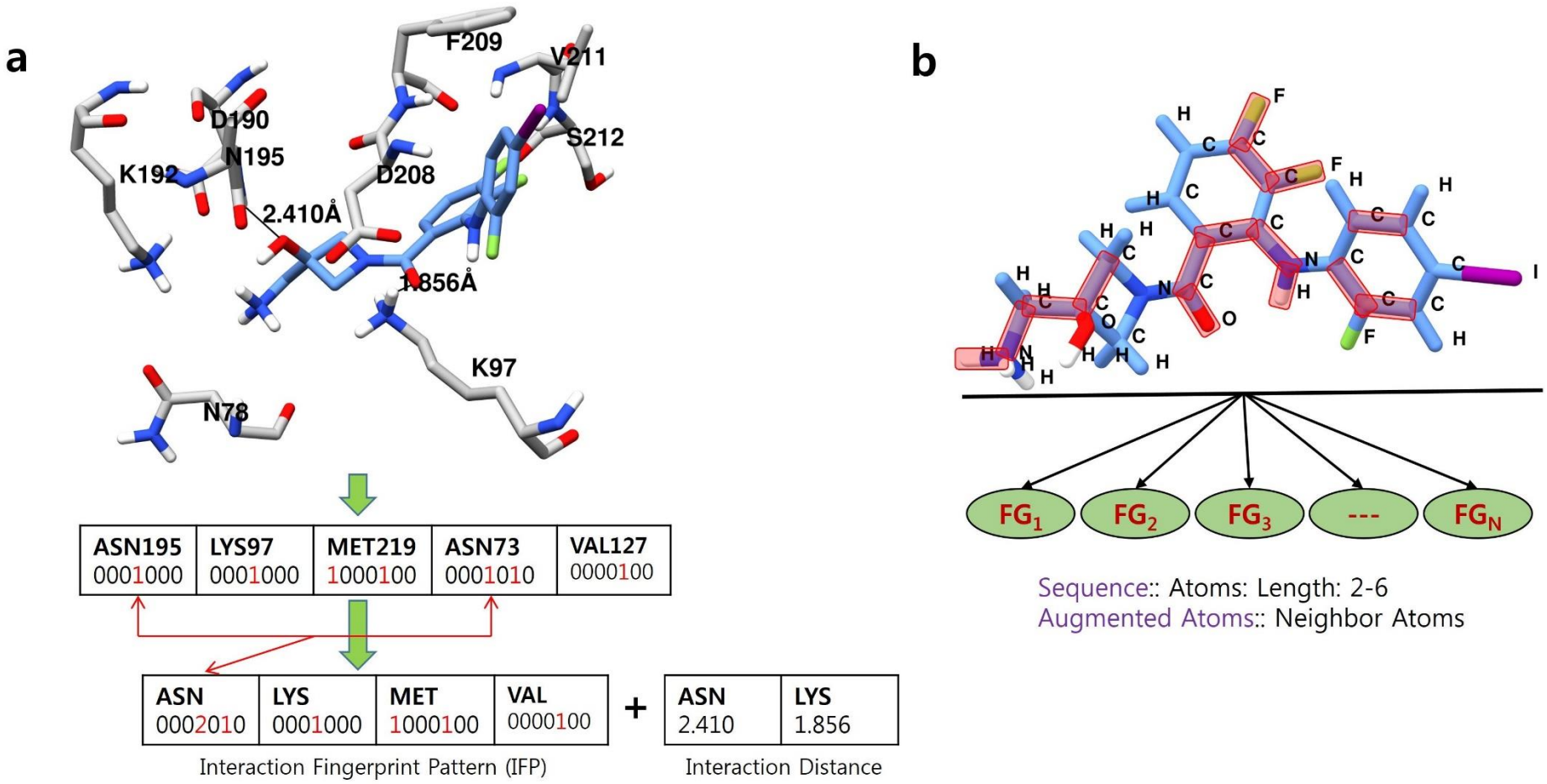

Figure 2: a) Interaction Fingerprint Pattern (IFP) generation. IFP represents the concatenation of interaction values from the same amino acid in a matrix with a fixed size of 140. Similarly, the interaction distances were calculated. **b)** The ligand fragmentation pattern (FG_1, FG_2, …, FG_N) for the input ligand, which corresponds to atoms sequence and augmented atoms.

individual learners. The main advantage of using bagging algorithms, as it decreases the prediction variance of the model and improves the accuracy of the ensemble. In this study, we used the RF as a bagging algorithm to build the regression model. RF is an ensemble of the decision tree (B) $\{T_1(X),....,T_B(X)\}$, as a base learning model, where $X=\{x_1,...,x_p\}$ is a p-dimensional vector of molecular properties. The ensemble learning produces output $\{\hat{Y}_1=T_1(X),....,\hat{Y}_B=T_B(X)\}$, where $\hat{Y}_b, b=1,...,B$, is the prediction for a molecule by the bth tree. The output obtained from all tress are aggregated to produce one final prediction for each molecule[51]. In regression modeling, it is the average of the individual tree predictions. The Scikit-learn package (version 0.22)[52] used to train and build the RF model. To build the models, different parameters from the RF were used (n_estimators = 100, 200, 300, 400, 500, and max_features =’auto’ and ‘sqrt’). The random_state parameters were fixed to different seed numbers during training and to reproduce the statistical result. (**See Table S1 for the selected random_state values**).

2.4.2. Deep Neural Network: In the last decades, deep learning has been used for image classification, video processing to speech recognition, and natural language processing. Also, these methods have already been used in drug design and discovery applications over the last few years[53,54]. A typical deep neural network (DNN) method uses an artificial neural network (ANN) to make a decision or solve the problem. The standard DNN architecture includes the input layer, hidden layers, and output layers. In this work, we used the DNN to build the model and perform the predictions. The DNN model was trained using Keras (version 2.2.4) with the Tensorflow backend module[55]. During the DNN training the early stopping criteria (Δ_{loss}), dropout, batch normalization, and L2 regularization was adopted to avoid the over-fitting of the DNN model. The DNN model was trained with tunable parameters that include dropout regularization (0.1, to 0.6), alpha (0.1 to 1.0), and batch sizes of 64, 128, and 256. During the learning, the best model was obtained as learning enters an over-fitting stage, which is based on a modified Loss formula (LOSS), adopted from Zheng et al.[31].

$$LOSS = \alpha\,(1\text{-}PCC) + (1-\alpha)\,RMSE$$

Furthermore, some additional parameters such as the learning rate of 0.001, decay constant of $1e^{-6,}$ and a momentum of 0.9, were kept constant during the learning. The ReLU as activation functions were used at each layer, and stochastic gradient descent (SGD) optimizer was selected to search for optimal weights in the model.

2.4.3. Evaluation metrics: The quality and performance of each machine learning or deep learning models were assessed using various evaluation metrics including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Pearson Correlation Coefficient (PCC). The detailed information is as follows;

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(BA_{predict} - BA_{true})^2}$$

Where RMSE measures the average magnitude of the error and it represents the square root of the average of squared differences between the predicted values and the experimental values. Mean absolute error (MAE) is another evaluation metrics and it differs from RMSE, as MAE is the average of the summed absolute differences of the predicted values to the experimental values.

$$MAE = \frac{1}{N}\sum |\, BA_{predict} - BA_{true} \,|$$

The Pearson correlation coefficient (PCC), was used to estimates the linear relationship between the predicted and experimental values. This metric also assesses the scoring function ability of the model.

$$PCC(R) = \frac{\sum[(BA_{predict} - \overline{BA}_{predict})(BA_{true} - \overline{BA}_{predict})]}{(SD_{\overline{BA}_{predict}})(SD_{\overline{BA}_{true}})}$$

2.4.4. Benchmark datasets for evaluation: We used five different benchmark datasets to assess the accuracy and efficiency of the SMPLIP-Score. Previously these datasets used by many researchers to measure the quality and performance of their ML/DL/CNN model.

2.4.4.1. Astex diverse dataset: These datasets comprised of diverse and high-quality protein-ligand pairs. Mooij et al.[56] manually curated these datasets to validate the protein-ligand docking program. After checking and comparing the overlapped protein-ligand pairs, 15 protein-ligand pairs were left for further processing.

2.4.4.2. Community structure-activity resource (CSAR): These in-house datasets were collected and managed by the University of Michigan. Among the CSAR datasets, we used CSAR-HiQ-NRC (Set01 and Set02) benchmark dataset from http://csardock.org.The original input dataset contains 176 and 167 protein-ligand pairs for Set01 and Set02 respectively with binding affinity data in K_d/K_i. After comparing and excluding the overlapped protein-ligand pairs from the refined set, a total of 56 and 64 pairs remained in Set01 and Set02 for further processing.

2.4.4.3. Comparative assessment of scoring functions (CASF): The CASF datasets are part of the PDBbind dataset and consists of a collection of a high-quality set of protein-ligand complexes provided to assess the scoring functions. We used CASF-2016 (http://www.pdbbind-cn.org/casf.php) and it comprises a total of 285 protein-ligand pairs with their experimental activity in K_d/K_i. A total of 122 protein-ligand pairs were selected after excluding the overlapped pairs from the training set.

2.4.4.4. FEP dataset: Originally, this dataset comprised of eight targets (BACE, CDK2, JNK1, MCL1, p38, PTP1B, Thrombin, and Tyk2) and contains a total of 199 compounds selected from various literature by Wang et al.[6] to predict the relative ligand-binding affinity using the free-energy perturbation method. Nevertheless, while there are 199 compounds for eight targets, binding affinity in K_i has only been reported for five targets (BACE, MCL1, PTP1B, Thrombin, and Tyk2). Therefore, in our work, we excluded CDK2, JNK1, and p38 from our FEP dataset. The remaining five targets were not part of the refined set, so all the reported compounds 36, 42, 22, 11, and 16 were selected respectively for each target BACE, MCL1, PTP1B, Thrombin, and Tyk2.

2.4.4.5. NMR PDBbind dataset: We also tested the performance of SMPLIP-Score on protein-ligand pairs that are resolved by NMR. The refined set and core set lacks structure resolved by NMR, thus were obtained from the general set. A total of 191 protein-ligand pairs were selected from the General set.

3. Results & Discussions:

We built the predictive models using IFP and SMF features, optimized the models, and evaluated their prediction power on the benchmark dataset. We further tested the robustness and effectiveness of our best models using the poses (input features) derived from the molecular docking simulations. We used the RF as an ensemble method and DNN as a deep learning method to build the predictive models, SMPLIP-RF and SMPLIP-DNN.

3.1. SMPLIP-RF Model: We first investigated the use of the single feature or combined features on the prediction performance of the model using the RF method. We trained six different models (with a different random seed) and the different combinations of feature(s), n_estimators, and max_features option were used for each model. All the parameters were kept on default in the RF models, except max_features ('auto' or 'sqrt') and n_estimators (100, 200, 300, 400, and 500). **Table S1** and **Figure S3-S10** report statistical results and the best model

(based on lowest RMSE reported on test data) were selected for each feature or combined features after comparing all the statistical results for different sets **(Table 1** and **Figure 3)**.

[Insert Table1 and Figure 3 here]

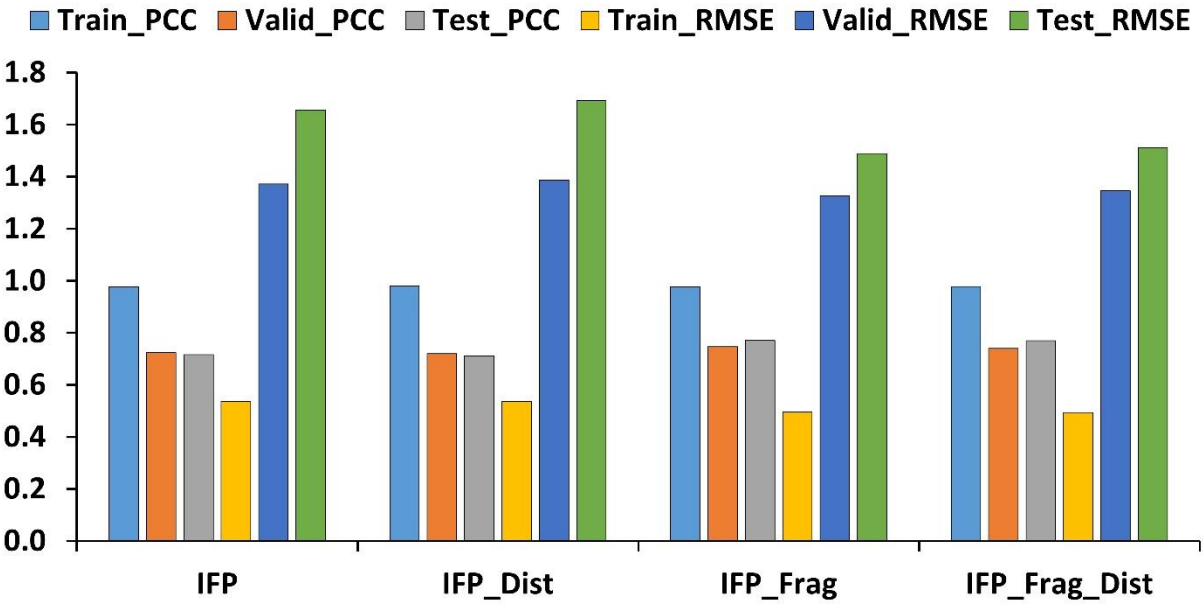

Figure 3: The bar plot for statistical comparison from different feature(s) combination. The PCC (Pearson-Correlation-Coefficient) and RMSE (Root-Mean-Squared-Error) were compared from the train, valid, and test data.

The RF model with the IFP feature alone has achieved PCC for the training data to 0.977 with an RMSE of 0.537. On the validation data, this model has a prediction power of 0.724 (PCC) with an RMSE of 1.372. Since validation data was from refined data, so we further evaluated its prediction power on the core data as a test set. Contrary to the validation data result, the IFP feature-based model for test set has achieved a PCC of 0.716 and an RMSE of 1.656 with a significance level (p_value) of 1.49E-29. Compared to the PCC of reported models for core set (n = 180) as a test set: Vina (0.676), RF-Score (0.710), RF-Score-v3 (0.746) and NNScore 2.0 (0.751), our model (0.716) showed obviously better performance than Vina, RF-Score (**Table1**). Furthermore, it is rationalized that the shape and size of bound ligands can be useful information with the improved predictive power (particularly, test data MAE: 1.227, RMSE: 1.489) through the significance of the ligand fragments that belong to neighbor atoms and augmented atoms as features. The model (IFP+Frag: 0.771) has higher PCC and comparable performance than Boyles's report (Vina + RDKit: 0.749; RF-Score + RDKit: 0.778), though being slightly lower than models from (RF-Score-v3 + RDKit: 0.780; NNScore 2.0 + RDKit: 0.786)[3].

Table 1: The statistical performance of **SMPLIP-RF** models on **PDBbind (Release 2015)** according to different features compositions

Features	Train				Valid				Test			
	RMSE	MAE	PCC	p_Value	RMSE	MAE	PCC	p_Value	RMSE	MAE	PCC	p_Value
IFP	0.537	0.420	0.977	0	1.372	1.066	0.724	4.17E-115	1.656	1.349	0.716	1.49E-29
IFP + Int-Dist	0.536	0.422	0.980	0	1.387	1.093	0.720	2.08E-112	1.692	1.388	0.711	5.15E-29
IFP + Frag	**0.496**	**0.381**	**0.977**	**0**	**1.327**	**1.035**	**0.747**	**2.62E-125**	**1.489**	**1.227**	**0.771**	**8.71E-37**
IFP + Int-Dist + Frag	0.494	0.382	0.978	0	1.346	1.054	0.740	8.07E-122	1.512	1.244	0.770	1.43E-36

Note: The Refined set (n =3481) used for training and validation, and core set (n=180) as a test data. The boldface represents the model with better statistics from different features combination and max_features options. **Abbreviations: RMSE:** Root-Mean-Square-Error; **MAE:** Mean Absolute Error; **PCC:** Pearson Correlation Coefficient; **p_value:** p_value for statistical significance

Table 2: The statistical performance of **SMPLIP-DNN** models on **PDBbind (Release 2015)** according to different features compositions

Features	Train			Valid			Test		
	LOSS	RMSE	PCC	LOSS	RMSE	PCC	LOSS	RMSE	PCC
IFP	0.563	0.871	0.899	1.019	1.483	0.678	1.032	1.538	0.726
IFP + Int-Dist	0.472	0.990	0.873	0.775	1.468	0.687	0.805	1.582	0.713
IFP + Frag	**0.209**	**0.595**	**0.956**	**0.631**	**1.402**	**0.699**	**0.646**	**1.530**	**0.733**
IFP + Int-Dist + Frag	0.212	0.403	0.979	0.834	1.400	0.733	0.923	1.559	0.714

Note: The Refined set (n =3481) used for training and validation, and core set (n=180) as a test data. The boldface represents the model with better statistics from different features combination.

On observation of improved statistics for IFP+Frag features, we further combined all three features (IFP+Int-Dist+Frag) to measure the prediction performance. Comparing the prediction power on test data, the random forest models with a different combination of feature(s), the features from IFP+Frag at n_estimators = 100 and max_features = 'auto' showed the best performance with lowest RMSE (1.489) and higher PCC (0.771) at significance value of 8.71E-37 as compared to other feature(s) models. Notably, this result suggests that facile recognizable SMPLIP features (especially, IFP+Frag) can show enough predictive power, never inferior to known scoring using complex features such as atom-centered or grid-based.

3.2. SMPLIP-DNN Model: We further evaluated prediction performance of SMPLIP features using the DNN method. For SMPLIP-DNN model, the same set of training, validation, and test data were used to compare the statistical performance from the random forest model. The predictive model was built in our DNN training with screened hyperparameter values for dropout, alpha, and batch-sizes. The models for each feature(s) or combined features with optimized hyperparameter values are shown in **Table S2 and Figure S11-S14.** In **Table S2**, the best model at each epoch was based on modified Loss formula, which prevents overfitting of the model. The best model for each feature(s) combination is shown in **Table 2.**

[Insert Table 2 here]

The IFP feature showed PCC of 0.899, 0.678, and 0.726, while RMSE was found to be 0.874, 1.483 and 1.538 for the train, valid and test data respectively. Like the random forest model, although the performance of the IFP feature model on training and validation data was not at par, this model performed well for test data in terms of RMSE (1.538) and PCC (0.726), with LOSS statistics at 1.032. With improved statistics on test data for IFP features, we further combined the Interaction distance features with IFP. Although the combined model did not improve its statistics for PCC (0.713) and RMSE (1.582) on test data, this model did perform better for SMPLIP-RF (IFP+Int-Dist). Notably to mention here that, in the RF model, the IFP+Frag features have higher predictive power for test data, so we further built the DNN model with these features. Even if SMPLIP-DNN (IFP+Frag) did not present dramatic improvement, it has the highest prediction power for test data (PCC: 0.733; RMSE: 1.530) than other DNN models. This performance was stable against epochs as shown in **Figure S15** to illustrates the comparison of PCC and RMSE of the train and valid data against epochs. The

best model was obtained for a batch size of 64; dropout of 0.1 and alpha value of 0.7 at 129 epochs.

3.3. Comparison of predictive performance with known models: We built SMPLIP-RF and SMPLIP-DNN models to predict the ligand-binding affinity. Under our investigated condition (different partitioning of the dataset, different hyperparameter options, and chosen features), all models presented statistical reliability with the distribution of PCC and RMSE values, from independently trained models (**Table S1 and S2**), and with their variance analysis (**Table S3**). Although IFP+Frag feature-based model outperforms on predicting a binding affinity in both random forest and DNN method, the statistical performance (PCC and RMSE) of the random forest model on test data was better than DNN model. **Figure 4** shows the correlation of predicted and experimental binding affinity from SMPLIP-RF (IFP+Frag) as a scatter plot for training, valid, and test data.

[Insert Figure 4 here]

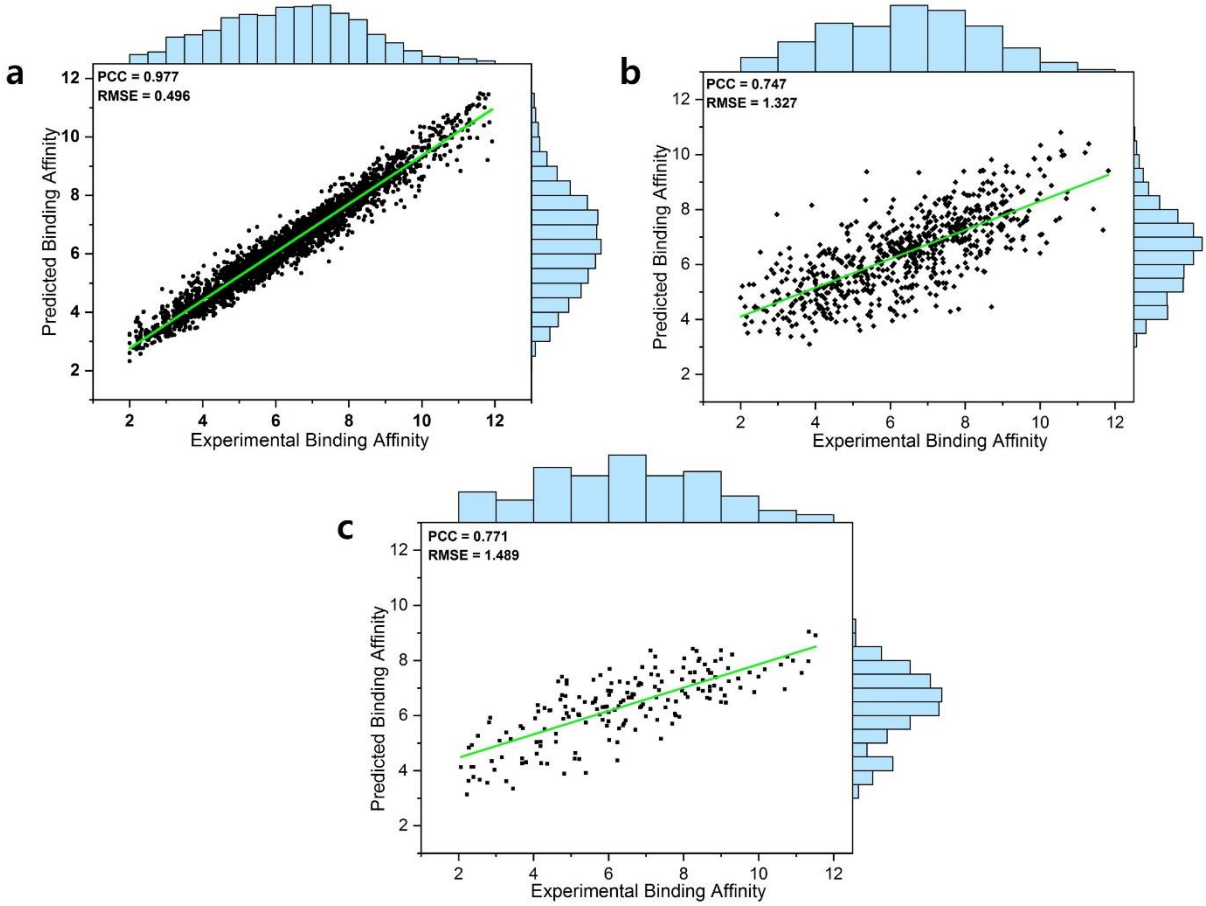

Figure 4: 2D-Scatter plot between predicted and experimental of SMPLIP-RF (IFP+Frag) for **a)** training, **b)** valid, and **c)** test data using

In consequence, the best SMPLIP-Score made us motivated in assessing its scoring and ranking power on other predictive models having different featurization methods. For this purpose, the performance of SMPLIP-Score was further compared with other state-of-art models (**Table 3**). Firstly, SMPLIP-Score was compared with Gomes's featurizer methods,[23] Atomic Convolution Neural Network (ACNN), GRID-RF, GRID-NN, GCNN, ECFP-RF, ECFP-NN. Their featurizer performs a 3-D spatial convolution operation to learn atomic-level chemical interactions from protein-ligand pair. From those featurization, they maximally achieved PCC of 0.739 (GRID-RF) and 0.669 (ACNN) for test set based on refined set and core set respectively. Obviously, our SMPILP-Score performed better than GRID-RF and ACNN models (**Table 3**). Secondly, Cang's [57] algebraic topology featurization were compared with SMPLIP-Score. It achieved PCC of 0.797 (TopBP-ML) and 0.799 (TopBP-DL) for core data as a test set. Despite their slightly higher PCC values than SMPLIP-Score (0.771), more notably, the RMSE of SMPLIP-Score (1.489) showed distinctly lower than their RMSE of 1.99 (TopBP-ML) and 1.91(TopBL-DL) to suggest better performance. Thirdly, Nguyen's rigidity index score[58] (RI-Score) reported a higher RMSE of 2.051 on core data as a test set. Dziubinska's 3D grid-based CNN model, pafnucy model [28] also was inferior to our SMPLIP-Score.

In turn, Wójcikowski's circular fingerprint featurization (PLEC-linear and PLEC-NN) having different depth level (protein depth of 5, and a ligand depth of 1) have achieved (PCC: 0.757; SD: 1.47), and (PCC: 0.774; SD: 1.43) respectively for PLEC-linear and PLEC-NN on PDBbind v.2013 core set. In the spite of similar performance, our SMPLIP-Score has less number of features than PLEC-linear and PLEC-NN. Furthermore, SMPLIP-Score is readily understandable based on interpretability of embedded feature matrix. Constantly, the SMPLIP-Score has even outperformed popular machine learning scoring functions, such as RF-Score-v3[59] (PCC: 0.74; SD: 1.51), X-Score[60] (PCC: 0.614; SD: 1.78), Autodock Vina[61] (PCC: 0.54; SD: 1.90) and Autodock[61] (PCC: 0.54; SD: 1.91) on the PDBbind v.2013 core set. Notably all the models/methods compared here are different in terms of their featurization process and require state-of-art architecture (ML/DL/CNN) to achieve predictive power. Eventually, SMPLIP-Score proved its cost-effectiveness based on these comparison: (1) computational complexity, (2) facile featurization using known tools (IChem & SMF), (3) easy interpretable for non-professional from embedded feature matrix to docking pose, and (4) predictive power

equal to state-of-art model.

[Insert Table 3 here]

Table 3: Performance comparison of SMPLIP-Score with reported models on the **PDBbind v.2015** dataset.

ML/DL Method	Refined Set		Core Set	Ref.
	Train	**Valid**	**Test**	
SMPLIP-Score	0.977 (0.496)	0.747 (1.327)	0.771 (1.489)	-
GRID-RF	0.981	0.739	-	Gomes et al.[23]
ACNN	-	-	0.669	
TopBP-ML	-	-	0.797 (1.99)	Cang et al.[57]
TopBP-DL	-	-	0.799 (1.91)	
RI-Score	-	-	0.782 (2.051)	Nguyen et al.[58]
Pafnucy	0.77(1.21)	0.72 (1.44)	0.70 (1.62)	Stepniewska Dziubinska et al.[28]
PLEC-Linear	-	-	0.757 (1.47)*	Wójcikowski et al.[32]
PLEC-NN	-	-	0.774 (1.43)*	Wójcikowski et al.[32]
RF-Score-v3	-	-	0.74 (1.51)*	Wójcikowski et al.[59]
X-Score	-	-	0.614 (1.78)*	Khamis et al.[60]
Autodock Vina	-	-	0.54 (1.90)*	Gaillard et al.[61]
Autodock	-	-	0.54 (1.91)*	

Note: Pearson correlation coefficients with **RMSE** in parentheses for predictions by different methods. SMPLIP-Score: Interaction Fingerprint Pattern and Ligand Fragment-based random forest (RF) model; GRID-RF: Grid featurizer based Random Forest; ACNN: Atomic Convolutional Neural Network model; TopBP: Topology based model; RI-Score: Rigidity Index based score; *values in parenthesis represent the standard deviation (SD).

3.4. Generalization of SMPLIP-Score evaluated through benchmark datasets: The generalization of the SMPLIP-Score was tested using additional benchmark datasets. These datasets were previously used by other researchers to evaluate the performance of their ML/DL/CNN models. The benchmark dataset used here belongs to the Astex Diverse set, CSAR NRC HiQ sets, CASF-2016, FEP, and PDBBind NMR dataset. SMPLIP-RF based on IFP+Frag features with n_estimators = 100; “max_features = ‘auto” was used for generalized

evaluation on the benchmark dataset. **Table 4** lists the comparative assessment on calculated scoring metrics (PCC, RMSE, and Sp) for these datasets. The scatter plot for predicted binding affinity against the experimental binding affinity for these benchmark datasets shown in **Figure S16**. The original Astex Diverse dataset consists of 93 pairs of protein-ligand and after removing the overlapped protein-ligand pair from PDBbind refined set, a small sets of 15 protein-ligand pair were left for prediction measurement. We achieved a PCC, Sp and RMSE of 0.724, 0.764, and 1.177 for this dataset respectively. The SMPLIP-Score has achieved a little lower scoring (**DeepAtom**: PCC = 0.768; RMSE = 1.027) and better scoring and ranking (**RF-Score:** PCC = 0.710; RMSE = 1.144; **Pafnucy:** PCC = 0.569; RMSE = 1.374; **Res4HTMD:** Sp = 0.41; RMSE = 1.54; **RosENet:** Sp = 0.29; RMSE = 1.84) performance on Astex Diverse dataset with previously reported models[25,28,29,62].

The second selected benchmark of CSAR NRC HiQ dataset consists of both Set01 and Set02. These sets were also used in docking program validation. After checking the overlapped protein-ligand pair, 56 and 64 pairs left for Set01 and Set02 respectively for predictive evaluation. The SMPLIP-Score has achieved the following prediction performance on Set01 (PCC = 0.785; Sp = 0.761; RMSE = 1.903) and Set02 (PCC = 0.803; Sp = 0.823; RMSE = 1.475). Compared with $\mathbf{K_{DEEP}}$ a CNN model (PCC = 0.72; RMSE = 2.09 (Set01); PCC = 0.65; RMSE = 1.92 (Set02)), the SMPLIP-Score has better performance. However, a little lower performance was observed in the form of spearman correlation (Sp) as compared to models from **Res4HTMD:** Sp = 0.84; RMSE = 1.75 (Set01); Sp = 0.83; RMSE = 1.34 (Set02); and **RosENet:** Sp = 0.87; RMSE = 1.71 (Set01); Sp = 0.85; RMSE = 1.38 (Set02) in predicting the binding affinity of CSAR NRC HiQ sets[24,29].

Another benchmark dataset, popularly known as FEP dataset, was selected from the work of Wang et al. These datasets were used to predict the relative binding potency using modern free-energy calculation protocol and forcefield. These datasets comprised of ligands from BACE, MCL1, PTP1B, Thrombin and Tyk2 targets. The predictive performance for these dataset on SMPLIP-Score were (**BACE:** PCC = 0.239; Sp = 0.250; RMSE = 0.639; **MCL1:** PCC = 0.077; Sp = 0.146; RMSE = 1.045; **PTP1B:** PCC = 0.634, Sp = 0.680; RMSE = 0.768; **Thrombin:** PCC = -0.645; Sp = -0.536; RMSE = 0.962; **Tyk2:** PCC = 0.469; Sp = 0.546; RMSE = 0.859). Notably, except to Thrombin, all prediction performance was positive, and prediction ranking was PTP1B > Tyk2 > BACE > MCL1 > Thrombin. Comparison of the

results of predictions with other methods ($\mathbf{K_{DEEP}}$: (**BACE**): PCC = -0.06; RMSE = 0.84; (**MCL1**): PCC = 0.34; RMSE = 1.04; (**PTP1B**): PCC = 0.58; RMSE = 0.93; (**Thrombin**): PCC = 0.58; RMSE = 0.44; (**Tyk2**): PCC = -0.22; RMSE = 1.13; **Res4HTMD:** (**BACE**): Sp = -0.19; RMSE = 1.27; (**MCL1**): Sp = 0.45; RMSE = 1.1; (**PTP1B**): Sp = 0.55; RMSE = 0.88; (**Thrombin**): Sp = 0.45; RMSE = 0.83; (**Tyk2**): Sp = 0.71; RMSE = 0.76), revealed that, for all FEP dataset, SMPLIP-Score performed better than K_{DEEP} model, while for BACE, PTP1B targets, Res4HTMD performed better[24,29].

Lastly, we further predicted the ligand-binding affinity of the dataset, for which different experimental techniques have been used. A total of 191 protein-ligand pairs derived from the NMR method were selected and predicted for their ligand-binding affinity. Compared with RosENet model (Sp = 0.56; RMSE = 1.37), our model did not predicted well (Sp = 0.234; RMSE = 1.857) the ligand-binding affinity derived from NMR method[29]. Overall, SMPLIP-Score performed very well in benchmark evaluation in most of the cases in predicting the ligand-binding affinity, which affirms, the reliability of our model for a diversified binding affinity prediction dataset and can be further use in the drug virtual screening method.

[Insert Table 4 here]

Table 4: The Prediction performance on benchmark datasets and statistical comparison of SMPLIP-Score with reported models.

Datasets	Models	SETS	PCC	RMSE	MAE	p_Value	Sp	Ref.
Astex Diverse Set	**SMPLIP-Score**	-	0.724	1.177	0.938	0.002	0.764	-
CSAR NRC HiQ		Set01	0.785	1.903	1.500	8.03E-13	0.761	
		Set02	0.803	1.475	1.134	1.54E-15	0.823	
FEP		BACE	0.239	0.639	0.505	0.160	0.250	
		MCL1	0.077	1.045	0.797	0.629	0.146	
		PTP1B	0.634	0.768	0.536	0.002	0.680	
		Thrombin	-0.645	0.962	0.780	0.321	-0.536	
		Tyk2	0.469	0.859	0.655	0.078	0.546	
PDBbind		-	0.209	1.857	1.552	0.004	0.234	

NMR								
Astex Diverse Set	**DeepAtom**	-	0.768	1.027	0.714	-	-	Li et al.[25]
	RF-Score	-	0.710	1.144	0.891	-	-	Ballester et al.[62]
	Pafnucy	-	0.569	1.374	1.110	-	-	Stepniewska Dziubinska et al.[28]
	Res4HTMD	-	-	1.54	-	0.07	0.41	Hassan Harrirou et al.[29]
	RosENet	-	-	1.84	-	0.21	0.29	
CSAR NRC HiQ	**K_{DEPP}**	Set01	0.72	2.09	-	-	-	Jiménez et al.[24]
		Set02	0.65	1.92	-	-	-	
	Res4HTMD	Set01	-	1.75	-	2E-15	0.84	Hassan Harrirou et al.[29]
		Set02	-	1.34	-	3E-13	0.83	
	RosENet	Set01	-	1.71	-	2E-17	0.87	
		Set02	-	1.38	-	2E-14	0.85	
FEP	**K_{DEEP}**	BACE	-0.06	0.84	-	-	-	Jiménez et al.[24]
		MCL1	0.34	1.04	-	-	-	
		PTP1B	0.58	0.93	-	-	-	
		Thrombin	0.58	0.44	-	-	-	
		Tyk2	-0.22	1.13	-	-	-	
	Res4HTMD	BACE	-	1.27	-	0.26	-0.19	Hassan Harrirou et al.[29]
		MCL1	-	1.1	-	2E-3	0.45	
		PTP1B	-	0.88	-	6E-3	0.55	
		Thrombin	-	0.83	-	0.16	0.45	
		Tyk2	-	0.76	-	2E-3	0.71	
PDBbind NMR	**RosENet**	-	-	1.37	-	-	0.56	

Abbreviations: Models: ML/DL method used to build the ligand binding affinity prediction model; **RMSE:** Root-Mean-Square-Error; **MAE:** Mean Absolute Error; **PCC:** Pearson Correlation Coefficient; **p_value:** p_value for statistical significance; **Sp:** Spearman Correlation Coefficient; **RF-Model:** RF parameters includes: n_estimators=500; max_features = “AUTO”.

3.5. Ranking power on benchmark dataset: The SMPLIP-Score was further assessed for its ranking power as indicated by the spearman's correlation coefficient (Sp) on the CASF-2016 benchmark dataset. The CASF benchmark dataset consists of high, medium, and low active crystal and locally optimized pose from each protein target with cluster number. Here, on these poses, two types of the ranking were calculated, first, we calculated the all Sp for the reduced set, and second, the individual Sp for each cluster was calculated, followed by an average of all clusters. The calculated PCC and Sp values for the CASF benchmark dataset are listed in **Table 5**. For the CASF-2016 benchmark dataset, the PCC on crystal pose and locally optimized pose remains the same, while there is a small increase in RMSE. Although the differences in RMSDs for crystal and locally optimized poses were not large, such small differences in prediction/error were expected as any machine/deep learning models are sensitive towards the input features. Notably, the prediction results on crystal pose and minimized pose are not even > 0.1, indicating that our model was less sensitive to minimized pose than crystal. Additionally, the ranking metrics indicate that the average value of Sp for all the cluster is lower than the overall Sp, which suggests that, while each cluster has high, medium and low active compounds, the difference in activity in some clusters may not be too large, resulting in the ranking of compounds prone to change by prediction error. We further compared the prediction performance of SMPLIP-Score on other ML/DL models. The statistical result from models based on Autodock Vina function, and ΔSAS (buried percentage of the solvent-accessible surface) function[63] and $\Delta_{Vina}RF_{20}$[33] listed in Table 5, which shows that the input features used in all the three ML/DL method are different from our features, nevertheless, the performance of our model was above to Autodock Vina function and ΔSAS function except $\Delta_{Vina}RF_{20}$. Notably, $\Delta_{Vina}RF_{20}$ using descriptors derived from Autodock Vina interaction term, ligand-dependent, and bSASA terms, which can result in its superior performance on the benchmark dataset. Nevertheless, our prediction model uses protein-ligand interaction fingerprint and ligand-dependent features for predicting the ligand-binding affinity, and we expect in the future, using additional interaction terms like desolvation, entropy effects, surface and shape matching property may further improve the prediction performance.

[Insert Table 5 here]

Table 5: The evaluation of the **CASF-2016** dataset.

Models	SETS	PCC	RMSE	Spearman (Sp)		Ref.
				All	Cluster Average	
SMPLIP-Score (IFP+Frag Features)	Crystal-Reduced	0.775	1.643	0.784	0.700	-
	Crystal-Minimized	0.775	1.647	0.780	0.682	
Autodock Vina	Crystal-Pose	0.600	-	0.60	0.53	Su et al.[63]
ΔSAS	Crystal-Pose	0.62	-	0.63	0.59	
$\Delta_{Vina}RF_{20}$	Crystal-Pose	0.82	1.27	0.82	0.75	Wang et al.[33]

Note: The crystal-reduced represent the dataset where protein-ligand pairs are obtained after removing the overlapped protein-ligand pair from the refined set. The crystal-minimized represent the dataset where ligands are locally optimized. The crystal-pose represents the experimental pose. **PCC:** Pearson Correlation Coefficient; **RMSE:** Root-Mean-Square-Error; **Sp:** Spearman Correlation Coefficient.

3.6. Robustness and effectiveness of SMPLIP-Score: Ligand-based or target-based discovery depend on the reliability of predicted binding poses either through experimental or computational methods. In particular, molecular docking simulations are the most popular method to generate such poses with a typical RMSD criterion (> 2.0 Å) compared to x-ray crystal poses. In addition, because every predictive model for binding affinity relies on features generated from input poses, robustness of such model is also affected by the input poses. Thus, we studied the robustness of SMPLIP-Score according corresponding to the change of input poses. To measure the robustness of SMPLIP-Score, we conducted docking simulations the selected dataset (PDBbind-Core Set, Astex Diverse Set, and CASF-2016) to generate docking poses. The molecular docking simulation was performed using the Autodock Vina program [13] (Exhaustiveness: 32; Num_modes: 50) and after docking, only poses with RMSD > 2.0 Å from crystal pose was selected as a reliable pose. For PDBbind Core Set, Astex Diverse Set, and CASF-2016, a total of 163, 15, and 100 poses presented RMSD > 2.0 Å. And then SMPLIP features (IFP+Frag) were extracted from molecular docking poses and the ligand-binding affinity were predicted. The prediction results are shown in **Table 6** and the scatter plot in

Figure 5.

[Insert Table 6 and Figure 5 here]

When comparing the statistical performance of models built on docked poses (**Table 6**) with crystal poses (**Table 1 and Table 4**), we found that the reasonable poses (RMSD > 2.0 Å) rarely made an effect on the affinity prediction and our result corroborated earlier work[31]. A slight variation on PCC and RMSE were observed between **Table 6** and **Table 1** (or **Table 4**). In addition, feature matrices encoded by IChem and SMF can be directly compared between docked poses and crystal poses to reveal interpretability of SMPLIP featurization. Thus to analyze such changes, we randomly selected few poses (crystal and docked) for comparison of interactions from these datasets. When two feature matrices were identical for both crystal and docked poses, they were excluded in this discussion. From PDBbind Core Set, the superimposed crystal and docked pose for PDBs 2JDM and 3VH9 and IFP shown in **Figure S17 (a, b),** and **Table S4, S5** respectively. The first selected 2JDM represents the protein-ligand pair from pseudomonas aeruginosa lectin II (PA-IIL) in complex with methyl-a-L-fucopyranoside with the reported experimental binding affinity of 5.4M[64]. Both the crystal pose and docked pose with RMSD of 0.376 show similar interactions (**Table S4**) and thus have the same predicted values of 5.739M. Similarly, another selected PDB, 3VH9 belongs to an Aeromonas proteolytica aminopeptidase enzyme bound with 8-quinolinol with an experimentally determined binding affinity value of 6.2M[65]. The IFP in **Table S5** shows that most of the interactions from crystal pose and docked poses were common, however, additional interactions such as H-bond (Asp117, Asp179) and Hydrophobic (Cys223) have been observed, resulting in an improvement in the predicted binding affinity value from 4.362 (crystal pose) to 4.579 (docked pose). Furthermore, the selected PDBs 1TT1 and 1SQN from Astex Diverse Set shows the binding interaction of superimposed crystal and docked pose in **Figure S18 (a, b)** and IFP in **Table S6, S7** respectively. The PDB 1TT1 is a GluR6 kainate receptor bound to 3-(carboxymethyl)-4-isopropenylproline with 4.19M experimental activity[66]. The binding interaction revealed that its docked pose has almost the same orientations of the groups. Interestingly, the crystal pose and docked pose have the same interactions with the binding site residues even with RMSD of 1.176Å, and both poses were predicted at 6.488M. Similarly, another selected PDB was 1SQN which belongs to the progesterone receptor in complex with norethindrone with the reported activity of 9.4M[67].

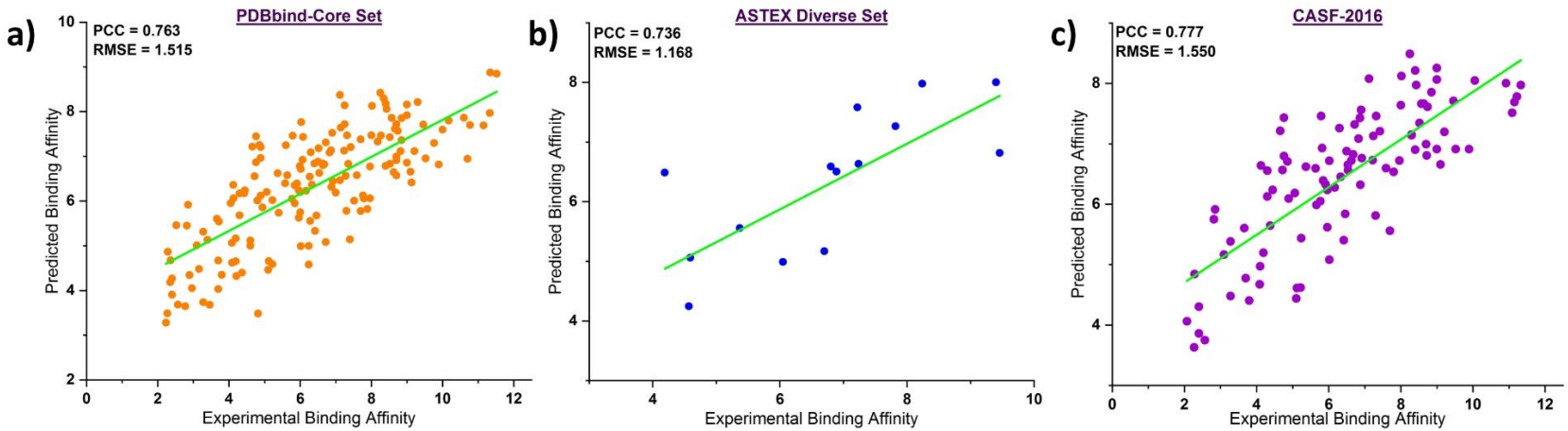

Figure 5: Thee Scatter plots of the predicted binding affinity against the experimental activity from the a) PDBbind Core Set; b) Astex Diverse Set; c) CASF-2016 Set.

Table 6: Prediction result based on simulated docking poses of the selected dataset.

Dataset	PCC	RMSE	MAE	P_value	Sp
PDBbind Core Set	0.763	1.515	1.251	2.70E-32	0.749
Astex Diverse Set	0.736	1.168	0.878	0.003	0.846
CASF-2016	0.779	1.550	1.253	1.49E-21	0.785

Abbreviations: PCC: Pearson Correlation Coefficient; **RMSE:** Root-Mean-Square-Error; **MAE:** Mean Absolute Error; **p_value:** p_value for statistical significance; **Sp:** Spearman Correlation Coefficient; **RF-Model:** Random Forest parameters includes: n_estimators=500; max_features = "AUTO".

Notably, most of the interactions are hydrophobic and crystal pose as well as docked pose have almost identical IFP except for additional hydrophobic (Leu763, Phe778) interactions for the docked pose. While with additional interaction with the docked pose, the predicted values of 8.086M and 8.002M observed for crystal and docked pose respectively. Lastly, we selected PDBs 2Y5H and 1W4O from CASF-2016 Set, to study the interaction fingerprint pattern. The first selected PDB: 2Y5H, belongs to Factor Xa, a serine protease from the blood coagulation cascade crystallized with derivatives of pyrrolo[3,4-a]pyrrolizin[68]. The superimposed crystal and docked pose and interaction fingerprint pattern (IFP) are shown in **Figure S19 (a)** and **Tables S8** respectively. The calculated IFP shows identical interactions for crystal and docked pose with the same predicted values of 7.459M. Another selected PDB: 1W4O is the Ribonuclease-A protein found in bound form with nonnatural 3'-Nucleotides[69] (**Figure S19 (b)**). The calculated IFP (**Table S9**) for docked pose shows that most of the interactions are shared by crystal pose, however, few interactions change their types while interacting with binding site residues, such as Phe_{120} residue formed H-bond with ligand (HBond_LIG) in the docked pose, whereas in crystal pose, the backbone of the same residue formed the HBond_PROT with the ligand. Similarly, in crystal pose, Lys_{41} forms Ionic (Ionic_PROT), while in the docked pose it forms H-bond (HBond_PROT) interactions. These trivial changes reflected in the predicted activity of crystal pose (4.460) and docked pose (4.621). Overall, the comparison of IFP results from both docked and crystal poses indicate that, during the virtual screening (molecular docking) experiments, the identified small molecules must have complementary interactions with crystal pose to be predicted accurately. Nonetheless, the additional interactions observed in protein-ligand complexes may cause changes for the observed prediction in the binding affinity.

4. Conclusions:

Herein, we reported SMPLIP-Score as a robust and effective predictor and compared it with state-of-art featurization processes/methods. SMPLIP features, originated from protein-ligand interaction pattern and ligand features, showed cost effectiveness as well as interpretability of the feature matrix embedded for learning. Most notably, the best SMPLIP features (IFP+Frag) proved scoring power, ranking power, and robustness on various benchmark datasets. Obviously, the comparison between crystal and docked poses proved

robustness of SMPLIP-Score against input poses and their interpretable feature matrices directly could be used to get an insight of a ligand binding to a protein and the integrated description of the binding mode with predicted affinity (of high accuracy) replaceable predictor of current scoring function.

Supplementary information:

Supplementary tables, figures, and methods are available.

Availability of data and materials:

Knime workflow, python code, and refined data will be available in GitHub.
https://github.com/college-of-pharmacy-gachon-university/SMPLIP-Score

Conflict of interests:

The authors confirm that this article content has no conflicts of interest.

Funding:

This study was supported by the Basic Science Research Program of the National Research Foundation of Korea (NRF), which is funded by the Ministry of Education, Science and Technology (No.:2017R1E1A1A01076642, 2020R1I1A1A01074750).

Acknowledgments:

The authors would like to thank OpenEye Scientific Software for providing an academic free license.

Authors' contributions:

M.-h. K. and S. K. conceived and designed the study. S. K. carried out all modeling & data work. M.-h. K. and S. K. analyzed results, wrote the manuscript, and revised it. M.-h. K. provided every research work facility. All authors read and approved the final manuscript.

References:

(1) Gilson, M. K.; Zhou, H.-X. Calculation of Protein-Ligand Binding Affinities. *Annual*

review of biophysics and biomolecular structure **2007**, *36*.

(2) Bajusz, D.; G Ferenczy, G.; M Keseru, G. Structure-Based Virtual Screening Approaches in Kinase-Directed Drug Discovery. *Current topics in medicinal chemistry* **2017**, *17* (20), 2235–2259.

(3) Boyles, F.; Deane, C. M.; Morris, G. M. Learning from the Ligand: Using Ligand-Based Features to Improve Binding Affinity Prediction. *Bioinformatics* **2020**, *36* (3), 758–764.

(4) Ripphausen, P.; Stumpfe, D.; Bajorath, J. Analysis of Structure-Based Virtual Screening Studies and Characterization of Identified Active Compounds. *Future Medicinal Chemistry* **2012**, *4* (5), 603–613.

(5) Huang, N.; Kalyanaraman, C.; Bernacki, K.; Jacobson, M. P. Molecular Mechanics Methods for Predicting Protein–Ligand Binding. *Physical Chemistry Chemical Physics* **2006**, *8* (44), 5166–5177.

(6) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *Journal of the American Chemical Society* **2015**, *137* (7), 2695–2703.

(7) Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities. *Expert opinion on drug discovery* **2015**, *10* (5), 449–461.

(8) Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling* **2011**, *51* (1), 69–82.

(9) Lyne, P. D.; Lamb, M. L.; Saeh, J. C. Accurate Prediction of the Relative Potencies of Members of a Series of Kinase Inhibitors Using Molecular Docking and MM-GBSA Scoring. *Journal of medicinal chemistry* **2006**, *49* (16), 4805–4808.

(10) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of medicinal chemistry* **2004**, *47* (7), 1739–1749.

(11) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation

of a Genetic Algorithm for Flexible Docking. *Journal of molecular biology* **1997**, *267* (3), 727–748.

(12) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *Journal of computational chemistry* **2009**, *30* (16), 2785–2791.

(13) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *Journal of computational chemistry* **2010**, *31* (2), 455–461.

(14) Li, G.-B.; Yang, L.-L.; Wang, W.-J.; Li, L.-L.; Yang, S.-Y. ID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein–Ligand Interactions. *Journal of Chemical Information and Modeling* **2013**, *53* (3), 592–600.

(15) Liu, J.; Wang, R. Classification of Current Scoring Functions. *Journal of Chemical Information and Modeling* **2015**, *55* (3), 475–482.

(16) Jain, A. N. Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *Journal of medicinal chemistry* **2003**, *46* (4), 499–511.

(17) Elokely, K. M.; Doerksen, R. J. Docking Challenge: Protein Sampling and Molecular Docking Performance. *Journal of Chemical Information and Modeling* **2013**, *53* (8), 1934–1945.

(18) Huang, S.-Y.; Grinter, S. Z.; Zou, X. Scoring Functions and Their Evaluation Methods for Protein–Ligand Docking: Recent Advances and Future Directions. *Physical Chemistry Chemical Physics* **2010**, *12* (40), 12899–12908.

(19) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein–Ligand Docking: Current Status and Future Challenges. *Proteins: Structure, Function, and Bioinformatics* **2006**, *65* (1), 15–26.

(20) Loh, W.-Y. Classification and Regression Tree Methods. *Wiley StatsRef: Statistics Reference Online* **2014**.

(21) Zhang, Q.; Yang, L. T.; Chen, Z.; Li, P. A Survey on Deep Learning for Big Data. *Information Fusion* **2018**, *42*, 146–157.

(22) Ellingson, S. R.; Davis, B.; Allen, J. Machine Learning and Ligand Binding Predictions: A Review of Data, Methods, and Obstacles. *Biochimica et Biophysica Acta (BBA)-*

General Subjects **2020**, *1864* (6), 129545.

(23) Gomes, J.; Ramsundar, B.; Feinberg, E. N.; Pande, V. S. Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity. *arXiv preprint arXiv:1703.10603* **2017**.

(24) Jiménez, J.; Skalic, M.; Martinez-Rosell, G.; De Fabritiis, G. K Deep: Protein–Ligand Absolute Binding Affinity Prediction via 3d-Convolutional Neural Networks. *Journal of Chemical Information and Modeling* **2018**, *58* (2), 287–296.

(25) Li, Y.; Rezaei, M. A.; Li, C.; Li, X. DeepAtom: A Framework for Protein-Ligand Binding Affinity Prediction; IEEE, 2019; pp 303–310.

(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) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein–Ligand Scoring with Convolutional Neural Networks. *Journal of Chemical Information and Modeling* **2017**, *57* (4), 942–957.

(28) 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.

(29) Hassan-Harrirou, H.; Zhang, C.; Lemmin, T. RosENet: Improving Binding Affinity Prediction by Leveraging Molecular Mechanics Energies with an Ensemble of 3D Convolutional Neural Networks. *Journal of Chemical Information and Modeling* **2020**.

(30) Nguyen, D. D.; Wei, G.-W. Agl-Score: Algebraic Graph Learning Score for Protein–Ligand Binding Scoring, Ranking, Docking, and Screening. *Journal of Chemical Information and Modeling* **2019**, *59* (7), 3291–3304.

(31) 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.

(32) Wójcikowski, M.; Kukiełka, M.; Stepniewska-Dziubinska, M. M.; Siedlecki, P. Development of a Protein–Ligand Extended Connectivity (PLEC) Fingerprint and Its Application for Binding Affinity Predictions. *Bioinformatics* **2019**, *35* (8), 1334–1341.

(33) Wang, C.; Zhang, Y. Improving Scoring-docking-screening Powers of Protein–Ligand Scoring Functions Using Random Forest. *Journal of computational chemistry* **2017**, *38*

(3), 169–177.

(34) Salentin, S.; Schreiber, S.; Haupt, V. J.; Adasme, M. F.; Schroeder, M. PLIP: Fully Automated Protein–Ligand Interaction Profiler. *Nucleic acids research* **2015**, *43* (W1), W443–W447.

(35) Da Silva, F.; Desaphy, J.; Rognan, D. IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein–Ligand Interactions. *ChemMedChem* **2018**, *13* (6), 507–510.

(36) Deng, Z.; Chuaqui, C.; Singh, J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein− Ligand Binding Interactions. *Journal of medicinal chemistry* **2004**, *47* (2), 337–344.

(37) Pérez-Nueno, V. I.; Rabal, O.; Borrell, J. I.; Teixidó, J. APIF: A New Interaction Fingerprint Based on Atom Pairs and Its Application to Virtual Screening. *Journal of Chemical Information and Modeling* **2009**, *49* (5), 1245–1260.

(38) Chuaqui, C.; Deng, Z.; Singh, J. Interaction Profiles of Protein Kinase− Inhibitor Complexes and Their Application to Virtual Screening. *Journal of medicinal chemistry* **2005**, *48* (1), 121–133.

(39) Marcou, G.; Rognan, D. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *Journal of Chemical Information and Modeling* **2007**, *47* (1), 195–207.

(40) Chalopin, M.; Tesse, A.; Martínez, M. C.; Rognan, D.; Arnal, J.-F.; Andriantsitohaina, R. Estrogen Receptor Alpha as a Key Target of Red Wine Polyphenols Action on the Endothelium. *PloS one* **2010**, *5* (1), e8554.

(41) de Graaf, C.; Rognan, D. Selective Structure-Based Virtual Screening for Full and Partial Agonists of the B2 Adrenergic Receptor. *Journal of medicinal chemistry* **2008**, *51* (16), 4978–4985.

(42) Chupakhin, V.; Marcou, G.; Baskin, I.; Varnek, A.; Rognan, D. Predicting Ligand Binding Modes from Neural Networks Trained on Protein–Ligand Interaction Fingerprints. *Journal of Chemical Information and Modeling* **2013**, *53* (4), 763–772.

(43) Deng, Z.; Chuaqui, C.; Singh, J. Knowledge-Based Design of Target-Focused Libraries Using Protein− Ligand Interaction Constraints. *Journal of medicinal chemistry* **2006**, *49* (2), 490–500.

(44) Durrant, J. D.; McCammon, J. A. NNScore 2.0: A Neural-Network Receptor–Ligand Scoring Function. *Journal of Chemical Information and Modeling* **2011**, *51* (11), 2897–2903.

(45) Lin, H.; Sassano, M. F.; Roth, B. L.; Shoichet, B. K. A Pharmacological Organization of G Protein–Coupled Receptors. *Nature methods* **2013**, *10* (2), 140.

(46) Biessen, E. A.; Bakkeren, H. F.; Beuting, D. M.; Kuiper, J.; Van Berkel, T. J. Ligand Size Is a Major Determinant of High-Affinity Binding of Fucose-and Galactose-Exposing (Lipo) Proteins by the Hepatic Fucose Receptor. *Biochemical Journal* **1994**, *299* (1), 291–296.

(47) Smith, R. D.; Engdahl, A. L.; Dunbar Jr, J. B.; Carlson, H. A. Biophysical Limits of Protein–Ligand Binding. *Journal of Chemical Information and Modeling* **2012**, *52* (8), 2098–2106.

(48) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME-the Konstanz Information Miner: Version 2.0 and Beyond. *AcM SIGKDD explorations Newsletter* **2009**, *11* (1), 26–31.

(49) Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments. *Journal of computer-aided molecular design* **2013**, *27* (3), 221–234.

(50) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural Fragments: An Universal Language to Encode Reactions, Molecular and Supramolecular Structures. *Journal of computer-aided molecular design* **2005**, *19* (9–10), 693–703.

(51) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of chemical information and computer sciences* **2003**, *43* (6), 1947–1958.

(52) 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. *the Journal of machine Learning research* **2011**, *12*, 2825–2830.

(53) Cao, C.; Liu, F.; Tan, H.; Song, D.; Shu, W.; Li, W.; Zhou, Y.; Bo, X.; Xie, Z. Deep Learning and Its Applications in Biomedicine. *Genomics, proteomics & bioinformatics* **2018**, *16* (1), 17–32.

(54) Jing, Y.; Bian, Y.; Hu, Z.; Wang, L.; Xie, X.-Q. S. Deep Learning for Drug Design: An Artificial Intelligence Paradigm for Drug Discovery in the Big Data Era. *The AAPS journal* **2018**, *20* (3), 58.

(55) Chollet, F. Keras. 2015.

(56) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, High-Quality Test Set for the Validation of Protein− Ligand Docking Performance. *Journal of medicinal chemistry* **2007**, *50* (4), 726–741.

(57) Cang, Z.; Mu, L.; Wei, G.-W. Representability of Algebraic Topology for Biomolecules in Machine Learning Based Scoring and Virtual Screening. *PLoS computational biology* **2018**, *14* (1), e1005929.

(58) Nguyen, D. D.; Xiao, T.; Wang, M.; Wei, G.-W. Rigidity Strengthening: A Mechanism for Protein–Ligand Binding. *Journal of Chemical Information and Modeling* **2017**, *57* (7), 1715–1721.

(59) Wójcikowski, M.; Zielenkiewicz, P.; Siedlecki, P. Open Drug Discovery Toolkit (ODDT): A New Open-Source Player in the Drug Discovery Field. *Journal of cheminformatics* **2015**, *7* (1), 1–6.

(60) Khamis, M. A.; Gomaa, W. Comparative Assessment of Machine-Learning Scoring Functions on PDBbind 2013. *Engineering Applications of Artificial Intelligence* **2015**, *45*, 136–151.

(61) Gaillard, T. Evaluation of AutoDock and AutoDock Vina on the CASF-2013 Benchmark. *Journal of Chemical Information and Modeling* **2018**, *58* (8), 1697–1706.

(62) Ballester, P. J.; Schreyer, A.; Blundell, T. L. Does a More Precise Chemical Description of Protein–Ligand Complexes Lead to More Accurate Prediction of Binding Affinity? *Journal of Chemical Information and Modeling* **2014**, *54* (3), 944–955.

(63) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative Assessment of Scoring Functions: The CASF-2016 Update. *Journal of Chemical Information and Modeling* **2018**, *59* (2), 895–913.

(64) Adam, J.; Pokorná, M.; Sabin, C.; Mitchell, E. P.; Imberty, A.; Wimmerová, M. Engineering of PA-IIL Lectin from Pseudomonas Aeruginosa–Unravelling the Role of the Specificity Loop for Sugar Preference. *BMC Structural Biology* **2007**, *7* (1), 36.

(65) Hanaya, K.; Suetsugu, M.; Saijo, S.; Yamato, I.; Aoki, S. Potent Inhibition of Dinuclear Zinc (II) Peptidase, an Aminopeptidase from Aeromonas Proteolytica, by 8-Quinolinol Derivatives: Inhibitor Design Based on Zn 2+ Fluorophores, Kinetic, and X-Ray Crystallographic Study. *JBIC Journal of Biological Inorganic Chemistry* **2012**, *17* (4), 517–529.

(66) Mayer, M. L. Crystal Structures of the GluR5 and GluR6 Ligand Binding Cores: Molecular Mechanisms Underlying Kainate Receptor Selectivity. *Neuron* **2005**, *45* (4), 539–552.

(67) Madauss, K. P.; Deng, S.-J.; Austin, R. J.; Lambert, M. H.; McLay, I.; Pritchard, J.; Short, S. A.; Stewart, E. L.; Uings, I. J.; Williams, S. P. Progesterone Receptor Ligand Binding Pocket Flexibility: Crystal Structures of the Norethindrone and Mometasone Furoate Complexes. *Journal of medicinal chemistry* **2004**, *47* (13), 3381–3387.

(68) Salonen, L. M.; Holland, M. C.; Kaib, P. S.; Haap, W.; Benz, J.; Mary, J.-L.; Kuster, O.; Schweizer, W. B.; Banner, D. W.; Diederich, F. Molecular Recognition at the Active Site of Factor Xa: Cation–π Interactions, Stacking on Planar Peptide Surfaces, and Replacement of Structural Water. *Chemistry–A European Journal* **2012**, *18* (1), 213–222.

(69) Jenkins, C. L.; Thiyagarajan, N.; Sweeney, R. Y.; Guy, M. P.; Kelemen, B. R.; Acharya, K. R.; Raines, R. T. Binding of Non-natural 3′-nucleotides to Ribonuclease A. *The FEBS Journal* **2005**, *272* (3), 744–755.

Figures

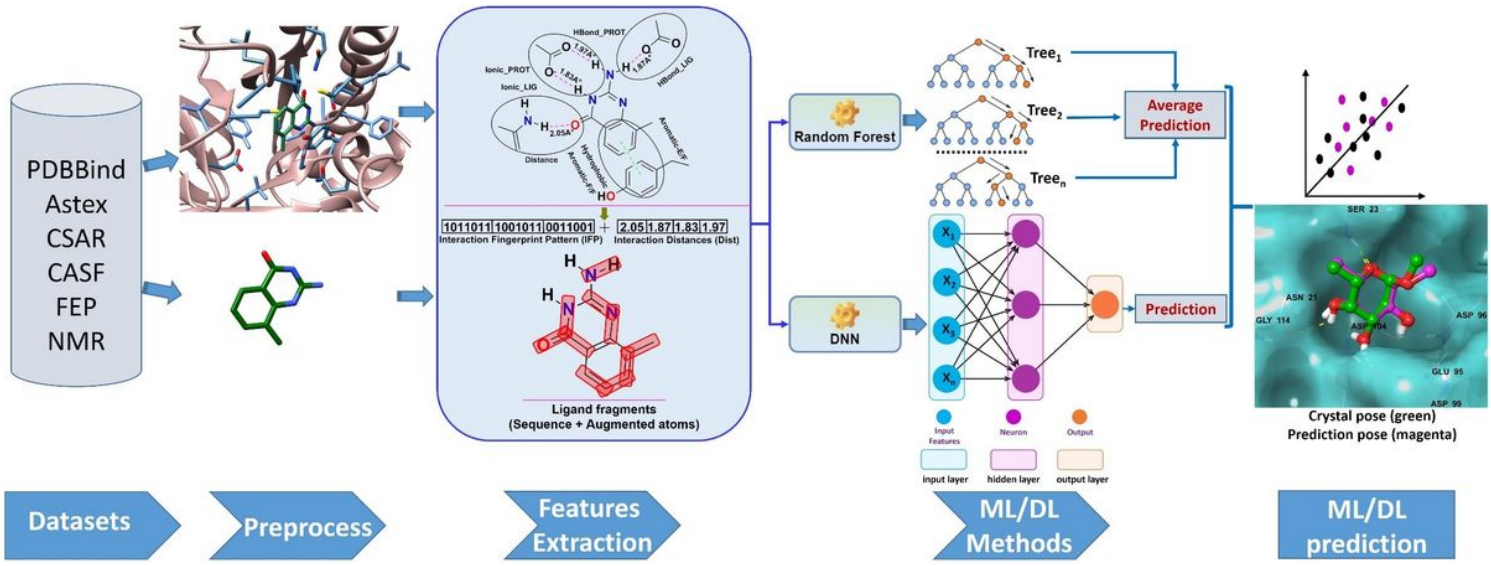

Figure 1

SMPLIP-Score workflow for binding affinity prediction. Publically available protein-ligand binding datasets were used to extract the information encoding the Interaction Fingerprint Pattern (IFP), Interaction Distances (Int-Dist), and ligand fragments to which machine learning or deep learning method may be applied for affinity prediction.

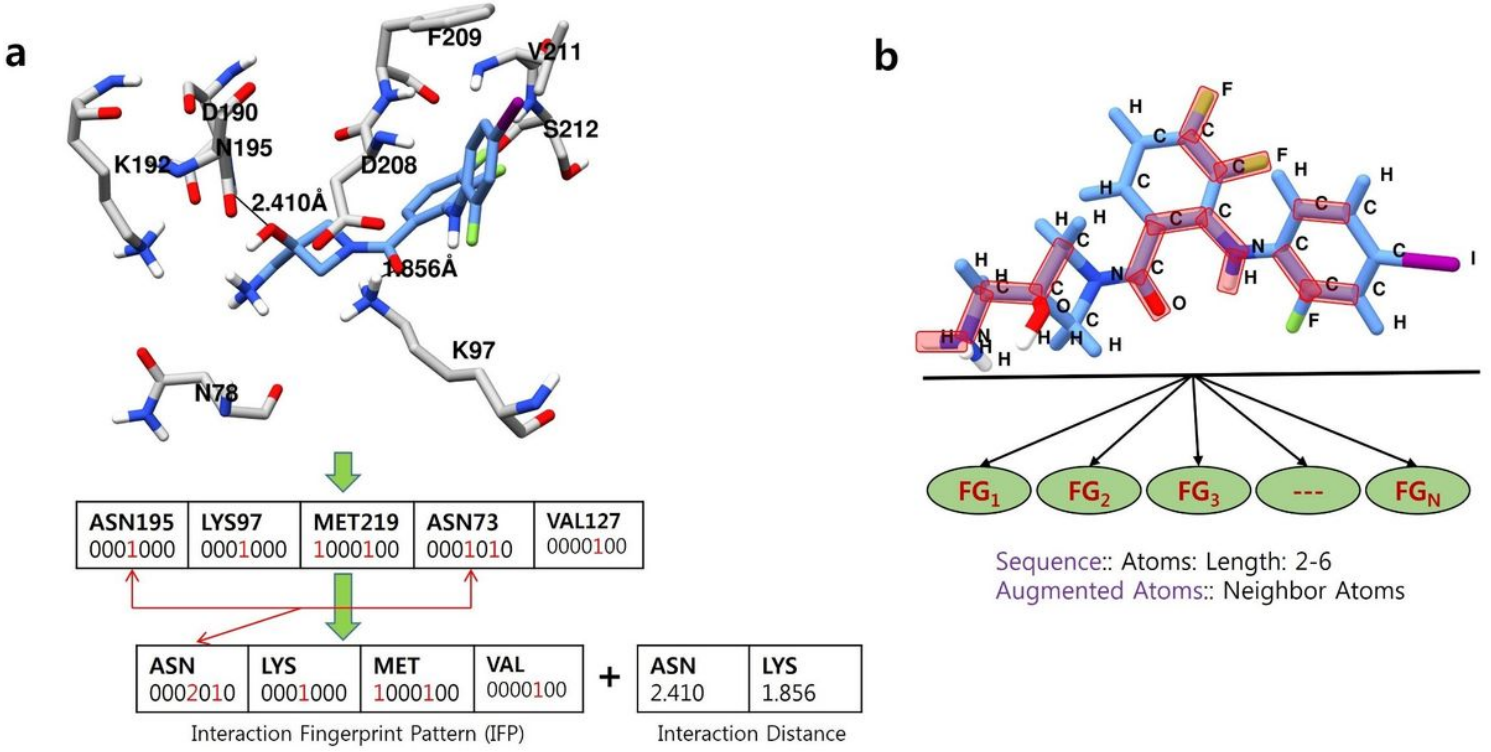

Figure 2

a) Interaction Fingerprint Pattern (IFP) generation. IFP represents the concatenation of interaction values from the same amino acid in a matrix with a fixed size of 140. Similarly, the interaction distances were calculated. b) The ligand fragmentation pattern (FG1, FG2, …, FGN) for the input ligand, which corresponds to atoms sequence and augmented atoms

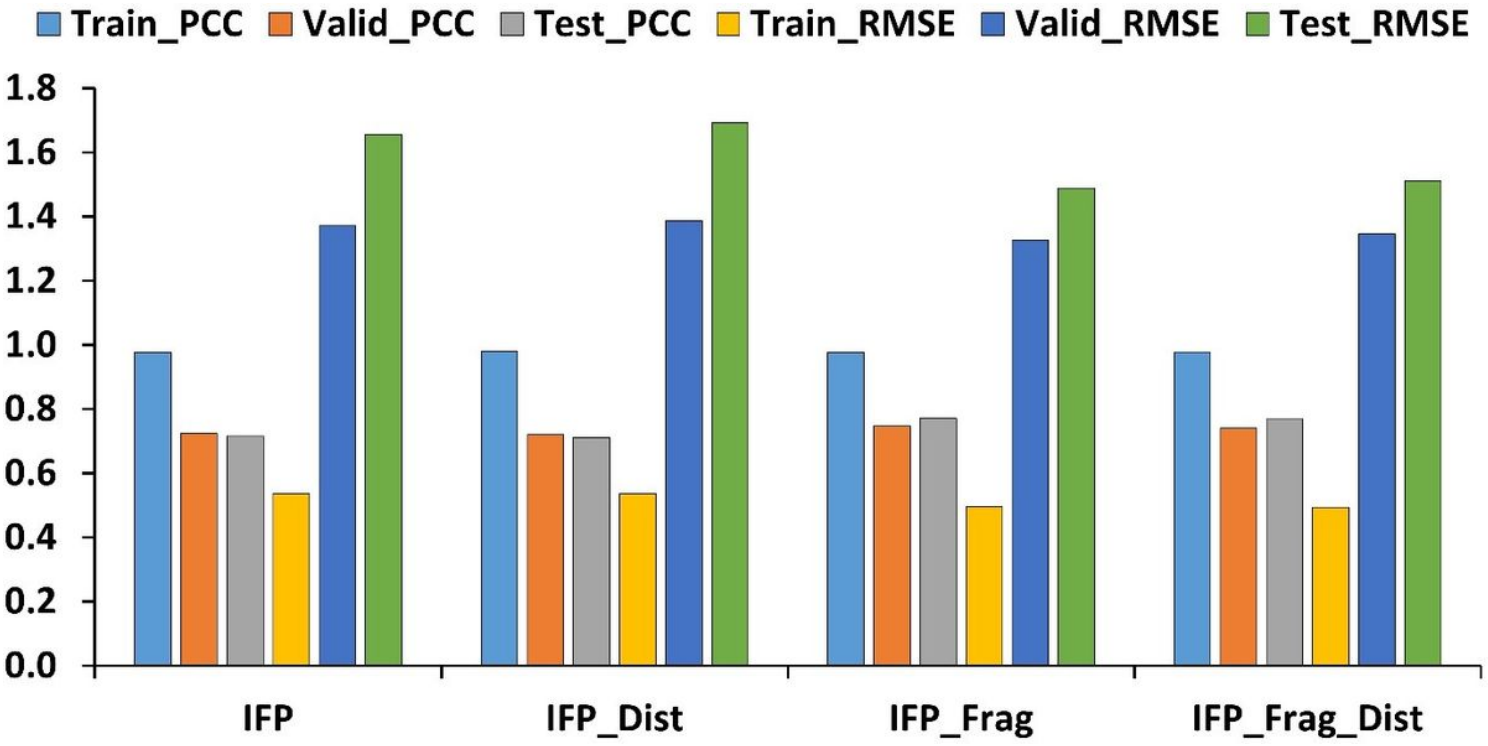

Figure 3

The bar plot for statistical comparison from different feature(s) combination. The PCC (Pearson-Correlation-Coefficient) and RMSE (Root-Mean-Squared-Error) were compared from the train, valid, and test data.

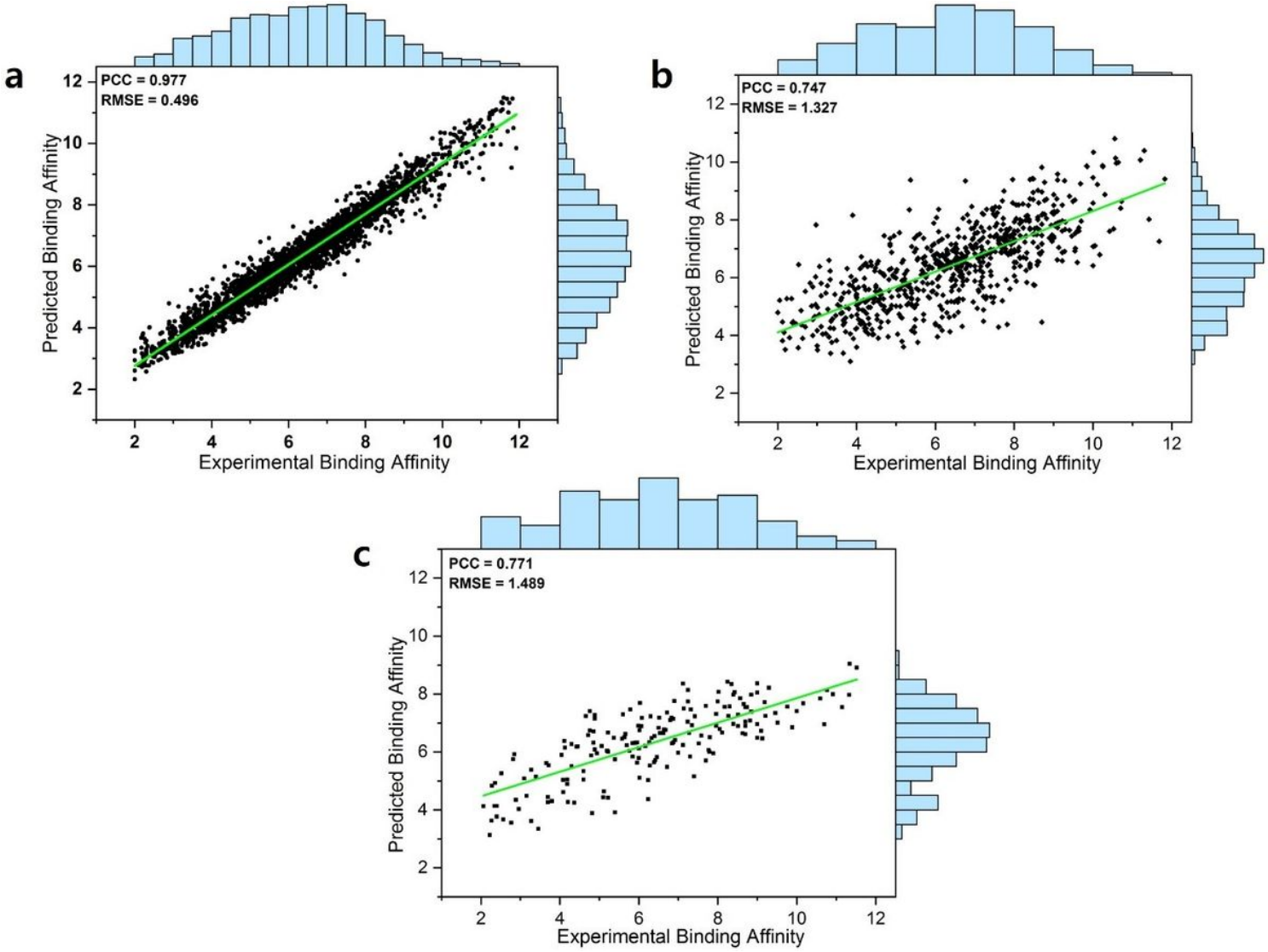

Figure 4

2D-Scatter plot between predicted and experimental of SMPLIP-RF (IFP+Frag) for a) training, b) valid, and c) test data using

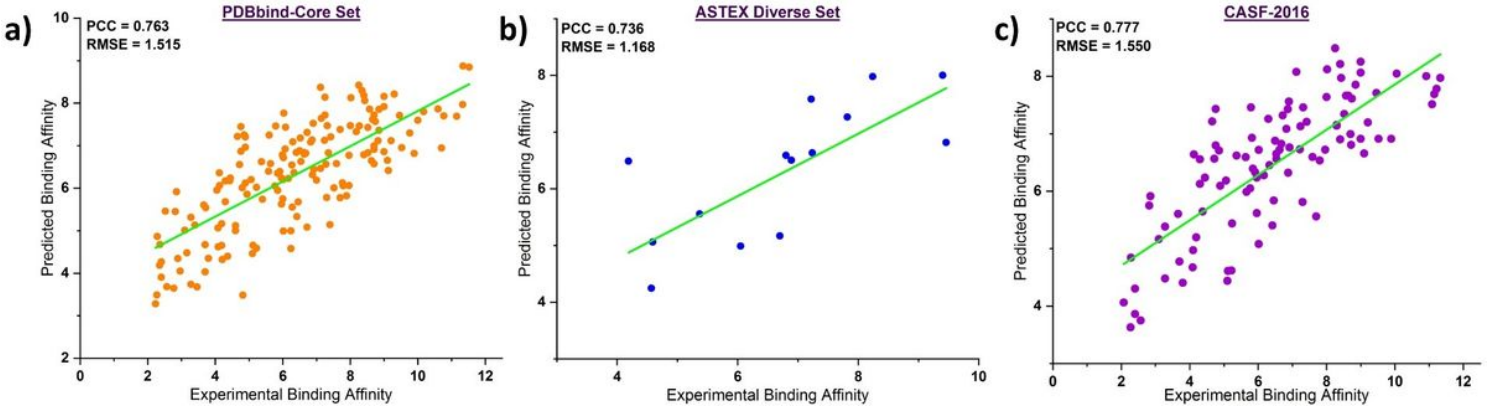

Figure 5

Thee Scatter plots of the predicted binding affinity against the experimental activity from the a) PDBbind Core Set; b) Astex Diverse Set; c) CASF-2016 Set.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryDataV2.docx