Assessing the generalization abilities of machine-learning scoring functions for structure-based virtual screening

Hui Zhu[1,2], Jincai Yang[2,*], Niu Huang[1,2,*]

[1]Tsinghua Institute of Multidisciplinary Biomedical Research, Tsinghua University, Beijing, China 102206, China

[2]National Institute of Biological Sciences, 7 Science Park Road, Zhongguancun Life Science Park, Beijing 102206, China

[*]Correspondence should be addressed to J. Y. (yangjincai@nibs.ac.cn) and N.H. (huangniu@nibs.ac.cn)

Abstract

In structure-based virtual screening (SBVS), it is critical for machine-learning scoring functions (MLSFs) to capture protein-ligand atomic interaction patterns. We generated a cross-target generalization ability benchmark for protein-ligand binding affinity prediction to assess whether MLSFs could capture these interactions. By focusing on the local domains in protein-ligand binding pockets, we developed standardized pocket Pfam-based clustering (Pfam-cluster) approach for the generalization ability benchmark. Subsequently, 11 typical MLSFs were tested using random cross-validation (Random-CV), protein sequence similarity-based cross-validation (Seq-CV), and pocket Pfam-based cross-validation (Pfam-CV) methods. Surprisingly, all of the tested models showed decreased performance as they were evaluated from Random-CV to Seq-CV to Pfam-CV experiments, without showing satisfactory generalization capacity. Interpretable analysis revealed that predictions on novel targets by MLSFs were relying on buried solvent accessible surface area (SASA)-related features in complex structures. By combining buried SASA-related information with ligand-specific patterns that were only shared among structurally similar compounds, higher performance in Random-CV tests was attained for Random forest (RF)-Score. Based on these findings, we strongly advise assessing the generalization ability of MLSFs with the Pfam-cluster approach and being cautious with the

features learned by MLSFs.

Introduction

Structure-based virtual screening (SBVS) has been increasingly applied to identify small molecular binders based on target structures in the field of drug discovery[1–5]. An accurate scoring function for estimating protein-ligand binding affinities is critical for the success of SBVS. Conventional scoring functions used in docking softwares are divided into force field–based scoring functions, empirical scoring functions, and knowledge-based scoring functions[5–10]. To develop a reliable scoring function, intensive efforts have been made in reproducing ligand crystal poses, discriminating between decoys and actives, and learning experimentally determined binding affinities[11–14]. Recently, with the rapid development of machine learning approaches, and more experimental protein-ligand binding complex structures available, machine-learning scoring functions (MLSFs) have been intensively studied for protein-ligand binding affinity prediction[15–17].

Unlike conventional scoring functions, MLSFs are developed to automatically learn protein-ligand structural and interaction features from large protein-ligand complex datasets using different ML algorithms, such as random forest (RF)[18–20], extreme gradient boosting (XGB)[21], convolutional neural network (CNN)[22–25], and graph neural networks (GNN)[26–30]. MLSFs have been demonstrated to achieve remarkable results in various benchmarking studies[12,13,31]. Moreover, MLSFs were also reported to assist in hit identification in practical virtual screening applications[32–34].

However, generalization ability, referring to the ability to make correct and stable decisions on a previously unseen dataset, is the main concern in data-driven algorithms. Machine learning models having huge numbers of parameters, are in principle rich enough to memorize the training data[35]. Besides, machine learning models can achieve encouraging but deceptive scores on an independent and identically distributed dataset by learning dataset bias[36,37]. Models that learn the shortcuts or other unintended features from training datasets would be extremely sensitive to small dataset distribution shifts and would be incapable of making reliable predictions on the out-of-distributed datasets. This instability hinders their practical applications in the unseen, unsought and uncertain world.

In SBVS applications, the main challenge is to discover true binders with novel chemical scaffolds against new drug targets[2,5], where the real testing datasets are out-of-distributed and require MLSFs to have sufficient generalization ability. An early attempt at evaluating generalization ability was the leave-cluster-out cross-validation experiment suggested by Kramer et al[38]. Entries in the PDBbind dataset were classified into different clusters based on protein sequence similarities, and the protein-ligand binding affinity prediction ability of RF-Score was found to decrease from Rp=0.76 on the PDBbind core set to Rp=0.46 on new targets. Thereafter, similar generalization assessments were carried out on many other MLSFs[39], such as SFCscore^{RF}[20], PotentialNet[26], ACNN[37], IGN[27], and SG-CNN[28]. All of these MLSFs demonstrated worse scoring power on new targets than on the PDBbind core set. However, the widely used protein sequence

similarity-based cluster method is not reliable in protein binding-site classification, especially the inability to handle structurally similar proteins with low protein sequence similarity. In addition, those assessment studies were only conducted on limited numbers of MLSFs, and lack of interpretable analysis to explain the poor generalization ability.

Here, we present a systematic study on the generalization ability of MLSFs. We proposed a more reliable data splitting approach, the pocket Pfam-based clustering (Pfam-cluster) approach. Then, 11 representative MLSFs were benchmarked with three, 3-fold cross-validation experiments on the latest version of PDBbind v2020 dataset: random cross-validation (Random-CV), protein sequence similarity-based cross-validation (Seq-CV), and pocket Pfam-based cross-validation (Pfam-CV). We found that all the assessed MLSFs showed declining binding affinity prediction performances from Random-CV to Seq-CV to Pfam-CV experiments, without showing a satisfying generalization ability. Further analysis of individual clusters revealed that all the models dominantly relied on the buried solvent accessible surface area (SASA) calculated in complex structures to make predictions on novel targets. The explicit descriptor of buried SASA was protein-ligand atomic interactions between carbon atoms (C-C interactions) in RF-Score. Moreover, we found that the remarkable binding affinity prediction ability of RF-Score in the Random-CV experiment was achieved by combining specific features that were only shared among structurally similar ligands.

Computational methods

1. PDBbind Dataset (Version 2020)

PDBbind is a comprehensive dataset to curate experimentally determined binding affinity data for the protein-ligand complexes deposited in the Protein Data Bank (PDB)[40]. The general set contains 19,443 protein-ligand complexes and the corresponding experimental binding affinity data, the refine set includes 5,316 high-quality complexes from the general set, and the core set is constructed with 266 representative complexes from the refine set based on protein sequence similarity clustering.

2. Data splitting methods

Random cross-validation (Random-CV). After removing complexes with ligand molecular weight greater than 1,000 Da, the PDBbind general set was randomly and equally divided into three subsets. Any two subsets were split into training and validation sets at an 80/20 ratio and the remaining subset was used for testing. This process was repeated ten times.

Sequence similarity-based cross-validation (Seq-CV). In Seq-CV, the dataset used in the Random-CV experiment was first clustered according to protein sequence similarity and ligand similarity. Pairwise2.align.globallxx module in biopython[41] was used to calculate all-chain-against-all-chain sequence identity and then protein similarity was determined by the result of dividing sequence identity by the longer sequence length. The ligand similarity was measured based on RDKit topological fingerprints[42]. The cutoff of protein similarity was set to 0.5, or 0.4 if the ligand similarity was over 0.9, as suggested previously[24]. The following steps were

the same as the steps in the Random-CV experiment, except that the complexes in the same cluster were grouped into the same subsets.

Pocket Pfam-based cross-validation (Pfam-CV). The protein sequences were systematically clustered into families and domains in Pfam database[43]. Sets of Pfam entries that are evolutionarily related are grouped into clans[44]. Pfam entries were assigned to each protein structure in the Protein Data Bank (PDB)[45]. As one protein may have several different Pfam entries, we only collected the Pfam entry that shared the most residues with the ligand-binding pocket defined by the residues within 7 Å of the crystal ligand, and this Pfam entry was named pocket Pfam. To group the proteins without a pocket Pfam entry into correct clusters, the protein sequences were collected and iteratively aligned to the Pfam sequences of other proteins with the help of the jackhammer search method and HMMER software[46]. Finally, the protein cluster results were manually annotated according to Pfams, clans, alignment results, and pocket structures. The Pfam-cluster returned three hierarchies: Pfam, clan, and cluster. The next processes were the same as the Random-CV experiment.

3. Models

Total of 11 open-source and representative MLSFs were benchmarked. The models and their reported Pearson correlation coefficient (Rp) on the PDBbind core set are summarized in Table 1.

Table 1: Summary of MLSFs

Name	Year	Model	Features	Training dataset	Test set	Rp
LR::V*	-	LR	6 descriptors	Refine 2007 (1,105)	Core 2007 (195)	0.62
LR::VR1*	-	LR	42 descriptors	-	-	-
RF-Score*[47]	2015	RF	42 descriptors	Refine 2007 (1,105)	Core 2007 (195)	0.80
XGB::VR1*	-	XGB	42 descriptors	-	-	-
NNScore*[48]	2011	Ensemble MLP	350 descriptors	-	-	-
Pafnucy[23]	2018	CNN	3D voxels	General 2013 (11,906)	Core 2016 (262)	0.78
OnionNet[49]	2019	CNN	Distance-based contacts and pharmacophore features	General 2016\Refine\ Core	Core 2016 (290)	0.82
SGCNN[28]	2021	GCN	Graph (Distance info)	General 2016 (11,308)	Core 2016 (290)	0.78
IGN[27]	2021	GCN	Graph (Distance + angle info)	General 2016 (8,298)	Core 2016 (262)	0.84
SIGN[30]	2021	GCN	Graph (Distance + angle info)	Refine 2016 (4,057)	Core 2016 (290)	0.80

GraphBAR[29]	2021	GCN	Graph (Distance info)	Refine 2016 (3,319)	Core 2016 (290)	0.75

*The source code of these models is available at https://github.com/hnlab/generalization_benckmark

LR::V and LR::VR1. To make a fair comparison with other MLSFs, two easier linear functions were tested in this study. In LR::V, the weighting factors of six Vina terms (gauss1, gauss2, repulsion, hydrophobic, hydrogen bonding, and the number of rotation bonds) were refitted. Except for six Vina features, 36 RF-Score features (described below) were also considered to construct LR::VR1. Here, six Vina features and 36 RF-Score features were computed from the protein-ligand complex structures with the ODDT toolkit[50].

RF-Score. RF is a representative ensemble ML algorithm, the final decision of which is determined by each decision tree in the RF[18]. RF-Score is an RF algorithm implementation in SBVS to predict protein-ligand binding affinity[47]. Briefly, four protein heavy atoms (C, N, O, S) and six ligand heavy atoms (C, N, O, F, P, S, Cl, Br, I) were selected to generate 36 dense atom pair features, representing the occurrence counts of intermolecular contacts between elemental specific atom pairs within 12 Å. Moreover, six Vina features were also considered in RF-Score. The RF model was constructed using the scikit-learn package[51]. All parameters were the defaults except that the number of estimators was set to 500 (n_estimators=500) and the number of features to consider when looking for the best split was set to 7 (max_features=7). These two parameters were determined by a grid hyperparameter search on the PDBbind core set.

XGB::VR1. XGBoost is another representative ensemble ML algorithm that uses gradient boosting algorithms to speed up computation[21]. The features used in XGB::VR1 were the same as in LR::VR1 and RF-Score. The XGBoost model was constructed with the XGBoost python package[21]. The number of gradient boosted trees was set to 300 (n_estimators=300) and the maximum tree depth for base learners was set to 6 (max_depth=6) based on a grid hyperparameter search on the PDBbind core set.

NNScore. NNScore is an ensemble multilayer perceptron to fit 5 Vina terms (without including the number of rotational bonds) and 345 BINANA descriptors simultaneously[48]. These features were extracted using the ODDT toolkit. In total, 1,000 models were trained every time and the top 20 models on the validation set were obtained. The details of the input and model were described previously[48].

OnionNet. OnionNet is a two-dimensional convolutional neural network (2D-CNN) for protein-ligand binding affinity prediction[49]. Similar to RF-Score features, the features of OnionNet are also based on atom pair contacts between ligands and protein atoms. But these contacts are further grouped into different distance ranges to cover both the local and nonlocal interactions between the ligand and the protein. Altogether, 3,840 features were generated and reshaped to a matrix to mimic image data, followed by two-dimensional CNNs to engineer features.

Pafnucy. Three-dimensional CNN (3D-CNN) has been widely employed in object recognition

because of its spatial representation advantage. Recently, many 3D-CNN models were proposed for protein-ligand binding affinity prediction as 3D-CNN implicitly represents pairwise protein-ligand interactions based on their relative positions in 3D voxel grids[22–24]. Typically, in 3D-CNN representation, each voxel grid contains atom information, such as atom type, partial charge, and atom radius. Then several 3D convolutional layers are followed to extract hierarchical features before fully connected neural networks, which are used to fit binding affinities. Here we chose Pafnucy[23] as the representation of 3D-CNN for benchmarking.

SG-CNN, GraphBAR, IGN, and SIGN. In graph-based neural networks, both proteins and ligands are represented in graphs using one-hot embedding atom feature vectors and an adjacency matrix that contains relationships to the neighboring atoms. Then the atom information is updated according to the neighboring atoms and bond types. After recursive atom feature updating, a read-out function is used to aggregate protein-ligand complex features and the fully connected neural networks are followed to identify and learn protein-ligand binding affinities. Owing to different updating, aggregation, read-out functions, and other model training strategies, a variety of graph neural networks have been proposed. Four graph-based models were considered here and each model has specific characteristics. Briefly, SGCNN utilizes a distance-aware graph attention algorithm to update atom features[28]. The atomic features in GraphBAR are fed into different graph convolution layers based on adjacency matrix type[29]. Instead of utilizing a ligand-based atom sum aggregation approach, IGN takes the sum of the edges as a readout[27]. SIGN considers not only the distance-based protein-ligand pairwise atomic interactions but also angle-related atom information[30]. Moreover, SIGN employs both supervised learning and unsupervised learning strategies.

4. Interpretable tool: Shapley Additive exPlanations

Shapley Additive exPlanations (SHAP) is a useful interpretable approach for understanding model behavior[52]. It decomposes the model output into base values and feature importance values (also called SHAP values). The base value is the mean prediction given by a model and is determined by training dataset distribution. Both the main and feature interaction effects are considered in SHAP values, and the positive or negative SHAP values represent the positive or negative contributions to model output.

5. Evaluation metrics

For all cross-validation schemes, Pearson correlation coefficient (Rp), mean absolute error (MAE), and coefficient of determination (R^2) were calculated between experimental binding affinities and predicted ones. As the dataset was split into three subsets, the overall performance was determined by the averaged results of the three subsets. In terms of the cluster performance, three test sets on one repetition were merged and individual cluster scores were calculated.

Results

1. Protein sequence similarity-based clustering

MLSF performances on the PDBbind core set have been considered overoptimistic. To evaluate model behavior objectively, many studies conducted Seq-CV experiments, where protein targets were clustered with certain protein sequence similarity cutoffs (commonly 0.5). However, we found that this protocol had certain drawbacks in protein clustering. Especially, proteins that shared low sequence similarities but contained conserved ligand-binding pockets were classified into different clusters. Therefore, some similar complexes were used for training, rendering the cross-target tests unreliable. Typical examples of misclassification by protein sequence similarity-based clustering (Seq-cluster) approach are illustrated in Figure 1.

It is well known that the kinase ATP-binding pocket is highly conserved among the kinase family[53]. However, the calculated pairwise sequence similarity between casein kinase II subunit alpha[54] (CK2α) and epidermal growth factor receptor[55] (EGFR) was only 0.36 (Figure 1A). Similarly, two ribonucleases[56,57] with similar ligand-binding pockets only shared 0.39 protein sequence similarity (Figure 1B). As a result, these similar kinases and ribonucleases were divided into different clusters with a protein sequence similarity cutoff of 0.5.

The next two misclassifications were the results of sequence lengths. The protein sequence similarity was determined by the result of dividing sequence identity by the longer sequence length. The computed protein sequence similarity would be misleading if one protein's sequence length was significantly longer than the other. For example, 63/99 residues in HIV-1 protease chain A[58] were mapped to renin[59]. Despite this, because renin was much longer than HIV-1 protease chain A, the protein sequence similarity between them was only 0.18 (Figure 1C). The conserved ligand-binding pockets were also shared by the human GPCR Angiotensin II type 2 receptor[60] and the β1-adrenoceptor[61] (Figure 1D), although the protein sequence similarity was 0.35 due to the inappropriate similarity calculating approach.

As the performances of MLSFs are significantly related to similar samples in the training dataset[62–64], generalization evaluation results with the Seq-cluster approach would be misleading.

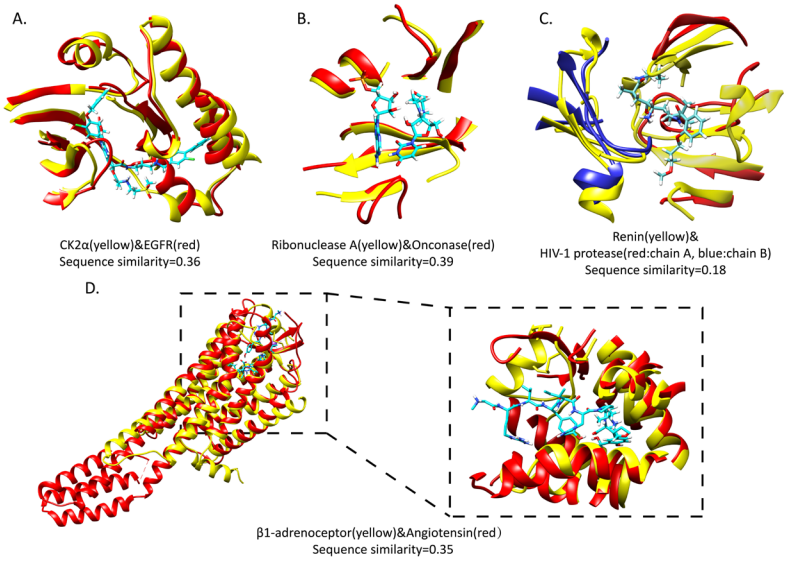

Figure 1. Typical examples of misclassification by Seq-cluster approach. A. structure alignment between casein kinase II alpha subunit (CK2a, yellow, PDB code: 5mo8) and epidermal growth factor receptor (EGFR, red, PDB code: 1xkk). B. structure alignment between Ribonuclease A (yellow, PDB code: 2g8r) and Onconase (red, PDB code: 2gmk). C. structure alignment of Renin (yellow, PDB code: 2v12) and HIV-1 protease (red: chain A, blue: chain B, PDB code: 1ec1). D. structure alignments between two GPCRs: β1-adrenoceptor (yellow, PDB code: 6h7m) and Angiotensin II (red, PDB code: 5xjm). All the images are rendered in CHIMERA[65].

2. Pocket Pfam-based clustering

Previous studies have suggested new approaches to overcome the limitations of the Seq-cluster approach, including using fold classification and the hidden Markov model to find sequences of distantly related samples[66,67]. Here, we clustered protein targets according to the Pafm entry as described in the methods section. Instead of only assembling proteins with high protein sequence similarity, pocket Pfam-based cluster (Pfam-cluster) approach focuses on the domains in binding pockets.

From Pfam-cluster results (Table S1), the PDBbind v2020 general set consisted of 939 clusters, among which 39 clusters contained more than 100 members. Figure 2A shows the distributions of clusters with more than 300 members. Protein kinase (Pkinase) superfamily, the biggest cluster, accounted for 15%. Peptidase also contributed a lot to PDBbind, with peptidase_AA, peptidase_PA, and peptidase_MA making up 5%, 5%, and 2%, respectively. However, based on the Seq-cluster results, the general set was classified into 2,359 clusters, 86.4% of which contained less than ten members.

The great advantage of the Pfam-cluster approach was that the small clusters in the Seq-cluster results were able to be grouped due to belonging to the same protein superfamily and containing similar pocket domains. For example, the Pfam-cluster approach classified 3,017

proteins from the kinase superfamily as Pkinase, whereas the Seq-cluster divided Pkinases into 83 different clusters, and divided 987 peptidase AA complexes into 7 clusters. In the Pkinase cluster, pairwise protein sequence similarities ranged from 0 to 1, with the majority around 0.4 (Figure 2B and Figure S1A). Proteins in Peptidase_AA cluster mainly contained RVP (retroviral aspartyl protease, 475/987) and Asp (Aspartate protease, 464/987) Pfam domains with pairwise similarities ranging from 0.16 to 1. However, the protein sequence similarities between these RVP and Asp Pfam domains were around 0.2 (Figure 2C and Figure S1B). The protein sequence similarity distributions in Pkinase and Peptidase_AA indicated that Seq-cluster approach was incapable of categorizing the structurally similar proteins correctly when the protein sequence similarity was lower than 0.5. Based on the Pfam-cluster results, we conducted a cross-target generalization ability benchmark for 11 MLSFs.

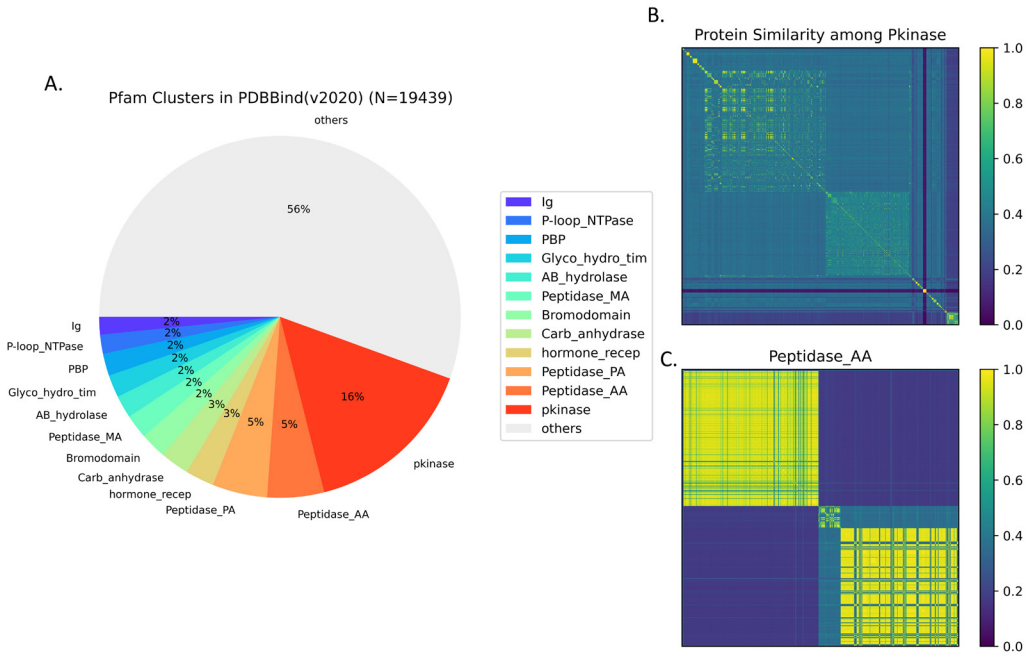

Figure 2. Pfam-cluster results. A. Pfam-cluster results of PDBbind dataset. Only the clusters with more than 300 complexes were shown and other clusters were merged into "others" in the figure. B and C. Pairwise protein sequence similarity distribution among Pkinase and Peptidase_AA cluster.

3. Evaluation of the generalization ability of MLSFs

Following other studies , we tested the performances of every model on the PDBbind v2020 core set using the PDBbind general set as the training dataset. The results are summarized in Figure 3 and Table S2. All MLSFs except two linear functions achieved excellent performances on the core set, with Rp ranging from 0.74 to 0.81, MAE ranging from 1.07 to 1.21, and R^2 ranging from 0.51 to 0.61. Two linear regression functions yielded Rp=0.61, MAE=1.54, R^2=0.16 for LR::V, and Rp=0.67, MAE=1.44, R^2=0.28 for LR::VR1. It was not surprising that more complicated MLSFs outperformed simple linear functions on the core set. However, whether these MLSFs capture the true protein-ligand interaction patterns remains a crucial concern. A robust and

reliable scoring function is supposed to be predictive in novel targets. To explore the cross-target generalization ability of MLSFs, we conducted three different cross-validation experiments: Random-CV, Seq-CV, and Pfam-CV experiments. The benchmark results were discussed from two aspects: the overall generalization ability on the PDBbind general set, and the individual cluster performances. Here, the average Rp, MAE, and R^2 on 3 testing folds in 3-fold cross validation experiment were used as the overall generalization ability evaluation metrics. As different clusters have different ranges of binding affinity, only the Rp metric was comparable on individual cluster performance analysis. The cluster Rp on three experiments was calculated on 3 merged testing folds, thus no training data was included.

3.1 Overall generalization ability comparison

Benchmark results on the PDBbind general set for MLSFs are shown in Figure 3 and Figure S2, where the models were sorted in order of the median Rp and MAE on ten repetitions in Random-CV experiments (Tables S3, S4). First, comparing the results of three cross-validation tests on the same model, the Rp of complicated MLSFs declined from core set to the Random-CV experiments to Seq-CV to Pfam-CV. In contrast, the linear regression model with six Vina terms behaved (LR::V) stably, with median Rp changing from 0.47 (Random-CV) to 0.45 (Seq-CV) to 0.46 (Pfam-CV). The reduced performance suggested that the scoring power of MLSFs would be significantly affected by similar structures, while simpler linear functions were stable and independent to the sample similarity, which was consistent with previous findings[62–64]. Furthermore, the reduced performance from Seq-CV to Pfam-CV highlighted the necessity of a more precise and rigorous protein clustering method, which would more reliably represent scoring functions' generalization ability on novel targets.

In summary, SIGN, a GCN model embedding both protein-ligand atom pair distance and angle information, obtained the best performance (Rp=0.72 and MAE=1.02) in Random-CV experiments, but the performance on novel targets was discouraging (Rp=0.53 and MAE=1.31). IGN, which demonstrated better than average performance in Random-CV tests (Rp=0.68 and MAE=1.06), showed a worse Rp score than LR::VR1 on novel targets (IGN on Pfam-CV: Rp=0.47 and MAE=1.35; LR::VR1 on Pfam-CV: Rp=0.49 and MAE=1.62). Moreover, OnionNet had the top two ranked scores on Random-CV tests (Rp=0.71 and MAE=1.03) and performed similarly with LR::VR1 on Pfam-CV tests. Interestingly, RF-Score, a random forest model featuring 36 distance-based descriptors and six Vina terms, behaved better. The Rp of RF-Score decreased from 0.70 on Random-CV to 0.58 in Pfam-CV experiments and the MAE changed from 1.06 to 1.21. These results indicated that models containing too many parameters were more likely to overfit training data. The complex overfitted models performed worse than the simpler model in the out-of-distribution dataset.

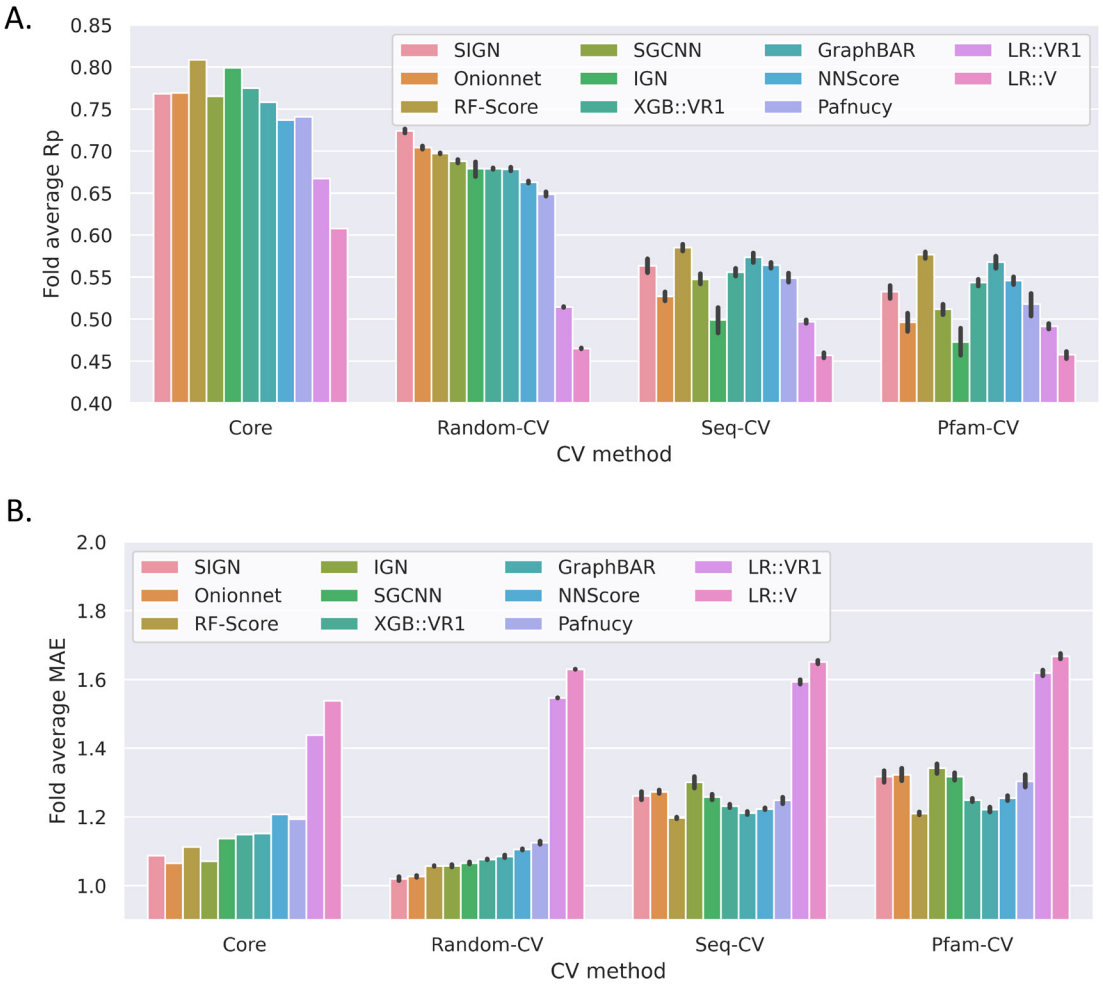

Figure 3. Cross-validation tests results of 11 MLSFs. A. The Rp on the PDBbind core set (only repeated once) and the average Rp performances on three folds (10 repetitions). MLSFs are sorted by the median Rp on Random-CV experiments. B. The same as panel A but the metric is MAE.

3.2 Individual cluster performance comparison

Here we selected LR::V, RF-Score, Pafnucy, and SIGN as the representative LR, traditional ML-based, 3D-CNN-based, and GCN-based scoring functions to investigate the individual cluster performances. Figure 4 shows the maximum, minimum, and averaged cluster Rp in ten repetitions of these models; the results of other models are included in Figure S3. All the models showed fluctuating performances, with smaller fluctuations in LR models and RF-Score compared to Pafnucy and SIGN. These observations represent model uncertainty. As the RF is an ensemble model, the output of which is related to the average prediction of every tree, its performance was more robust. Moreover, compared to the fluctuations between Pfam-CV and Random-CV experiments, the cluster Rp in Pfam-CV tests changed more significantly because the training and testing datasets were distributed differently in Pfam-CV experiments.

In addition, each model's cluster performance varied from target to target, with Rp ranging from -0.40 to 0.80. The varied cross-target cluster performances were reported previously[20,38].

Moreover, different models shared many well-predicted clusters. The top five predicted clusters (Alk_phosphatase, Cyclophil-like, POLO_box, Calcineurin, and S5) in the RF-Score Pfam-CV test are highlighted as black dots in Figure 4. Except for Pafnucy, which had an unstable prediction on the Calcineurin cluster, the other models performed well. In particular, LR::V, RF-Score, Pafnucy, and SIGN achieved more than Rp scores of 0.8 on the Cyclophil-like cluster in Pfam-CV tests. Even the simplest LR::V scoring function also demonstrated outstanding binding affinity prediction ability on these targets (Calcineurin: Rp=0.87, POLO_box: Rp=0.82, Cyclophil-like: Rp=0.80 and Alk_phosphatase: Rp=0.72). We hypothesized that these clusters contains the simplest samples, whose binding affinities could be predicted using a few basic properties, and that all of the MLSFs had captured these features from other proteins in the training dataset.

To understand how similar protein-ligand complexes influence model behavior, we also compared the individual cluster performances between Random-CV and Pfam-CV tests (Figure 5 and Table S5). Consistent with the above observation, performances of the easier linear functions (LR::V and LR::VR1) were independent of whether similar structures existed in the training dataset or not; the individual cluster performance of Pfam-CV tests was almost the same as that in Random-CV tests (Figure 5A). Surprisingly, no significant improvement was observed in other complicated MLSFs when adding similar complexes to the training dataset. For example, Figure 5B showed the comparison results of RF-Score in the first repetition, where the cluster performance changes were small between two cross-validation tests, and the Rp was 0.86. In detail, RF-Score only achieved significantly better scores on clusters Avidin, Sialidase, una_570, and APC in Random-CV tests. Overall, the Rp between cluster performance on two cross-validation tests for all the models and all the repetitions ranged from 0.45 to 1.00 (Figure 5C). The comparable cluster performances between Random-CV tests and Pfam-CV tests indicated that the features learned by Random-CV models and Pfam-CV models were similar. However, on some targets, similar complexes helped Random-CV models behave better.

Interpretable analysis was a useful tool for us to identify the basic features captured by MLSFs and determine the target-specific features captured from similar targets. As RF-Score showed relatively stable cluster performances and was easier to interpret, we took it as an example to address these questions.

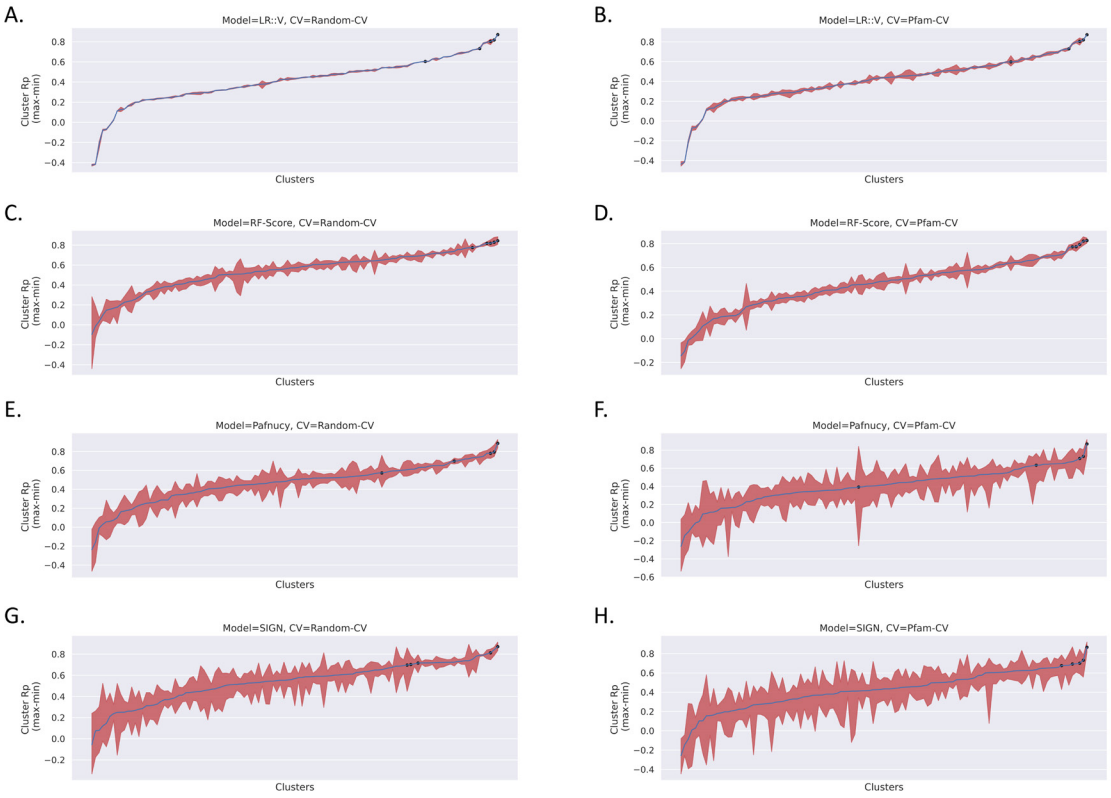

Figure 4. Individual cluster performances (Rp) in Random-CV and Pfam-CV experiments. The maximum, mean and the minimum cluster performance (Rp) on 10 repetitions were plotted. The annotated 5 black dots were the top 5 clusters in RF-Score Pfam-CV experiments: Alk_phosphatase, Cyclophil-like, POLO_box, Calcineurin, and S5

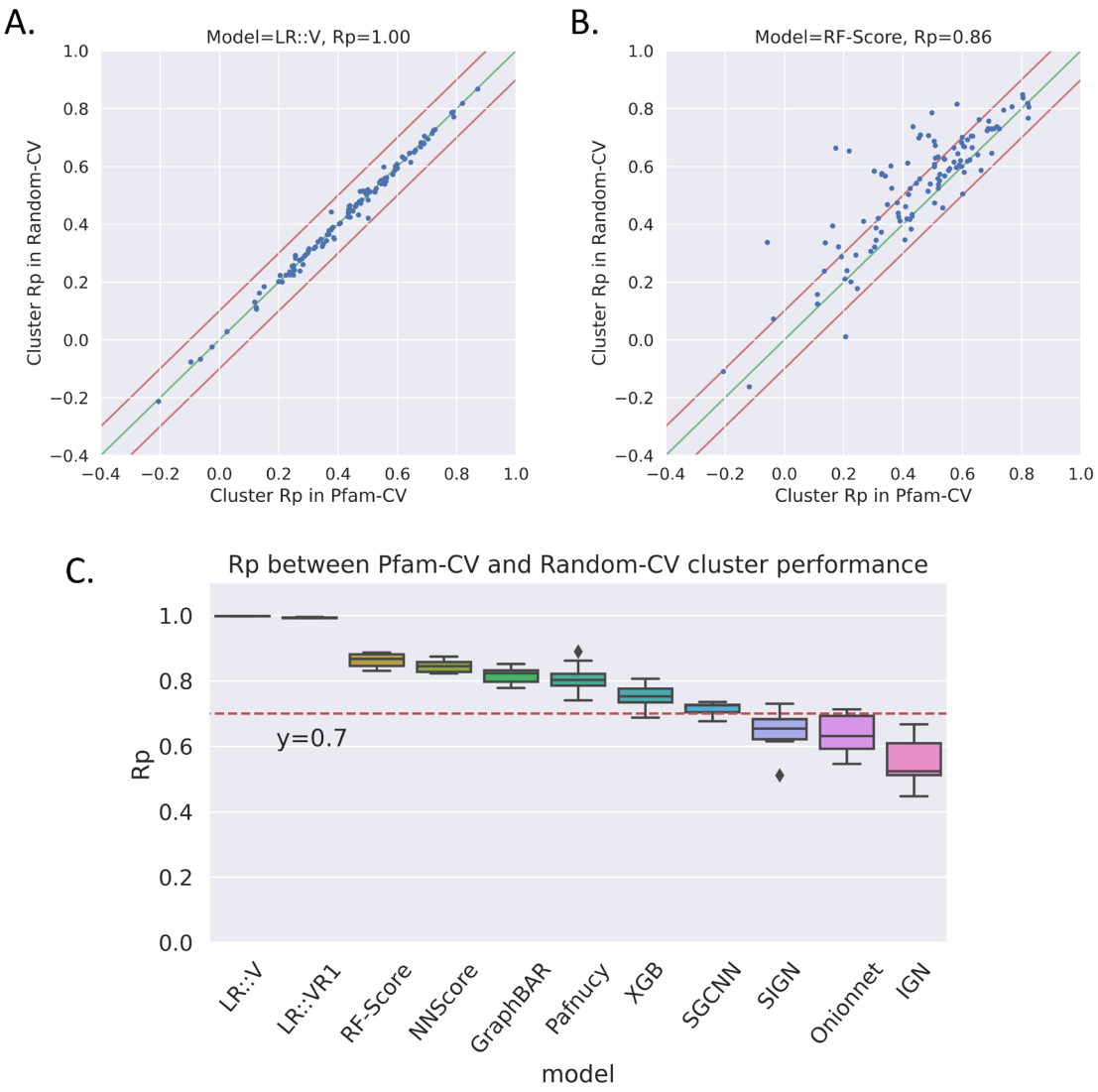

Figure 5. Cluster performance (Rp) comparison between Pfam-CV and Random-CV experiments. A. Individual cluster performance (Rp) achieved by LR::V on the first Random-CV and Pfam-CV experiments. The red lines are Y=X+0.1 and Y=X-0.1. B. The same as A except the model is RF-Score. C. Rp between Pfam-CV and Random-CV cluster performances on ten repetitions.

4. Interpretable analysis on RF-Score

We first conducted SHAP analysis on RF-Score Pfam-CV and Random-CV models on the training dataset to explore the basic features and then compared clusters with different performances between Pfam-CV and Random-CV tests to identify the target-specific features.

4.1 Basic features: from C-C interactions to buried SASA

The beeswarm plots in Figures 6A and Figure S4 are an overview of the seven most important features of RF-Score on Pfam-CV and Random-CV tests, where the SHAP values of the seven features of every sample from the training dataset are plotted. Two models, Pfam-CV and

Random-CV RF-Score, shared many important features: protein carbon-ligand carbon interaction (C-C), protein nitrogen-ligand carbon interaction (N-C), and four Vina terms (gauss1, gauss2, hydrophobic, and the number of rotation bonds). It was acceptable that Pfam-CV and Random-CV RF-Score learned similar features because their training datasets were both sampled from the PDBbind dataset. Based on these findings, we used the RF-Score interpretable results in Pfam-CV tests as examples to investigate feature importance in detail.

Larger values of feature N-C interaction, C-C interaction, gauss1, gauss2, and hydrophobic terms represented the greater positive contributions to the final scores (Figure 6B, Figure S5, S6). In terms of ligand rotation bond number, the model gives higher scores to complexes containing ligands with increasing numbers of rotation bonds ranging from 0 to 10. However, the model penalizes the complexes with rotation bonds more than ten (Figure 6C). The mean absolute SHAP values of other features are represented in Figure 6D (Pfam-CV) and Figure S4 (Random-CV). It was clear that the pair interactions between protein atoms and ligand phosphorous atom (feature X-P), bromine atom (feature X-Br), and iodine atom (feature X-I) did not affect the model's output. It is likely that these atom pairs account for an extremely low percentage in crystal structures and thus models were insensitive to these features. Our studies also showed that the feature interaction effects (off-diagonal values in Figure 6D) had a minor impact on the final scores. Thus, the sum of single feature and feature interaction SHAP values was used to analyze feature importance, unless specifically annotated.

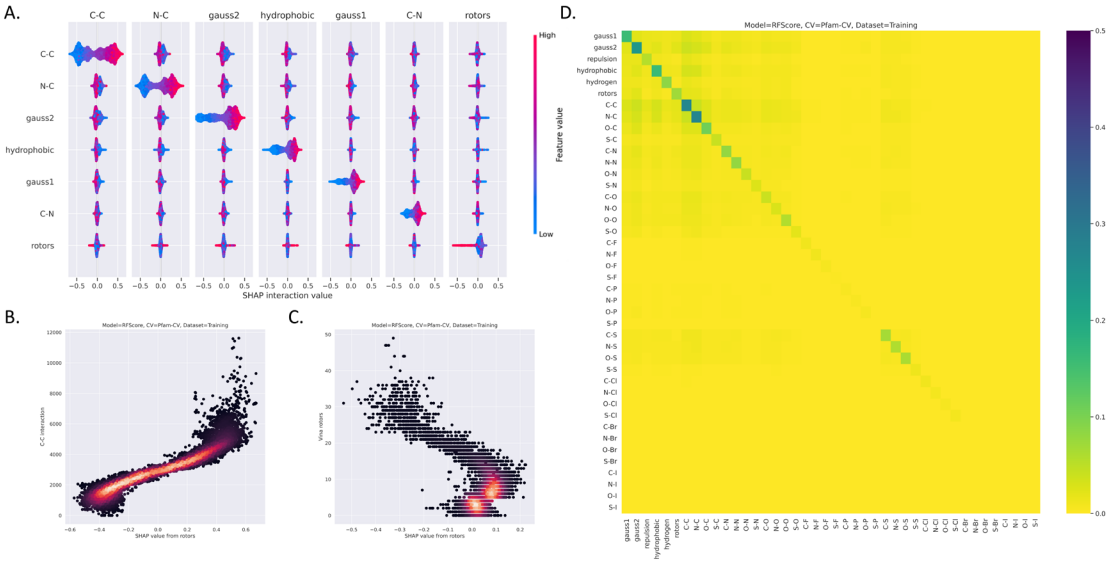

Figure 6. SHAP analysis of RF-Score on Pfam-CV training set. A. The most important 7 features learned by RF-Score in Pfam-CV test. Every dot represents a sample and the color means the relative feature value (red high, blue low). The diagonal SHAP values are single feature importance and the off-diagonal SHAP values are the feature interaction effect. B. Correlations between C-C interaction feature value and C-C interaction SHAP values on Pfam-CV training set. C. Correlations between ligand rotation bonds number and ligand rotation bonds number SHAP values on Pfam-CV training set. D. The mean absolute SHAP value of 42 features and feature interaction effects in Pfam-CV experiment.

The SHAP analysis results highlighted the importance of protein carbon-ligand carbon (C-C) interaction (the protein carbon-ligand carbon pair counts within 12 Å) on both Pfam-CV and Random-CV experiments on RF-Score. In terms of physical properties, the feature of C-C interaction corresponds to the buried SASA of the protein-ligand complex. The Rp between feature C-C values and buried SASA was 0.84 in the PDBbind dataset (Figure 7A). Thus, we speculated that buried SASA was the basic feature learned by RF-Score and other models. To test this hypothesis, we compared the cluster Rp scores calculated with only buried SASA and the cluster Rp predicted by models. Interestingly, all models showed high positive correlations between buried SASA scores and model predicted scores in three cross-validation experiments and the average Rp for ten repetitions ranged between 0.40 and 0.91 (Figure 7B). Compared with the cluster performances on Random-CV experiments, the buried SASA scores–model scores correlations were higher in Seq-CV and Pfam-CV experiments.

Moreover, linear models showed higher correlations to buried SASA scores compared with other models. Here, we took RF-Score in the first Pfam-CV experiment as an example to analyze the relationship between model cluster performance and buried SASA cluster performance (Figure 7C) and similar scatterplots of other models and other cross-validation experiments are shown in Figure S7. The Rp between buried SASA cluster scores and Pfam-CV RF-Score cluster scores was 0.66. It was clear that RF-Score could only make clear predictions on clusters with high binding affinity-buried SASA correlations. For example, the experimental binding affinity increased with larger buried SASA in the SH2-like cluster and the model gave higher scores to complexes with larger protein-ligand interfaces (Figure 7D). However, in terms of clusters without a binding affinity-buried SASA correlation like Peripal_BP, models were incapable of making predictions (Figure 7E). RF-Score assigned higher labels to complexes with high buried SASA, however, the experimental binding affinity did not increase as buried SASA increased, and the model performed poorly.

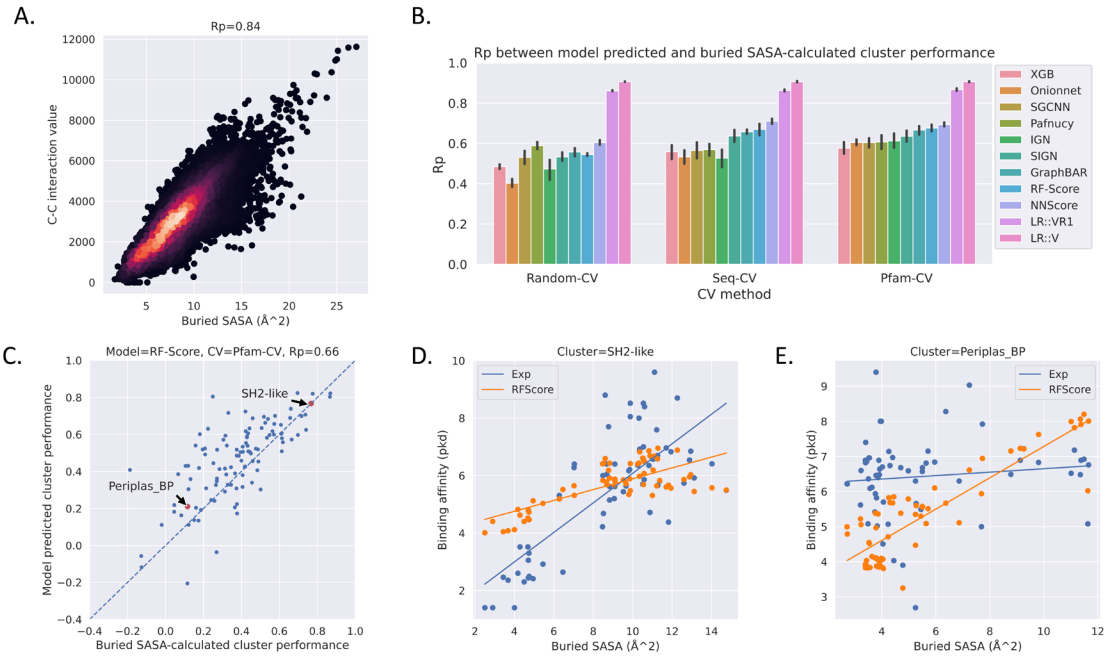

Figure 7. Correlations between models predicted cluster performance and buried SASA calculated cluster performance. A. Correlation between complexes buried SASA and C-C interaction feature on PDBbind dataset. B. Pearson correlation coefficients of 11 MLSFs in 3 CV experiments on 10 repetitions. C. Correlation between RF-Score predicted individual cluster performances on Pfam-CV experiment and buried SASA-calculated cluster performances. D and E. Specific RF-Score predicted binding affinities, experimental binding affinities, and buried SASA of complexes in SH2-like and Periplas_BP cluster.

4.2 Binding affinity-buried SASA correlation in the PDBbind dataset

It is not surprising that RF-Score and other models assign higher scores to complexes with larger protein-ligand binding interfaces. It is common in medicinal chemistry studies that improved ligand binding affinity is always coupled with increased molecular size during the ligand optimization process. The binding affinity-buried SASA correlation coefficient is 0.26 in the PDBbind dataset (Figure S8) and 0.39 if only considering complexes with ligands of less than 1,000 Da (Figure 8A); the Rp of most clusters ranges between 0.2 and 0.6 (Figure 8B). However, a significant binding affinity-buried SASA correlation was identified in some clusters, such as SH2-like (Rp = 0.77, Figure 8C), and Cyclophil-like (Rp =0.87, Figure 8D), and Calcineurin (Rp = 0.87, Figure 8E). The fragment-to-lead strategy was applied when designing nonpeptidic inhibitors in the SH2-like cluster[68–70]. The ligands in Calcineurin are large toxins or fragments identified by fragment-based screening[71–75]. The Cyclophil-like cluster includes peptides, macrocycle, and simplified macrocycles[76–79]. As a result, all the models, even LR::V, demonstrate excellent binding affinity prediction performances in these clusters.

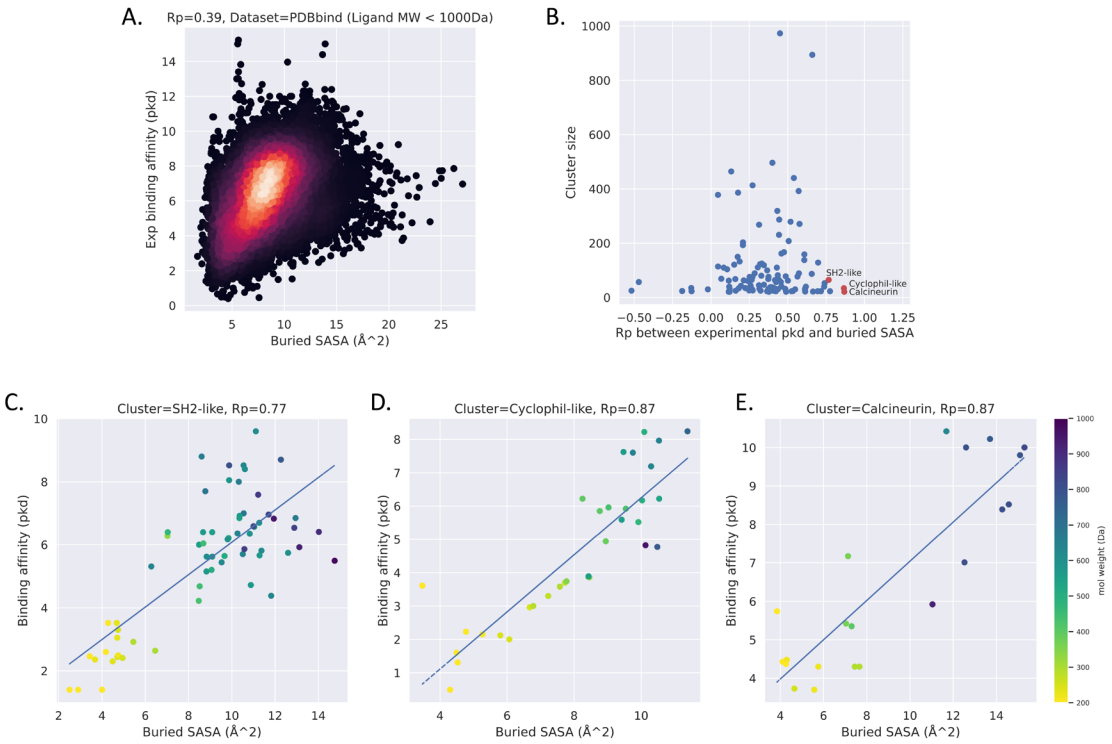

Figure 8. Experimental binding affinity-buried SASA correlation. A. Overall Rp between buried SASA and experimental binding affinity in PDBBind dataset (ligand molecular weight < 1000Da). B. Experimental binding affinity-buried

SASA correlations of individual clusters in PDBbind dataset. Only the clusters with more than 20 members were plotted. For overall clarity, the biggest Pkinase cluster was not plotted (cluster size=2973, Rp between binding affinity and buried SASA =0.43). C, D and E. Details of binding affinities, buried SASA, and ligand molecular weight of complexes in SH2-like, Cyclophil-like, and Calcineurin cluster.

4.3 Target-specific features

From the above analysis, we knew that the individual cluster performances were mainly achieved by learning buried SASA-related information. Nevertheless, we also observed that the cluster performances in Random-CV experiments were slightly better than that in Pfam-CV experiments, which indicated that similar complexes made a difference in protein-ligand binding affinity prediction. So, we next focused on three clusters (Avidin, Sialidase, and Peptidase AA), of which the cluster performances were steadily improved in RF-Score Random-CV tests and the complex numbers were sufficient (more than 50), to interpret the target-specific features (Figure S9).

Most of the ligands in the Avidin cluster were biotins and epi-biotins, which formed the strongest known non-covalent interactions with avidins and showed bioactivities at a range between 4 and 14 pKd in PDBbind dataset. However, RF-Score predicted relatively weaker binding affinities on biotin-avidin complexes on Pfam-CV test because of small values of ligand size-related features (biotin MW= 240 Da and buried SASA = 5.56 $Å^2$, Figure 9A). For example, the penalties was 0.29 pKd for low C-C interaction values and 0.19 pKd for low N-C interaction values on one testing avidin-biotin complex (PDB code: 1swp) and the final predicted label was 6.63 pKd (Figure 9B). As the similar biotin-avidin complexes were presented in the training dataset on Random-CV test, RF-Score recognized the specific features of these complexes and assigned them with reasonable labels. Comparing the SHAP analysis results of 1swp between two tests, the features of C-S, O-S, and N-S interactions made much more positive contributions to the final predictions in the Random-CV model, indicating that the local environment of the S atom in the biotin was highlighted by the Random-CV model (Figure 9C). Interestingly, when decomposing C-C interactions importance into feature interaction effects, we found the low C-C interactions value also made huge negative contributions to the final scores in the Random-CV model. However, the positive interaction effects between C-C interactions and other features, such as C-S, O-S, and gauss2, offset these negative effects (Figure 9D).

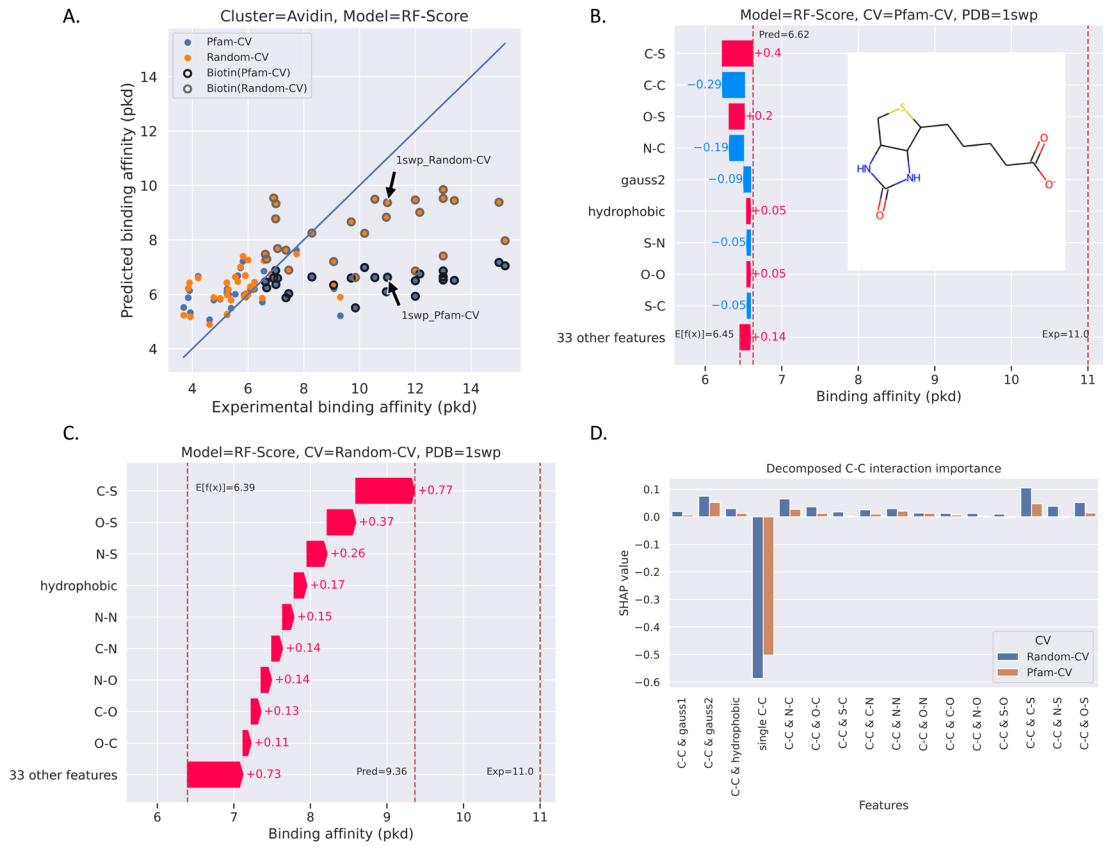

Figure 9. RF-Score important features in Avidin. A. Experimental binding affinities and RF-Score predicted binding affinities in Avidin cluster. Avidin-biotin complexes were circled. B and C. SHAP analysis of 1swp on RF-Score Pfam-CV and Random-CV experiments. Exp represents experimentally determined binding affinity, Pred means RF-Score predicted label, and E[f(x)] is the base value of SHAP analysis. D. Decomposed C-C interactions SHAP value into single C-C interaction importance and feature interaction effect between C-C interactions and other features on 1swp.

In terms of Sialidase and Peptidase_AA cluster, SHAP analysis results also demonstrated that the improvements in Random-CV tests were tightly related to shared features between similar complexes. In Sialidase cluster, RF-Score could roughly discriminate the stronger complexes from weaker complexes (Rp=0.65) in Random-CV tests, while assigned them with similar binding affinities (Rp=0.27) in Pfam-CV because of similar buried SASA (Figure 10A and S10). Some similar complexes with ligand RDKit topological fingerprint similarities more than 0.8 were annotated in Figure 10A and these complexes all formed strong hydrogen bond interactions. For example, in 3ti5, hydrogen bonds were identified between ligand carboxyl group and three protein Arginines, and between ligand guanidyl group and carbonyl groups in protein backbone (Figure 10B). Random-CV RF-Score recognized these patterns from similar structures in the training dataset and assigned greater importance to the Vina hydrogen term compared with the SHAP values in the Pfam-CV RF-Score. Moreover, the specific C-N and N-N interactions from the ligand guanidyl group were also essential in the final scores (Figure 10C and 10D). From these patterns in the training dataset, the Random-CV RF-Score assessed that the binding affinity range of these complexes was higher than the average one and predicted them with higher values.

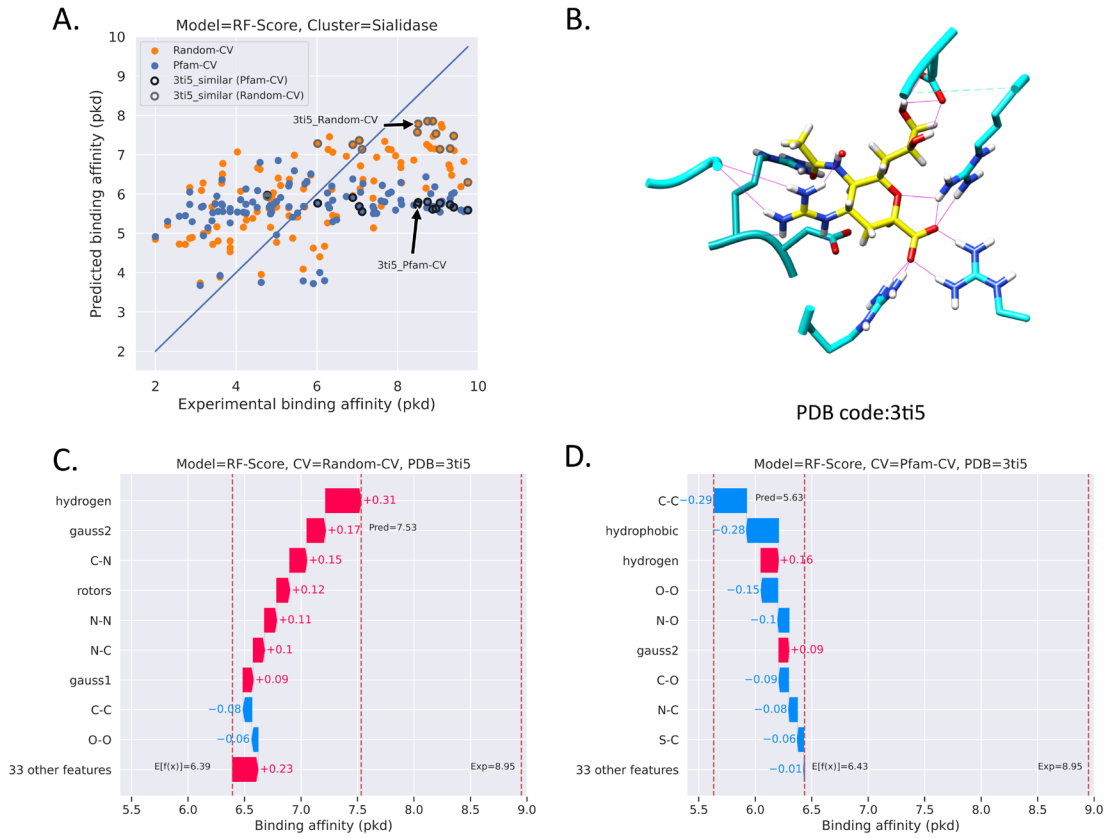

Figure 10. RF-Score important features in Sialidase. A. RF-Score predicted labels on Sialidase in Pfam-CV and Random-CV experiments. Complexes with highly similar ligands to 3ti5 ligand were annotated. B. 3ti5 binding pocket. G and H. SHAP analysis results on 3ti5 in Random-CV and Pfam-CV tests.

In Peptidase_AA, RF-Score performed similarly on complexes with experimental binding affinities lower than 8 pKd in Pfam-CV and Random-CV tests. In terms of other complexes, the predicted labels were much higher in Random-CV test than Random-CV test (Figure 11A). It was not surprising that these well predicted complexes had a highly similar ligand in the training dataset. For example, the proteins of annotated complexes in Figure 11A were HIV proteins and the ligands were darunavir (DRV) or were similar to DRV (Figure 11B). We discovered that though the most critical features of 6oou (one of DRV-HIV complexes) were identical in two cross-validation tests: C-C interaction, N-C interaction, and gauss2 (Figures 11C and 11D), the weights were slightly different. These characteristics contributed more positively in Random-CV test. Besides, several specific interactions with ligand S atom, such as C-S, N-S, and O-S interactions, were also captured in Random-CV test. According to the unique characteristics of DRV-HIV, RF-Score allocated specific weights to these characteristics in Random-CV test and performed better predictions on these complexes.

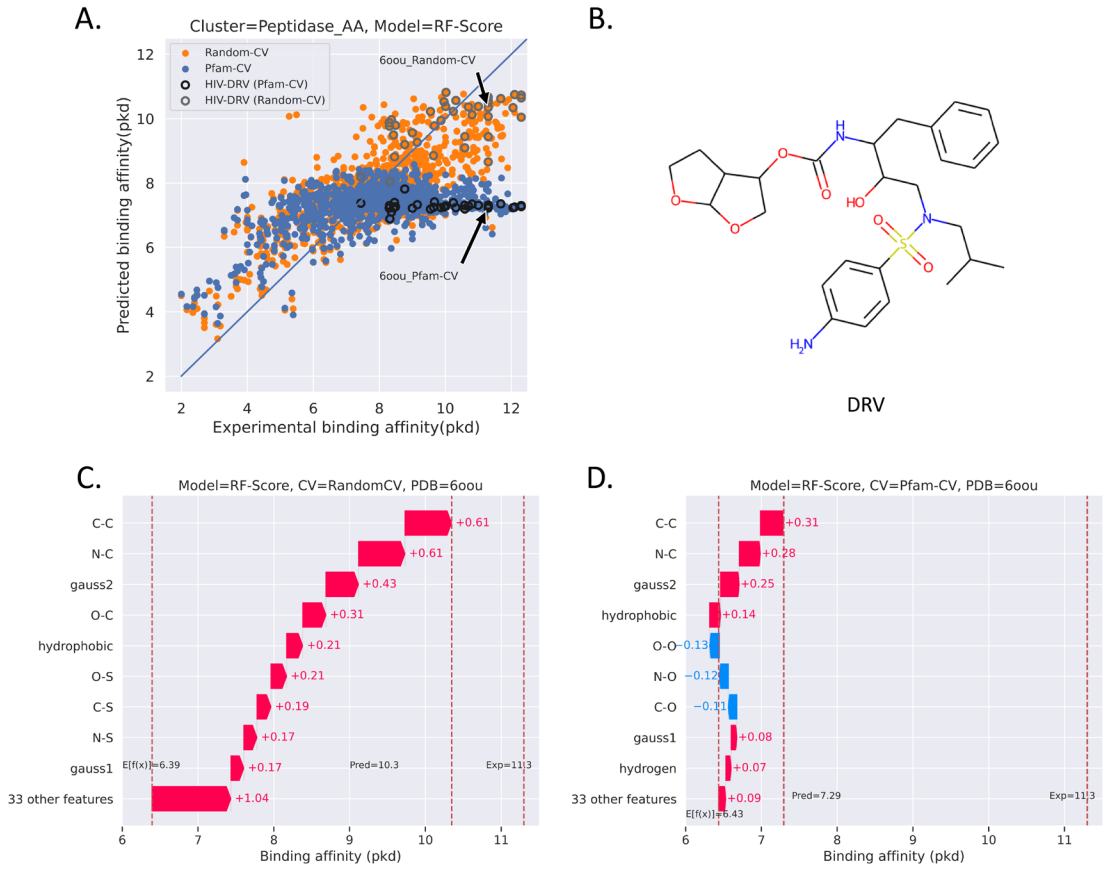

Figure 11. RF-Score important features in Peptidase_AA. A. RF-Score predicted labels on Peptidase_AA in Pfam-CV and Random-CV tests. The HIV-DRV complexes were annotated. B. 2D structure of DRV. C and D. SHAP analysis results on 6oou in Random-CV and Pfam-CV experiments.

Conclusion and Discussion

In this work, we first investigated the limitations of the protein sequence similarity-based cluster approach and established the first standardized protein cluster method for evaluating generalization ability of MLSFs. The Pfam-cluster approach rigorously and clearly assembled proteins with similar ligand-binding pockets based on the pocket Pfam domains. Thus, we highly recommend considering the Pfam-cluster approach when evaluating MLSF performances on novel targets.

The generalization ability benchmark results showed that the cross-target scoring power of MLSFs was far from satisfying, though both the PDBbind dataset and model algorithms improved a lot. In addition, the further interpretable analysis indicated that MLSFs mainly relied on buried SASA-related information to make decisions on novel targets, with higher predicted binding affinities on complexes with larger protein-ligand interfaces. The intrinsic property that larger ligands performing higher binding affinities of PDBbind dataset accounts for this preference. Besides, the ligand sizes-binding affinities correlations were especially noticeable when targeting the protein targets, such as Cyclophil-like, SH2-like, and Calcineurin. Data-driven algorithms

preferred such a hidden correlation and relied on it to make decisions. This is consistent with statistical knowledge-based SFs, such as PMF@DS, DrugScore, and Convex-PL, which also showed a strong preference for ligands with larger protein-ligand interfaces in CASF virtual screening benchmarks[80].

It was difficult to conclude in terms of whether buried SASA-related information was useful or not in SBVS. Unfortunately, we did not identify any reliable features when interpreting similar complexes. When randomly splitting the PDBbind dataset, both similar proteins and similar ligands were added to the training dataset. RF-Score made better predictions on the Random-CV test by recognizing the specific feature combinations of similar ligands, such as the ligand S atom environment in Avidin-biotin complexes and HIV-DRV complexes, the high hydrogen-bonding interactions in Sialidase, and the ligand guanidyl group in some Sialidase complexes. Combining these specific features with the basic buried SASA-related features, RF-Score remembered the experimental binding affinity ranges of these highly similar structures and predicted better scores. Nevertheless, most of these patterns lack physical knowledge and were nonsensical.

According to our benchmark and interpretable analytical results, the implementations of the MLSFs on real SBVS scenarios were concerning. It is critical to build an unbiased dataset and rigorously evaluate MLSFs in order to generate reliable and robust MLSFs. Because the PDBbind dataset only contains crystal structures of true binders, it revealed a significantly positive binding affinity-buried SASA connection in some clusters. However, in practice, the observation that larger molecules have higher binding affinities does not always hold true. Thus, adding more negative data to the dataset helps to dilute this artificial link. The ChEMBL database has more than 17 million samples with experimentally determined binding affinities[81,82]. The 3D structures would be achieved by proper docking and modeling methods. Additionally, artificially redocked poses also help models minimize the importance of ligand size. Many attempts have been made to distinguish between crystal and decoy poses, as well as to predict binding affinity with redocked poses[24,83–86]. However, considerable care should be taken in loss function and model design because of data imbalance. Models could iteratively train the hard samples and be more robust in unseen datasets by combining these augmented datasets with some specific methodologies, such as active learning, adversarial learning, and uncertainty evaluation[87].

Although the underlying assumption of ML algorithms is that training data is distributed identically to the testing data, this assumption would never hold in real-world applications. This is particularly true in the scenario of SBVS, the main challenge of which is to discover true binders with novel chemical scaffolds against new drug targets. As a result, it is crucial to assess MLSF generalization capacity and improve it through data augmentation and model optimization. Our generalization benchmark was designed from the beginning as an open-access benchmark. The complete Pfam-cluster approach, 3-fold dataset split, and SHAP analysis processes are available on https://github.com/hnlab/generalization_benckmark.

Supporting information

Supplementary_figures.pdf

Supplementary_tables.xlsx

Author Information

Corresponding Authors

Niu Huang - National Institute of Biological Sciences, Beijing, Beijing, 102206, China; Tsinghua Institute of Multidisciplinary Biomedical Research, Tsinghua University, Beijing, 102206, China; http://orcid.org/0000-0002-6912-033X; Email: huangniu@nibs.ac.cn

Jincai Yang - National Institute of Biological Sciences, Beijing, Beijing, 102206, China; http://orcid.org/0000-0002-0033-0187; Email: yangjincai@nibs.ac.cn

Authors

Hui Zhu - National Institute of Biological Sciences, Beijing, Beijing, 102206, China; Tsinghua Institute of Multidisciplinary Biomedical Research, Tsinghua University, Beijing, 102206, China; http://orcid.org/0000-0002-4024-374X; Email: zhuhui@nibs.ac.cn

Notes

The authors declare no competing financial interest.

The complete Pfam-cluster approach, 3-fold dataset split, and SHAP analysis processes are available on https://github.com/hnlab/generalization_benckmark. All other data are also available upon request.

Acknowledgments. This work was supported by Beijing Municipal Science & Technology Commission (Z201100005320012 to N.H.), Beijing Postdoctoral Research Foundation (2022-ZZ-012 to J.Y.) and Tsinghua University.

Abbreviations

SBVS: structure-based virtual screening

MLSF: machine-learning scoring function

Pfam-cluster: pocket Pfam-based cluster

Seq-cluster: protein sequence similarity-based cluster

Random-CV: random cross-validation

Pfam-CV: protein sequence similarity-based cross-validation

Seq-CV: pocket Pfam-based cross-validation

SASA: solvent accessible surface area

RF: random forest

XGB: extreme gradient boosting

CNN: convolutional neural network

GNN: graph neural networks

References

(1) Shoichet, B. K. Virtual Screening of Chemical Libraries. *Nature* **2004**, *432* (7019), 862–865.

https://doi.org/10.1038/nature03197.

(2) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566* (7743), 224–229. https://doi.org/10.1038/s41586-019-0917-9.

(3) Zhou, Y.; Ma, J.; Lin, X.; Huang, X.-P.; Wu, K.; Huang, N. Structure-Based Discovery of Novel and Selective 5-Hydroxytryptamine 2B Receptor Antagonists for the Treatment of Irritable Bowel Syndrome. *J. Med. Chem.* **2016**, *59* (2), 707–720. https://doi.org/10.1021/acs.jmedchem.5b01631.

(4) Peng, S.; Xiao, W.; Ju, D.; Sun, B.; Hou, N.; Liu, Q.; Wang, Y.; Zhao, H.; Gao, C.; Zhang, S.; Cao, R.; Li, P.; Huang, H.; Ma, Y.; Wang, Y.; Lai, W.; Ma, Z.; Zhang, W.; Huang, S.; Wang, H.; Zhang, Z.; Zhao, L.; Cai, T.; Zhao, Y.-L.; Wang, F.; Nie, Y.; Zhi, G.; Yang, Y.-G.; Zhang, E. E.; Huang, N. Identification of Entacapone as a Chemical Inhibitor of FTO Mediating Metabolic Regulation through FOXO1. *Science Translational Medicine* **2019**, *11* (488). https://doi.org/10.1126/scitranslmed.aau7116.

(5) Irwin, J. J.; Shoichet, B. K. Docking Screens for Novel Ligands Conferring New Biology. *J. Med. Chem.* **2016**, *59* (9), 4103–4120. https://doi.org/10.1021/acs.jmedchem.5b02008.

(6) Guedes, I. A.; Pereira, F. S. S.; Dardenne, L. E. Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Front Pharmacol* **2018**, *9*, 1089. https://doi.org/10.3389/fphar.2018.01089.

(7) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739–1749. https://doi.org/10.1021/jm0306430.

(8) Dittrich, J.; Schmidt, D.; Pfleger, C.; Gohlke, H. Converging a Knowledge-Based Scoring Function: DrugScore2018. *J. Chem. Inf. Model.* **2019**, *59* (1), 509–521. https://doi.org/10.1021/acs.jcim.8b00582.

(9) Velec, H. F. G.; Gohlke, H.; Klebe, G. DrugScore(CSD)-Knowledge-Based Scoring Function Derived from Small Molecule Crystal Data with Superior Recognition Rate of near-Native Ligand Poses and Better Affinity Prediction. *J Med Chem* **2005**, *48* (20), 6296–6303. https://doi.org/10.1021/jm050436v.

(10) Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44* (7), 1035–1042. https://doi.org/10.1021/jm0003992.

(11) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on a Diverse Test Set. *J Chem Inf Model* **2009**, *49* (4), 1079–1093. https://doi.org/10.1021/ci9000053.

(12) Khamis, M. A.; Gomaa, W. Comparative Assessment of Machine-Learning Scoring Functions on PDBbind 2013. *Engineering Applications of Artificial Intelligence* **2015**, *45*, 136–151. https://doi.org/10.1016/j.engappai.2015.06.021.

(13) 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. https://doi.org/10.1021/acs.jcim.8b00545.

(14) Huang, N.; Kalyanaraman, C.; Bernacki, K.; Jacobson, M. P. Molecular Mechanics Methods for Predicting Protein-Ligand Binding. *Phys Chem Chem Phys* **2006**, *8* (44), 5166–5177. https://doi.org/10.1039/b608269f.

(15) 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. *WIREs Computational Molecular Science* **2020**, *10* (1), e1429. https://doi.org/10.1002/wcms.1429.

(16) Li, J.; Fu, A.; Zhang, L. An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. *Interdiscip Sci Comput Life Sci* **2019**, *11* (2), 320–328. https://doi.org/10.1007/s12539-019-00327-w.

(17) Ain, Q. U.; Aleksandrova, A.; Roessler, F. D.; Ballester, P. J. Machine-Learning Scoring Functions to Improve Structure-Based Binding Affinity Prediction and Virtual Screening. *WIREs Computational Molecular Science* **2015**, *5* (6), 405–424. https://doi.org/10.1002/wcms.1225.

(18) Breiman, L. Random Forests. *Machine Learning* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324.

(19) Ballester, P. J.; Mitchell, J. B. O. A Machine Learning Approach to Predicting Protein-Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics* **2010**, *26* (9), 1169–1175. https://doi.org/10.1093/bioinformatics/btq112.

(20) Zilian, D.; Sotriffer, C. A. SFCscoreRF: A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein–Ligand Complexes. *J. Chem. Inf. Model.* **2013**, *53* (8), 1923–1933. https://doi.org/10.1021/ci400120b.

(21) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: San Francisco California USA, 2016; 785–794. https://doi.org/10.1145/2939672.2939785.

(22) Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; De Fabritiis, G. KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* **2018**, *58* (2), 287–296. https://doi.org/10.1021/acs.jcim.7b00650.

(23) Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P. Development and Evaluation of a Deep Learning Model for Protein-Ligand Binding Affinity Prediction. **2017**. https://doi.org/10.1093/bioinformatics/bty374/4994792.

(24) Francoeur, P. G.; Masuda, T.; Sunseri, J.; Jia, A.; Iovanisci, R. B.; Snyder, I.; Koes, D. R. Three-Dimensional Convolutional Neural Networks and a Cross-Docked Data Set for Structure-Based Drug Design. *J. Chem. Inf. Model.* **2020**, *60* (9), 4200–4215. https://doi.org/10.1021/acs.jcim.0c00411.

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

(26) Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V. S. PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **2018**, *4* (11), 1520–1530. https://doi.org/10.1021/acscentsci.8b00507

(27) Jiang, D.; Hsieh, C.-Y.; Wu, Z.; Kang, Y.; Wang, J.; Wang, E.; Liao, B.; Shen, C.; Xu, L.; Wu, J.; Cao, D.; Hou, T. InteractionGraphNet: A Novel and Efficient Deep Graph Representation Learning Framework for Accurate Protein-Ligand Interaction Predictions. *J Med Chem* **2021**, *64* (24), 18209–18232. https://doi.org/10.1021/acs.jmedchem.1c01830.

(28) Jones, D.; Kim, H.; Zhang, X.; Zemla, A.; Stevenson, G.; Bennett, W. F. D.; Kirshner, D.; Wong, S. E.; Lightstone, F. C.; Allen, J. E. Improved Protein–Ligand Binding Affinity Prediction with Structure-Based Deep Fusion Inference. *J. Chem. Inf. Model.* **2021**, *61* (4), 1583–1592. https://doi.org/10.1021/acs.jcim.0c01306.

(29) Son, J.; Kim, D.; Development of a graph convolutional neural network model for efficient prediction of protein-ligand binding affinities. *PLoS ONE.* **2021**, *16 (4)*, e0249404. https://doi.org/10.1371/journal.pone.0249404.

(30) Li, S.; Zhou, J.; Xu, T.; Huang, L.; Wang, F.; Xiong, H.; Huang, W.; Dou, D.; Xiong, H. Structure-Aware Interactive Graph Neural Networks for the Prediction of Protein-Ligand Binding Affinity. **2021**. https://doi.org/10.1145/3447548.3467311.

(31) Ashtawy, H. M.; Mahapatra, N. R. A Comparative Assessment of Ranking Accuracies of Conventional and Machine-Learning-Based Scoring Functions for Protein-Ligand Binding Affinity Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2012**, *9* (5), 1301–1313. https://doi.org/10.1109/TCBB.2012.36.

(32) Wicht, K. J.; Combrinck, J. M.; Smith, P. J.; Egan, T. J. Bayesian Models Trained with HTS

Data for Predicting β-Haematin Inhibition and in Vitro Antimalarial Activity. *Bioorg Med Chem* **2015**, *23* (16), 5210–5217. https://doi.org/10.1016/j.bmc.2014.12.020.

(33) Zhang, H.; Liu, W.; Liu, Z.; Ju, Y.; Xu, M.; Zhang, Y.; Wu, X.; Gu, Q.; Wang, Z.; Xu, J. Discovery of Indoleamine 2,3-Dioxygenase Inhibitors Using Machine Learning Based Virtual Screening. *MedChemComm* **2018**. https://doi.org/10.1039/c7md00642j.

(34) Wang, Y.; Dai, Y.; Wu, X.; Li, F.; Liu, B.; Li, C.; Liu, Q.; Zhou, Y.; Wang, B.; Zhu, M.; Cui, R.; Tan, X.; Xiong, Z.; Liu, J.; Tan, M.; Xu, Y.; Geng, M.; Jiang, H.; Liu, H.; Ai, J.; Zheng, M. Discovery and Development of a Series of Pyrazolo[3,4-d]Pyridazinone Compounds as the Novel Covalent Fibroblast Growth Factor Receptor Inhibitors by the Rational Drug Design. *J. Med. Chem.* **2019**, *62* (16), 7473–7488. https://doi.org/10.1021/acs.jmedchem.9b00510.

(35) Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; Vinyals, O. Understanding Deep Learning (Still) Requires Rethinking Generalization. *Commun. ACM* **2021**, *64* (3), 107–115. https://doi.org/10.1145/3446776.

(36) Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; Wichmann, F. A. Shortcut Learning in Deep Neural Networks. *Nat Mach Intell* **2020**, *2* (11), 665–673. https://doi.org/10.1038/s42256-020-00257-z.

(37) Yang, J.; Shen, C.; Huang, N. Predicting or Pretending: Artificial Intelligence for Protein-Ligand Interactions Lack of Sufficiently Large and Unbiased Datasets. *Frontiers in Pharmacology* **2020**, *11*, 69. https://doi.org/10.3389/fphar.2020.00069.

(38) Kramer, C.; Gedeck, P. Leave-Cluster-Out Cross-Validation Is Appropriate for Scoring Functions Derived from Diverse Protein Data Sets. *J. Chem. Inf. Model.* **2010**, *50* (11), 1961–1969. https://doi.org/10.1021/ci100264e.

(39) Nikolaienko, T.; Gurbych, O.; Druchok, M. Complex Machine Learning Model Needs Complex Testing: Examining Predictability of Molecular Binding Affinity by a Graph Neural Network. *Journal of Computational Chemistry* **2022**, *43* (10), 728–739. https://doi.org/10.1002/jcc.26831.

(40) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *Journal of medicinal chemistry* **2005**, *48*, 4111–4119. https://doi.org/10.1021/jm048957q.

(41) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25* (11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.

(42) *RDKit*. https://www.rdkit.org/ (accessed 2022-05-24).

(43) Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G. A.; Sonnhammer, E. L. L.; Tosatto, S. C. E.; Paladin, L.; Raj, S.; Richardson, L. J.; Finn, R. D.; Bateman, A. Pfam: The Protein Families Database in 2021. *Nucleic Acids Research* **2021**, *49* (D1), D412–D419. https://doi.org/10.1093/nar/gkaa913.

(44) Finn, R. D.; Mistry, J.; Schuster-Böckler, B.; Griffiths-Jones, S.; Hollich, V.; Lassmann, T.; Moxon, S.; Marshall, M.; Khanna, A.; Durbin, R.; Eddy, S. R.; Sonnhammer, E. L. L.; Bateman, A. Pfam: Clans, Web Tools and Services. *Nucleic Acids Res* **2006**, *34* (Database issue), D247-251. https://doi.org/10.1093/nar/gkj149.

(45) Xu, Q.; Dunbrack, R. L. Assignment of Protein Sequences to Existing Domain and Family Classification Systems: Pfam and the PDB. *Bioinformatics* **2012**, *28* (21), 2763–2772. https://doi.org/10.1093/bioinformatics/bts533.

(46) Finn, R. D.; Clements, J.; Eddy, S. R. HMMER Web Server: Interactive Sequence Similarity Searching. *Nucleic Acids Res* **2011**, *39* (Web Server issue), W29–W37. https://doi.org/10.1093/nar/gkr367.

(47) Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P. J. Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets. *Molecular Informatics* **2015**, *34* (2–3), 115–126. https://doi.org/10.1002/minf.201400132.

(48) Durrant, J. D.; McCammon, J. A. NNScore 2.0: A Neural-Network Receptor–Ligand Scoring Function. *J. Chem. Inf. Model.* **2011**, *51* (11), 2897–2903. https://doi.org/10.1021/ci2003889.

(49) 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. https://doi.org/10.1021/acsomega.9b01997.

(50) 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), 26. https://doi.org/10.1186/s13321-015-0078-2.

(51) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12* (85), 2825–2830.

(52) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems* **2017**, *30*.

(53) Noble, M. E. M.; Endicott, J. A.; Johnson, L. N. Protein Kinase Inhibitors: Insights into Drug Design from Structure. *Science* **2004**, *303* (5665), 1800–1805. https://doi.org/10.1126/science.1095920.

(54) De Fusco, C.; Brear, P.; Iegre, J.; Georgiou, K. H.; Sore, H. F.; Hyvönen, M.; Spring, D. R. A Fragment-Based Approach Leading to the Discovery of a Novel Binding Site and the Selective CK2 Inhibitor CAM4066. *Bioorganic & Medicinal Chemistry* **2017**, *25* (13), 3471–3482. https://doi.org/10.1016/j.bmc.2017.04.037.

(55) Wood, E. R.; Truesdale, A. T.; McDonald, O. B.; Yuan, D.; Hassell, A.; Dickerson, S. H.; Ellis, B.; Pennisi, C.; Horne, E.; Lackey, K.; Alligood, K. J.; Rusnak, D. W.; Gilmer, T. M.; Shewchuk, L. A Unique Structure for Epidermal Growth Factor Receptor Bound to GW572016 (Lapatinib): Relationships among Protein Conformation, Inhibitor Off-Rate, and Receptor Activity in Tumor Cells. *Cancer Research* **2004**, *64* (18), 6652–6659. https://doi.org/10.1158/0008-5472.CAN-04-1168.

(56) Leonidas, D. D.; Maiti, T. K.; Samanta, A.; Dasgupta, S.; Pathak, T.; Zographos, S. E.; Oikonomakos, N. G.; The binding of 3′-N-piperidine-4-carboxyl-3′-deoxy-ara-uridine to ribonuclease A in the crystal. Bioorganic & medicinal chemistry **2006**, *14* (17), 6055-6064. https://doi.org/10.1016/j.bmc.2006.05.011

(57) Varghese, J. N.; Smith, P. W.; Sollis, S. L.; Blick, T. J.; Sahasrabudhe, A.; McKimm-Breschkin, J. L.; Colman, P. M. Drug Design against a Shifting Target: A Structural Basis for Resistance to Inhibitors in a Variant of Influenza Virus Neuraminidase. *Structure* **1998**, *6* (6), 735–746. https://doi.org/10.1016/S0969-2126(98)00075-6.

(58) Andersson, H. O.; Fridborg, K.; Löwgren, S.; Alterman, M.; Mühlman, A.; Björsne, M.; Garg, N.; Kvarnström, I.; Schaal, W.; Classon, B.; Karlén, A.; Danielsson, U. H.; Ahlsén, G.; Nillroth, U.; Vrang, L.; Öberg, B.; Samuelsson, B.; Hallberg, A.; Unge, T. Optimization of P1–P3 Groups in Symmetric and Asymmetric HIV-1 Protease Inhibitors. *European Journal of Biochemistry* **2003**, *270* (8), 1746–1758. https://doi.org/10.1046/j.1432-1033.2003.03533.x.

(59) Rahuel, J.; Rasetti, V.; Maibaum, J.; Rüeger, H.; Göschke, R.; Cohen, N.-C.; Stutz, S.; Cumin, F.; Fuhrer, W.; Wood, J.; Grütter, M. Structure-Based Drug Design: The Discovery of Novel Nonpeptide Orally Active Inhibitors of Human Renin. *Chemistry & Biology* **2000**, *7* (7), 493–504. https://doi.org/10.1016/S1074-5521(00)00134-4.

(60) Asada, H.; Horita, S.; Hirata, K.; Shiroishi, M.; Shiimura, Y.; Iwanari, H.; Hamakubo, T.; Shimamura, T.; Nomura, N.; Kusano-Arai, O.; Uemura, T.; Suno, C.; Kobayashi, T.; Iwata, S. Crystal Structure of the Human Angiotensin II Type 2 Receptor Bound to an Angiotensin II Analog. *Nat Struct Mol Biol* **2018**, *25* (7), 570–576. https://doi.org/10.1038/s41594-018-0079-8.

(61) Dwivedi, A. K.; Gurjar, V.; Kumar, S.; Singhl, N. Molecular Basis for Nonspecificity of Nonsteroidal Anti-Inflammatory Drugs (NSAIDs). *Drug Discovery Today* **2015**, *20* (7), 863–873. https://doi.org/10.1016/j.drudis.2015.03.004.

(62) Li, H.; Peng, J.; Leung, Y.; Leung, K.-S.; Wong, M.-H.; Lu, G.; Ballester, P. J. The Impact of Protein Structure and Sequence Similarity on the Accuracy of Machine-Learning Scoring Functions for Binding Affinity Prediction. *Biomolecules* **2018**, *8* (1), 12. https://doi.org/10.3390/biom8010012.

(63) Li, Y.; Yang, J. Structural and Sequence Similarity Makes a Significant Impact on Machine-Learning-Based Scoring Functions for Protein–Ligand Interactions. *J. Chem. Inf. Model.* **2017**, *57* (4), 1007–1012. https://doi.org/10.1021/acs.jcim.7b00049.

(64) Su, M.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Tapping on the Black Box: How Is the Scoring Power of a Machine-Learning Scoring Function Dependent on the Training Set? *J. Chem. Inf. Model.* **2020**, *60* (3), 1122–1136. https://doi.org/10.1021/acs.jcim.9b00714.

(65) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera--a Visualization System for Exploratory Research and Analysis. *J Comput Chem* **2004**, *25* (13), 1605–1612. https://doi.org/10.1002/jcc.20084.

(66) Greener, J. G.; Kandathil, S. M.; Moffat, L.; Jones, D. T. A Guide to Machine Learning for Biologists. *Nat Rev Mol Cell Biol* **2022**, *23* (1), 40–55. https://doi.org/10.1038/s41580-021-00407-0.

(67) Jones, D. T.; Thornton, J. M. The Impact of AlphaFold2 One Year On. *Nat Methods* **2022**, *19* (1), 15–20. https://doi.org/10.1038/s41592-021-01365-3.

(68) Fodor, M.; Price, E.; Wang, P.; Lu, H.; Argintaru, A.; Chen, Z.; Glick, M.; Hao, H.-X.; Kato, M.; Koenig, R.; LaRochelle, J. R.; Liu, G.; McNeill, E.; Majumdar, D.; Nishiguchi, G. A.; Perez, L. B.; Paris, G.; Quinn, C. M.; Ramsey, T.; Sendzik, M.; Shultz, M. D.; Williams, S. L.; Stams, T.; Blacklow, S. C.; Acker, M. G.; LaMarche, M. J. Dual Allosteric Inhibition of SHP2 Phosphatase. *ACS Chem. Biol.* **2018**, *13* (3), 647–656. https://doi.org/10.1021/acschembio.7b00980.

(69) Chen, Y.-N. P.; LaMarche, M. J.; Chan, H. M.; Fekkes, P.; Garcia-Fortanet, J.; Acker, M. G.; Antonakos, B.; Chen, C. H.-T.; Chen, Z.; Cooke, V. G.; Dobson, J. R.; Deng, Z.; Fei, F.; Firestone, B.; Fodor, M.; Fridrich, C.; Gao, H.; Grunenfelder, D.; Hao, H.-X.; Jacob, J.; Ho, S.; Hsiao, K.; Kang, Z. B.; Karki, R.; Kato, M.; Larrow, J.; La Bonte, L. R.; Lenoir, F.; Liu, G.; Liu, S.; Majumdar, D.; Meyer, M. J.; Palermo, M.; Perez, L.; Pu, M.; Price, E.; Quinn, C.; Shakya, S.; Shultz, M. D.; Slisz, J.; Venkatesan, K.; Wang, P.; Warmuth, M.; Williams, S.; Yang, G.; Yuan, J.; Zhang, J.-H.; Zhu, P.; Ramsey, T.; Keen, N. J.; Sellers, W. R.; Stams, T.; Fortin, P. D. Allosteric Inhibition of SHP2 Phosphatase Inhibits Cancers Driven by Receptor Tyrosine Kinases. *Nature* **2016**, *535* (7610), 148–152. https://doi.org/10.1038/nature18621.

(70) Lange, G.; Lesuisse, D.; Deprez, P.; Schoot, B.; Loenze, P.; Bénard, D.; Marquette, J.-P.; Broto, P.; Sarubbi, E.; Mandine, E. Requirements for Specific Binding of Low Affinity Inhibitor Fragments to the SH2 Domain of (Pp60)Src Are Identical to Those for High Affinity Binding of Full Length Inhibitors. *J Med Chem* **2003**, *46* (24), 5184–5195. https://doi.org/10.1021/jm020970s.

(71) Maynes, J. T.; Luu, H. A.; Cherney, M. M.; Andersen, R. J.; Williams, D.; Holmes, C. F. B.; James, M. N. G. Crystal Structures of Protein Phosphatase-1 Bound to Motuporin and Dihydromicrocystin-LA: Elucidation of the Mechanism of Enzyme Inhibition by Cyanobacterial Toxins. *Journal of Molecular Biology* **2006**, *356* (1), 111–120. https://doi.org/10.1016/j.jmb.2005.11.019.

(72) Kita, A.; Matsunaga, S.; Takai, A.; Kataiwa, H.; Wakimoto, T.; Fusetani, N.; Isobe, M.; Miki, K. Crystal Structure of the Complex between Calyculin A and the Catalytic Subunit of Protein Phosphatase 1. *Structure* **2002**, *10* (5), 715–724. https://doi.org/10.1016/S0969-2126(02)00764-5.

(73) Feder, D.; Hussein, W. M.; Clayton, D. J.; Kan, M.-W.; Schenk, G.; McGeary, R. P.; Guddat, L. W. Identification of Purple Acid Phosphatase Inhibitors by Fragment-Based Screening: Promising New Leads for Osteoporosis Therapeutics. *Chemical Biology & Drug Design* **2012**, *80* (5), 665–674. https://doi.org/10.1111/cbdd.12001.

(74) Kelker, M. S.; Page, R.; Peti, W. Crystal Structures of Protein Phosphatase-1 Bound to Nodularin-R and Tautomycin: A Novel Scaffold for Structure Based Drug Design of Serine/Threonine Phosphatase Inhibitors. *J Mol Biol* **2009**, *385* (1), 11–21. https://doi.org/10.1016/j.jmb.2008.10.053.

(75) Shibata, A.; Moiani, D.; Arvai, A. S.; Perry, J.; Harding, S. M.; Genois, M.-M.; Maity, R.; van Rossum-Fikkert, S.; Kertokalio, A.; Romoli, F.; Ismail, A.; Ismalaj, E.; Petricci, E.; Neale, M. J.; Bristow, R. G.; Masson, J.-Y.; Wyman, C.; Jeggo, P. A.; Tainer, J. A. DNA Double-Strand Break Repair Pathway Choice Is Directed by Distinct MRE11 Nuclease Activities. *Molecular Cell* **2014**, *53* (1), 7–18. https://doi.org/10.1016/j.molcel.2013.11.003.

(76) Sedrani, R.; Kallen, J.; Martin Cabrejas, L. M.; Papageorgiou, C. D.; Senia, F.; Rohrbach, S.; Wagner, D.; Thai, B.; Jutzi Eme, A.-M.; France, J.; Oberer, L.; Rihs, G.; Zenke, G.; Wagner, J. Sanglifehrin−Cyclophilin Interaction: Degradation Work, Synthetic Macrocyclic Analogues, X-Ray Crystal Structure, and Binding Data. *J. Am. Chem. Soc.* **2003**, *125* (13), 3849–3859. https://doi.org/10.1021/ja021327y.

(77) Steadman, V. A.; Pettit, S. B.; Poullennec, K. G.; Lazarides, L.; Keats, A. J.; Dean, D. K.; Stanway, S. J.; Austin, C. A.; Sanvoisin, J. A.; Watt, G. M.; Fliri, H. G.; Liclican, A. C.; Jin, D.; Wong, M. H.; Leavitt, S. A.; Lee, Y.-J.; Tian, Y.; Frey, C. R.; Appleby, T. C.; Schmitz, U.; Jansa, P.; Mackman, R. L.; Schultz, B. E. Discovery of Potent Cyclophilin Inhibitors Based on the Structural Simplification of Sanglifehrin A. J. Med. Chem. 2017, 60 (3), 1000–1017.

https://doi.org/10.1021/acs.jmedchem.6b01329.

(78) Kontopidis, G.; Taylor, P.; Walkinshaw, M. D.; Enzymatic and Structural Characterization of Non-Peptide Ligand-Cyclophilin Complexes. *Acta Cryst.* **2004**, *D60*, 479-485. https://doi.org/10.1107/S0907444904000174

(79) Wear, M. A.; Nowicki, M. W.; Blackburn, E. A.; McNae, I. W.; Walkinshaw, M. D. Thermo-Kinetic Analysis Space Expansion for Cyclophilin-Ligand Interactions – Identification of a New Nonpeptide Inhibitor Using Biacore™ T200. *FEBS Open Bio* **2017**, *7* (4), 533–549. https://doi.org/10.1002/2211-5463.12201.

(80) Kadukova, M.; Grudinin, S. Convex-PL: A Novel Knowledge-Based Potential for Protein-Ligand Interactions Deduced from Structural Databases Using Convex Optimization. *Journal of computer-aided molecular design* **2017**, *31*. https://doi.org/10.1007/s10822-017-0068-8.

(81) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res* **2017**, *45* (D1), D945–D954. https://doi.org/10.1093/nar/gkw1074.

(82) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Research* **2012**, *40* (D1), D1100–D1107. https://doi.org/10.1093/nar/gkr777.

(83) Shen, C.; Hu, X.; Gao, J.; Zhang, X.; Zhong, H.; Wang, Z.; Xu, L.; Kang, Y.; Cao, D.; Hou, T. The Impact of Cross-Docked Poses on Performance of Machine Learning Classifier for Protein–Ligand Binding Pose Prediction. *J Cheminform* **2021**, *13* (1), 81. https://doi.org/10.1186/s13321-021-00560-w.

(84) Gabel, J.; Desaphy, J.; Rognan, D. Beware of Machine Learning-Based Scoring Functions—On the Danger of Developing Black Boxes. *J. Chem. Inf. Model.* **2014**, *54* (10), 2807–2815. https://doi.org/10.1021/ci500406k.

(85) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein–Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57* (4), 942–957. https://doi.org/10.1021/acs.jcim.6b00740.

(86) Ashtawy, H. M.; Mahapatra, N. R. Task-Specific Scoring Functions for Predicting Ligand Binding Poses and Affinity and for Screening Enrichment. *J. Chem. Inf. Model.* **2018**, *58* (1), 119–133. https://doi.org/10.1021/acs.jcim.7b00309.

(87) Qiao, F.; Peng, X. Uncertainty-Guided Model Generalization to Unseen Domains. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Nashville, TN, USA, 2021; 6786–6796. https://doi.org/10.1109/CVPR46437.2021.00672.