

Bioinformatics, 2023, 1–7

doi: DOI HERE

Advance Access Publication Date: Day Month Year

Original Paper



Robustly interrogating machine learning-based scoring functions: what are they learning?

Guy Durant¹,¹ Fergus Boyles¹,¹ Kristian Birchall²,² Brian Marsden³
and Charlotte M. Deane¹

¹Department of Statistics, University of Oxford, St Giles', OX1 3LB, Oxford, United Kingdom, ²LifeArc, Gunnels Wood Road, SG1 2FX, Stevenage, United Kingdom and ³Nuffield Department of Medicine, University of Oxford, Old Road, OX3 7BN, Oxford, United Kingdom

*Corresponding author. deane@stats.ox.ac.uk

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Motivation: Machine learning-based scoring functions (MLBSFs) have been found to exhibit inconsistent performance on different benchmarks and be prone to learning dataset bias. For the field to develop MLBSFs that learn a generalisable understanding of physics, a more rigorous understanding of how they perform is required.

Results: In this work, we compared the performance of a diverse set of popular MLBSFs (RFScore, SIGN, OnionNet-2, Pafnucy, and PointVS) to our proposed baseline models that can only learn dataset biases on a range of benchmarks. We found that these baseline models were competitive in accuracy to these MLBSFs in almost all proposed benchmarks, indicating these models only learn dataset biases. Our tests and provided platform, ToolBoxSF, will enable researchers to robustly interrogate MLBSF performance and determine the effect of dataset biases on their predictions.

Availability and Implementation: <https://github.com/guydurant/toolboxsf>

Contact: deane@stats.ox.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

Introduction

Predicting the binding affinity of a protein-ligand complex from its 3D structure has been extensively researched in the past decade (Meli *et al.*, 2022). However, doing so accurately and for any protein-ligand complex still poses a significant challenge in computational chemistry (Mobley and Gilson, 2017). Accurately predicting binding affinity would aid in structure-based drug discovery, where the chemical structure of a drug is designed based on the structure of its target, as it would allow design hypotheses to be tested in silico. One proposed methodology, scoring functions (Goodsell *et al.*, 1996), which estimate binding affinity based on the features of a single protein-ligand complex structure, offer fast predictions and are suited for high throughput virtual screening and lead optimisation (Bissantz *et al.*, 2000).

Docking software, such as AutoDock 4 (Morris *et al.*, 2009), AutoDock Vina (Trott and Olson, 2010), GOLD (Verdonk *et al.*, 2003), and Glide (Friesner *et al.*, 2004) commonly use scoring functions to predict the structure of the bound ligand (the pose), its binding affinity and its rank compared to other proposed poses. These scoring functions use either

molecular force fields (Huang *et al.*, 2006), statistical potentials (Gohlke *et al.*, 2000) or linear combinations of empirical terms (Krammer *et al.*, 2005). Advancements in machine learning (ML) have enabled the development of ML-based scoring functions (MLBSFs) that appear to outperform other scoring functions in accuracy for predicting binding affinity. Initially, these scoring functions employed classical ML techniques, e.g. tree-based models, and simple features extracted from the protein-ligand complex structure (Ballester and Mitchell, 2010; Wang and Zhang, 2017; Ballester *et al.*, 2014; Li *et al.*, 2015; Zilian and Sottriffer, 2013; Durrant and McCammon, 2011; Meli *et al.*, 2021).

With the emergence of deep learning techniques, scoring functions based on the convolutional neural network (CNN) architecture were built and trained on explicit, voxelised representations of the ligand-protein complex (Francoeur *et al.*, 2020) (e.g. Pafnucy (Stepniewska-Dziubinska *et al.*, 2018) and KDeep (Jiménez *et al.*, 2018)). Newer deep learning methods such as graph neural networks (GNNs) represented atoms as nodes, bonds as edges and used message-passing to pass feature vectors across the graphs to learn higher representations for predicting binding affinity (Moon *et al.*, 2022; Karlov *et al.*, 2020; Li *et al.*,

2021; Scantlebury *et al.*, 2023; Volkov *et al.*, 2022). Despite the plethora of methods published, there is no clear consensus on which architecture should be used and how to improve scoring function accuracy, given the small differences in performance observed between the methods on the standard benchmarks (Su *et al.*, 2019; Carlson *et al.*, 2016).

There have been increasing calls for the field to rethink how we measure MLBSF performance and accuracy (Scantlebury *et al.*, 2023; Volkov *et al.*, 2022). Most MLBSFs are trained on the PDBBind database (Wang *et al.*, 2004), which consists of thousands of protein-ligand complex crystal structures with binding affinity data extracted from the literature. Complexes in CASF 2016, the most popular benchmark for scoring function performance, have very high similarity to data points within the standard training dataset (PDBBind) resulting in an over-optimistic measurement of accuracy as MLBSFs can learn data similarity or “bias” instead of relevant biophysics (Scantlebury *et al.*, 2023). Alternative methods of interrogation have been proposed, these include clustered cross-validation (Zhu *et al.*, 2022), leave-cluster-out cross-validation (Kramer and Gedeck, 2010), time-splits (Volkov *et al.*, 2022) and removing training data similar to the test data (Scantlebury *et al.*, 2023; Boyles *et al.*, 2020). Unfortunately, due to the widespread use of the CASF 2016 benchmark for evaluating models, researchers can only compare their proposed model to others using that benchmark, exacerbating the problem of inadequate scoring function evaluation. Furthermore, MLBSFs are benchmarked and tested on accurate crystal structures but often will be used for scoring predicted docked ligand poses against non-cognate or predicted structures in a real-world drug discovery setting. These noisy structures are likely to be less accurately predicted compared to the crystal structure, yet this impact has been explored in a limited manner for a few scoring functions (Boyles *et al.*, 2022; Scardino *et al.*, 2023; Wong *et al.*, 2022; Shen *et al.*, 2021; McNutt *et al.*, 2021; Francoeur *et al.*, 2020). It has been demonstrated that models trained only on ligand and/or protein identities without explicitly including the interactions between them perform surprisingly well on CASF 2016 (Boyles *et al.*, 2020; Volkov *et al.*, 2022). It can be difficult to definitively prove that models are learning bias in the dataset due to the “black box” nature of many machine learning models.

Here we present a platform for interrogating scoring function performance, called ToolBoxSF. First, we reimplemented a diverse set of MLBSFs: RFScore (Ballester and Mitchell, 2010), Pafnucy (Stepniowska-Dziubinska *et al.*, 2018), PointVS (Scantlebury *et al.*, 2023), SIGN (Li *et al.*, 2021) and OnionNet-2 (Wang *et al.*, 2021b), to use a consistent API and provide new tests and baseline models to interrogate their performance. We found that simple baseline models trained on only protein and ligand “dataset biases” and unable to learn interactions or the 3D nature of the protein-ligand complex outperformed or had competitive performance to the tested scoring functions in accuracy on a range of benchmarks. We also found that MLBSFs were still partially accurate in nonsensical scenarios like scoring very high root mean squared deviation (RMSD) error poses, ligand docked into a random, wrong protein or complexes with deliberately induced clashes. These results indicate that they are learning dataset bias instead of interpreting the 3D structure. Scoring functions pre-trained or developed for pose classification or ranking did demonstrate sensitivity to clashes though. These results prove that current evaluations of MLBSFs are inadequate at

proving usefulness in drug discovery. The provided platform and results should enable researchers to fully and robustly interrogate their models and determine the effect of dataset biases on their predictions.

Methods

Training dataset

For consistency, we trained all models on PDBBind 2020 General, the most recent release at the time of writing (Liu *et al.*, 2014). It consists of crystal structures of bound protein-ligand complexes with an associated binding affinity label (K_i , K_D or IC_{50}). We excluded complexes that could not be processed by the latest version of RDKit (2023.03.01) (Landrum, 2023) or by OpenBabel (3.1.1) (O’Boyle *et al.*, 2011). This left 19079 complexes for training and testing. Structures were prepared as described below (Docking) except the ligand coordinates were not recalculated. In this work, IC_{50} , K_i and K_D were treated as the same, a common approach in the field (Meli *et al.*, 2022) despite the values not being strictly interchangeable (Kalliokoski *et al.*, 2013). The pK for each compound was calculated by the following equation:

$$pK = -\log_{10}(K_i \text{ or } K_d \text{ or } IC_{50}) \quad (1)$$

Docking

Docking was done using Smina, a fork of AutoDock Vina (Koes *et al.*, 2013). The default parameters were chosen except “exhaustiveness” (set to 12) and “autobox_add” (set to 8Å). The protonation of the ligand and protein were kept consistent with those provided by PDBBind. The MOL2 ligand files provided in PDBBind General 2020 were converted into SDF format for consistency with the docked poses. Their 3D coordinates were recalculated using the ETKDG method from RDKit (Riniker and Landrum, 2015) before docking to ensure the docking software was not able to use the crystal pose to influence its conformational search. Protein files had water molecules and any other non-chain atoms removed by only retaining lines in the PDB file starting with ‘ATOM’.

Benchmark preparation

To generate a benchmark where ligand bias cannot be used for accurate predictions, the 0 Ligand Bias benchmark, we clustered identical molecules that were bound to different proteins by matching their InChI-Key (Pletnev *et al.*, 2012) and took clusters whose mean pK value was within 6 and 7 pK units and whose variance was larger than 1 pK unit. This left 363 complexes as a test set. These two final steps were done to remove identical ligands that had highly similar pK values and to ensure that predicting the mean of the clusters did not artificially increase the accuracy. For example, if two clusters had values concentrated around a low pK and a high pK value respectively, predicting the mean of each cluster would result in a high Pearson’s R between the predicted and true pK values. Peptides, defined as any entry in PDBBind with a ligand code with the letters ‘MER’, were held out to create the Peptides Holdout (2574 complexes). This benchmark tested the scoring functions’ ability to score peptides having never been exposed to them in the training dataset, to be accurate on this benchmark a scoring function must learn an understanding of biophysics that generalises from smaller molecules to peptides. For the 2019 Holdout set, as done in Volkov *et al.* (2022), we

took any PDBBind data point with a crystal structure produced from 2019 or later as a test set (1511 complexes). This time split was designed to create a tougher test for scoring functions compared to CASF 2016. To determine the effect of protein structure accuracy on performance, we redocked (Redocked) and crossdocked CASF 2016 ligands into protein conformations that were either bound but with high pocket similarity (CrossDocked (Best)), low pocket similarity (CrossDocked (Worst)), into apo structures (Apo), predicted AlphaFold 2 structures (Alphafold 2) and a random wrong protein (Wrong Protein). The details of how the conformations were picked can be found in the Supplementary Information.

To generate a diverse range of docking errors, we redocked the ligand back into the cognate structure of each protein-ligand complex of CASF 2016, 2019 Holdout and 0 Ligand Bias. We increased the 'autobox_add' parameter to 20Å and 'num_modes' to 1000 for Smina. To generate more poses close in accuracy to the true pose, we also minimised the crystal pose using the 'minimize' option. Poses were binned by RMSD to the crystal pose using the following ranges: 0-1Å, 1-2Å, 2-4Å, 4-6Å, 6-8Å, 8-10Å, 10-15Å, 15-20Å, 20-25Å and 25-30Å. If available, the pose closest to the mean of the bin was chosen for each test set. To explore the impact of clashes on the models, the crystal pose of the ligand for each complex from CASF 2016, 2019 Holdout and 0 Ligand Bias was progressively translated into the protein 1Å at a time, ten times. To calculate a normal vector for this translation, we normalised the direction vector between the closest ligand and closest protein atom in each protein-ligand complex. Each ligand atom was translated using this normal vector.

Implementation of scoring functions and models

To compare performance across a range of scoring functions, five popular and diverse models were selected from the literature: RFScore (Ballester and Mitchell, 2010), PointVS (Scantlebury *et al.*, 2023), Pafnucy (Stepniewska-Dziubinska *et al.*, 2018), SIGN (Li *et al.*, 2021) and OnionNet-2 (Wang *et al.*, 2021b). RFScore was one of the first methods to use machine learning to predict binding affinity, it uses Random Forest models and counts of protein and ligand elements that are within 12Å of each other. Pafnucy employs a CNN architecture and 3D voxelized representations of the protein-ligand complex. OnionNet-2 also uses a CNN with a 2D image of the counts of each specific amino acid-ligand atom interaction with differing thresholded distances. SIGN and PointVS both use graph neural networks with attention layers for the edges. PointVS is also pre-trained to classify pose accuracy within 2Å and uses this as a prior for its prediction of binding affinity. The pre-trained weights for pose classification of PointVS (Scantlebury *et al.*, 2023) were provided by the authors; the model was then fine-tuned using our training set to predict binding affinity. All differences between the original implementations and our modified implementations can be found in the Supplementary Information.

We developed three separate baseline models that represent models that can only learn "bias" in the dataset. All three were developed using tree-based models with architecture and hyperparameters chosen by the FLAML package (Wang *et al.*, 2021a) using five-fold cross-validation of the training dataset with CASF 2016 excluded. The LigandBias model is based on the RDKit model from Boyles *et al.* (Boyles *et al.*, 2020), where

Method	CASF 2016	2019 Holdout	Peptides Holdout	0 Ligand Bias
LigandBias	0.76 ± 0.06	0.59 ± 0.03	0.23 ± 0.04	0.08 ± 0.11
ProteinBias	0.75 ± 0.07	0.59 ± 0.04	0.32 ± 0.04	0.41 ± 0.10
BothBias	0.85 ± 0.03	0.68 ± 0.03	0.35 ± 0.04	0.27 ± 0.12
Smina	0.59 ± 0.08	0.36 ± 0.04	0.19 ± 0.04	0.12 ± 0.10
RFScore	<u>0.82 ± 0.04</u>	0.64 ± 0.03	0.33 ± 0.04	0.24 ± 0.10
PointVS	0.79 ± 0.04	0.66 ± 0.03	0.27 ± 0.06	0.28 ± 0.10
Pafnucy	0.74 ± 0.06	0.60 ± 0.04	0.25 ± 0.06	0.17 ± 0.11
SIGN	<u>0.82 ± 0.04</u>	0.66 ± 0.03	0.34 ± 0.04	0.27 ± 0.10
OnionNet-2	<u>0.82 ± 0.04</u>	0.70 ± 0.03	0.36 ± 0.04	0.33 ± 0.11

Table 1. Pearson's R between predicted and true pK values for protein-ligand complexes for our baseline models (LigandBias, ProteinBias and BothBias), a non-ML-based scoring function (Smina) and five commonly used MLBSFs (RFScore, PointVS, Pafnucy, SIGN and OnionNet-2) on four benchmark datasets (CASF 2016, 2019 Holdout, Peptides Holdout and 0 Ligand Bias). See methods for further details of scoring functions and dataset creation. The highest values are in bold and underlined, second highest is underlined. Error ranges represent the 95% confidence intervals from bootstrapped Pearson's R (N=10000)

1D and 2D descriptors from the RDKit package were calculated to featurise only the ligand. Any descriptor that produced NaN values or extremely large values was excluded, leaving 195 features. The ProteinBias model used counts of each amino acid within the pocket as a feature vector. We defined the protein pocket as any amino acid that had an atom within 15Å of any ligand atom. The impact of this threshold on model performance is explored in the Supplementary Information. This gives the model the identity of the amino acids but not proximity to each other or ligand atoms. Our final model, the BothBias model uses features from both the ProteinBias and LigandBias models. The full information of algorithm and hyperparameters for each baseline model are provided in the Supplementary Information. We also scored all the test sets using Smina (Koes *et al.*, 2013) as a baseline for the performance of a non-ML-based scoring function.

The data produced in this work and code for these models have been developed into an easy-to-use platform, called ToolBoxSF, to robustly compare to proposed models from the community and examine if they are learning more than bias. All models have been installed into separate Singularity containers to allow instant and easy use of the models for training or predictions. These wrappers are available on GitHub and as pre-built Singularity containers (<https://github.com/guydurant/toolboxsf>).

Metrics

Scoring function accuracy was calculated between their predicted and true values using bootstrapped Pearson R values where data points were sampled with replacement 10000 times to produce 95% confidence intervals.

Results

Existing Benchmarks

Evaluation of scoring function accuracy has typically been done using the CASF 2016 benchmark, so we first benchmarked each

model on this set (Table 1, CASF 2016). However, the similarity between training (PDBBind) and test set (CASF 2016) makes this an unsuitable benchmark for assessing MLBSF generalisability (Scantlebury *et al.*, 2023). We compared five different models that feature the protein-ligand complex differently: RFScore (Ballester and Mitchell, 2010), Pafnucy (Stepniewska-Dziubinska *et al.*, 2018), PointVS (Scantlebury *et al.*, 2023), SIGN (Li *et al.*, 2021) and OnionNet-2 (Wang *et al.*, 2021b). We retrained these scoring functions on our training sets and compared their performance against our baseline models which are unable to learn anything about the structure or interactions of the protein-ligand complex. The model trained on both protein and ligand features that contain no 3D information and so only dataset biases (“BothBias”) had the highest Pearson’s R (0.85) demonstrating that learning biophysics from 3D information is not necessary for close to state-of-the-art performance on the standard CASF 2016 benchmark (0.86) (Wang *et al.*, 2021b). High performance on this benchmark has been shown to not be indicative of generalisability (Volkov *et al.*, 2022; Scantlebury *et al.*, 2023; Zhu *et al.*, 2022) but our result goes one step further and demonstrates that even attempting to learn biophysics from structures of the protein-ligand complex provides no additional accuracy. Volkov *et al.* (2022) in an attempt to account for this bias, proposed a time-split where PDBBind data points from 2019 and later were held out as a test set (Table 1 2019 Holdout). BothBias baseline still exceeds the performance of every other model tested except OnionNet-2 with (0.68). This outcome indicates that a time-based split may not be suitable for demonstrating that a scoring function has learnt concepts of biophysics instead of dataset bias. Although both of these benchmarks have value in evaluating the accuracy of scoring functions, it is clear that other benchmarks are required to determine whether a model would be capable of generalising to novel protein or ligand families.

New Proposed Benchmarks

We propose two benchmarks which evaluate the generalisability of ML scoring functions in different ways. The first utilises the difference between the properties of peptide-protein complexes and ligand-protein complexes found within PDBBind 2020. We removed any peptide-containing complex as a hold-out set from the training dataset. Peptides are difficult to score due to their inherent flexibility and are often much larger than the other ligands in PDBBind (London *et al.*, 2010). This makes it a difficult benchmark but success would demonstrate that the models have learnt an understanding of biophysics, such as entropy and changes in solvation, that generalises to peptides. The results in Table 1 show that the BothBias model has similar performance to the best scoring function (OnionNet-2) but neither scores peptides accurately. Our second benchmark takes advantage of scoring functions tending to learn ligand-specific bias in that they are poor at differentiating between the same ligand bound to different proteins (Boyles *et al.*, 2020). We identified identical ligands within PDBBind 2020 General that had existed two or more times in the dataset and filtered to ensure these identical ligands’ mean and variance of pKs were centred but spread across the mean pK of the PDBBind dataset (i.e. the training dataset). These groups of identical ligands were then combined into a single set as the 0 Ligand Bias set. On this test set, the BothBias model (0.27) is only as accurate as SIGN (0.27) and is outperformed by OnionNet-2 (0.33) and PointVS (0.28), showing that it is possible

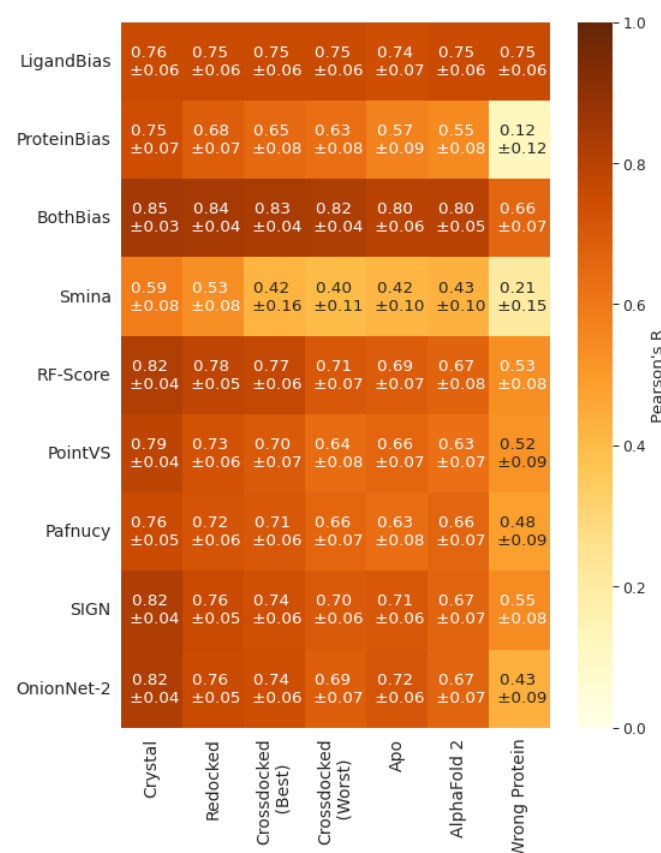


Fig. 1: Pearson’s R between predicted and true pK values for protein-ligand complexes for our baseline models (LigandBias, ProteinBias and BothBias), a non-ML-based scoring function (Smina) and five commonly used MLBSFs (RFScore, PointVS, Pafnucy, SIGN and OnionNet-2) on alternate CASF 2016 complex type test sets. Errors are the 95% confidence intervals from the bootstrapped Pearson’s R (N=10000).

to beat this baseline model. However, this is due to learning ligand bias explicitly harming its accuracy, as we can see with the poor performance of the LigandBias model (0.08). The most accurate model is the ProteinBias model (0.41) demonstrating that simply ignoring the ligand entirely out-performs other models currently on this test set. Low performance across all models tested, with Pearson’s R from 0.08 to 0.41, indicates that this is a challenging benchmark. These benchmarks demonstrate that current scoring functions are not able to significantly outperform models trained on bias. Therefore these MLBSFs are both learning bias that does not generalise to this test set and learning little or nothing further.

Effect of Protein Structure Accuracy on Performance

One deficiency in using the test sets employed above as benchmarks or held-out tests is that they only measure accuracy for scoring crystal structures. Typically scoring functions are used to score docked poses against crystal or predicted structures that might not have an accurate active site conformation for the docked ligand. This introduces noise into the structure as docking predictions may not find the specific interactions or recapture the true binding pose of the crystal structure. To explore the impact

of this noise on accuracy, we created alternate docked versions of the CASF 2016 benchmark, which is made up of five structures, each bound to a different ligand, for each of 57 types of proteins (so 285 complexes total) and so contains alternate conformations for the same protein to dock into. We produced six test sets where we re-docked the ligand back into the cognate protein structure (Figure 1 Redocked), cross-docked it into a conformation most similar to its own (Figure 1 Crossdocked (Best)) and again into a conformation most dissimilar (Figure 1 Crossdocked (Worst)). We also docked the ligand into apo (unbound) structures (Figure 1 Apo), predicted AF2 structures (Figure 1 Alphafold 2), and a random protein from CASF 2016 not from its family as a baseline (Figure 1 Wrong Pocket). The differences in structure are shown for a case study (PDB:1E66) in the Supplementary Information to display both the change in 3D structure and how this affects which interactions are formed. These increasingly noisy types of structure demonstrated decreased accuracy when scored by all scoring functions, as shown in Figure 1. The scoring functions were able to retain some accuracy even if the ligand was docked into a completely different protein demonstrating a lower bound of accuracy caused by predictions being dominated by identifying the ligand rather than the nature of the complex. The BothBias model does not appear to be affected as much by the increasing noise as its ligand features are not impacted by changes in conformation and the number of amino acids in the protein pocket does not change significantly across the complex types. These results also suggest as expected that measuring performance on crystal structures provides an upper limit of the ability of scoring functions that is unlikely to be replicated if used in a virtual screen.

Effect of Docking Accuracy on Performance

To measure the impact of docking accuracy, we considered a diverse set of poses for the CASF 2016 complexes, measured by RMSD. We binned the poses by RMSD first in small ranges (0-1Å and 1-2Å) and then increasingly larger bins with higher inaccuracy to produce 10 test sets. We tested all scoring functions and baseline models but here highlight the results of PointVS, RFScore, SIGN and OnionNet-2 (Figure 2). The results for all other methods can be found in the Supplementary Information. When high-accuracy poses were used, models retained high predictive accuracy relative to scoring the crystal structures when scored by different scoring functions. However, as docking error increased, accuracy decreased and ultimately plateaued at 10Å, except for RFScore which continued to decline beyond this point. Similar to the complex type tests, there was a lower bound for this decrease in performance even at extreme docking errors (25-30Å), where the ligand is no longer bound in the correct site, showing again the models were relying on ligand bias to score protein-ligand complexes. We also explored this effect on 2019 Holdout and 0 Ligand Bias complexes and found the same trend (Supplementary Information). This demonstrates that although docking accuracy is important for binding affinity prediction accuracy, bias is currently a more significant driver of scoring performance as there is still high accuracy for highly inaccurate poses.

Clashes

Finally, we investigated scoring function performance when there were clashes in the protein-ligand complex by creating a series of structures where the ligand was translated into the protein for

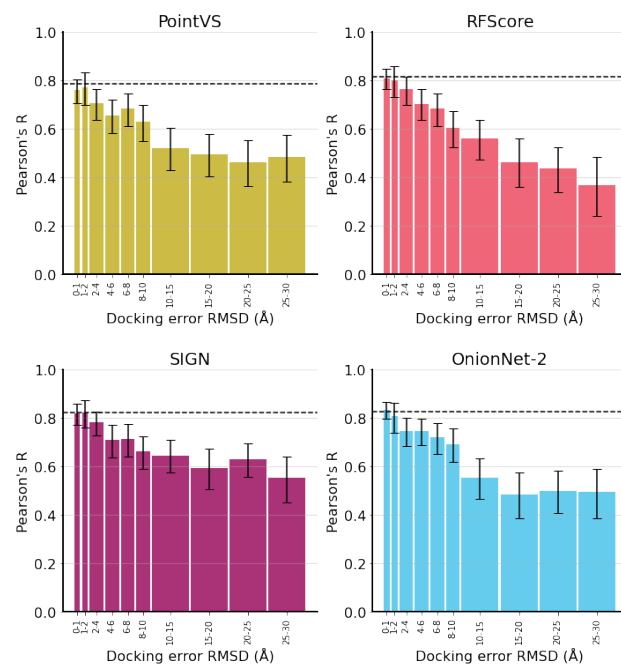


Fig. 2: Pearson's R between predicted and true pK values for protein-ligand complexes for four selected MLBSFs, PointVS, RFScore, SIGN and OnionNet-2, on different accuracy poses of CASF 2016 complexes. Accuracy on the crystal structures of CASF 2016 is shown as a dashed black line. Errors are the 95% confidence intervals from the bootstrapped Pearson's R (N=10000).

each CASF 2016 complex. Although it is unlikely these scoring functions will come across these types of structures in a drug discovery scenario, the overlap of ligand and protein structure provides such an unrealistic structure with many clashes and few interactions between the protein and ligand. Therefore it is expected that the scoring function should fail to accurately predict binding affinity. The MLBSFs displayed greater sensitivity to translation than the BothBias baseline model; however, most scoring functions displayed only a gradually decreasing performance as the clashes became increasingly severe, again indicating a lower bound (Figure 3). This indicates that the scoring functions only recognise that the ligand is further from the binding site, rather than detecting the unphysical clashes with the protein. The exceptions to these trends are Smina and PointVS, which are both co-trained or pre-trained for pose prediction and demonstrate higher sensitivity to clashes with low or no accuracy on complexes with significant clashes. Again, we also explored this effect on 2019 Holdout and 0 Ligand Bias complexes and found the same trends as for CASF 2016 (Supplementary Information). This suggests that considering pose quality in the training process provides scoring functions with the ability to discriminate between clashing, overlapping structures, and true protein-ligand complex structures.

Conclusion

In this work, we have demonstrated that state-of-the-art performance on CASF 2016 can be achieved by baseline models

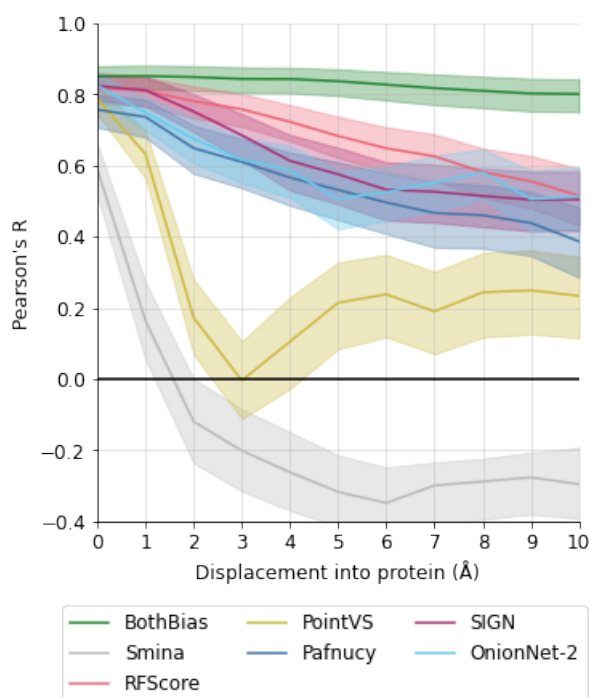


Fig. 3: Pearson's R between predicted and true pK values for protein-ligand complexes for one baseline model (BothBias), a non-ML-based scoring function (Smina) and five commonly used MLBSFs (RFScore, PointVS, Pafnucy, SIGN and OnionNet-2) on progressively displaced ligands into the protein originally from CASF 2016 crystal structures. Errors are the 95% confidence intervals from the bootstrapped Pearson's R (N=10000).

using only protein and ligand bias. We propose the 0 Ligand Bias and Peptide Holdout test sets which either explicitly penalise learning ligand bias or require a greater understanding of biophysics, as tougher benchmarks and novel thresholds for improvement from the field. Five popular MLBSFs were equalled or outperformed by baseline models in our tests, indicating that the performance of these scoring functions may be the result of learning dataset bias. We believe our baseline models offer a yardstick for the field as if any proposed scoring function can outperform them, they will have learned more than simple dataset bias.

We examined the effect of noise in the 3D structure of the protein-ligand complex on scoring function performance. The noise introduced by using inaccurate active site conformations or docked poses both resulted in degradation of accuracy in relationship to the amount of noise. This noise being either how dissimilar the active sites are to the cognate crystal structure or the RMSD difference of the pose to the crystal pose. However, both decreases in accuracy had a lower bound showing indifference to the 3D structure input and instead relying on recognising the identity of the ligand.

A further proof that these models are not necessarily learning relevant biophysics is their insensitivity to serious steric clashes between protein and ligands. Translation of the crystal pose into the surface of the protein resulted in a gradual decrease in performance indicating that the scoring functions were only able

to recognise that the ligand was further from its true location. The exceptions to this trend, PointVS and Smina, were either pre-trained or developed for pose classification or ranking respectively. These exceptions suggest that scoring functions trained to predict only binding affinity do not learn how sensible a pose is, whilst co-training for another task, such as pose classification, forces it to appreciate clashes.

Overall, this work has provided a meta-analysis of scoring functions and created baseline models that equal existing scoring function accuracy and has provided train-test splits that can help identify if proposed models have learnt more than this simple dataset bias. For the field to progress it will be necessary to design and train models in such a way that they cannot achieve apparent success simply by learning dataset biases. For researchers to prove their proposed scoring functions have learnt more than dataset bias, we have presented rigorous tests and baseline models that can be used for comparisons. All code and dataset splits can be accessed here: (<https://github.com/guydurant/toolboxsf>).

Competing interests

No competing interest is declared.

Acknowledgments

This work was supported by funding from the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/S024093/1]. G.D. thanks Dr. Carlos Outeiral for his generation of AlphaFold 2 models for this work.

References

- Ballester, P. J. and Mitchell, J. B. (2010). A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*, **26**(9), 1169–1175.
- Ballester, P. J., Schreyer, A., and Blundell, T. L. (2014). Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity? *Journal of Chemical Information and Modeling*, **54**, 944–955.
- Bissantz, C., Folkers, G., and Rognan, D. (2000). Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *Journal of medicinal chemistry*, **43**(25), 4759–4767.
- Boyles, F., Deane, C. M., and Morris, G. M. (2020). Learning from the ligand: using ligand-based features to improve binding affinity prediction. *Bioinformatics*, **36**, 758–764.
- Boyles, F., Deane, C. M., and Morris, G. M. (2022). Learning from Docked Ligands: Ligand-Based Features Rescue Structure-Based Scoring Functions When Trained on Docked Poses. *Journal of Chemical Information and Modeling*, **62**, 5329–5341.
- Carlson, H. A., Smith, R. D., Damm-Ganamet, K. L., Stuckey, J. A., Ahmed, A., Convery, M. A., Somers, D. O., Kranz, M., Elkins, P. A., Cui, G., et al. (2016). CSAR 2014: a benchmark exercise using unpublished data from pharma. *Journal of chemical information and modeling*, **56**(6), 1063–1077.
- Durrant, J. D. and McCammon, J. A. (2011). NNScore 2.0: A neural-network receptor-ligand scoring function. *Journal of Chemical Information and Modeling*, **51**, 2897–2903.
- Francoeur, P. G., Masuda, T., Sunseri, J., Jia, A., Iovanisci, R. B., Snyder, I., and Koes, D. R. (2020). Three-dimensional convolutional neural networks and a crossdocked data set for structure-based drug design. *Journal of Chemical Information and Modeling*, **60**, 4200–4215.
- 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., and Shenkin, P. S. (2004). Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1.

- Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry*, **47**, 1739–1749.
- Gohlke, H., Hendlich, M., and Klebe, G. (2000). Knowledge-based scoring function to predict protein-ligand interactions. *Journal of molecular biology*, **295**, 337–356.
- Goodsell, D. S., Morris, G. M., and Olson, A. J. (1996). Automated Docking of Flexible Ligands: Applications of AutoDock. *JOURNAL OF MOLECULAR RECOGNITION*, **9**, 1–5.
- Huang, N., Kalyanaraman, C., Jacobson, M. P., and Bernacki, K. (2006). Molecular mechanics methods for predicting protein-ligand binding. *Article in Physical Chemistry Chemical Physics*.
- Jiménez, J., Škalič, M., Martínez-Rosell, G., and Fabritiis, G. D. (2018). KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *Journal of Chemical Information and Modeling*, **58**, 287–296.
- Kalliokoski, T., Kramer, C., Vulpetti, A., and Gedeck, P. (2013). Comparability of mixed IC50 data—a statistical analysis. *PloS one*, **8**(4), e61007.
- Karlov, D. S., Sosnin, S., Fedorov, M. V., and Popov, P. (2020). GraphDelta: MPNN Scoring Function for the Affinity Prediction of Protein-Ligand Complexes. *ACS Omega*, **5**, 5150–5159.
- Koes, D. R., Baumgartner, M. P., and Camacho, C. J. (2013). Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. *Journal of Chemical Information and Modeling*, **53**(8), 1893–1904. PMID: 23379370.
- Kramer, C. and Gedeck, P. (2010). Leave-cluster-out cross-validation is appropriate for scoring functions derived from diverse protein data sets. *Journal of Chemical Information and Modeling*, **50**, 1961–1969.
- Krammer, A., Kirchhoff, P. D., Jiang, X., Venkatachalam, C. M., and Waldman, M. (2005). LigScore: a novel scoring function for predicting binding affinities. *Journal of molecular graphics modelling*, **23**, 395–407.
- Landrum, G. (2023). RDKit: Open-source cheminformatics.
- Li, H., Leung, K. S., Wong, M. H., and Ballester, P. J. (2015). Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets. *Molecular Informatics*, **34**, 115–126.
- Li, S., Zhou, J., Xu, T., Huang, L., Wang, F., Xiong, H., Huang, W., Dou, D., and Xiong, H. (2021). Structure-aware Interactive Graph Neural Networks for the Prediction of Protein-Ligand Binding Affinity; Structure-aware Interactive Graph Neural Networks for the Prediction of Protein-Ligand Binding Affinity. *KDD*, **21**.
- Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., Nie, W., Liu, Y., and Wang, R. (2014). PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, **31**(3), 405–412.
- London, N., Movshovitz-Attias, D., and Schueler-Furman, O. (2010). The Structural Basis of Peptide-Protein Binding Strategies. *Structure*, **18**(2), 188–199.
- McNutt, A. T., Francoeur, P., Aggarwal, R., Masuda, T., Meli, R., Ragoza, M., Sunseri, J., and Koes, D. R. (2021). GNINA 1.0: molecular docking with deep learning. *Journal of Cheminformatics*, **13**, 1–20.
- Meli, R., Anighoro, A., Bodkin, M. J., Morris, G. M., and Biggin, P. C. (2021). Learning protein-ligand binding affinity with atomic environment vectors. *Journal of Cheminformatics*, **13**, 1–19.
- Meli, R., Morris, G. M., and Biggin, P. C. (2022). Scoring Functions for Protein-Ligand Binding Affinity Prediction Using Structure-based Deep Learning: A Review. *Frontiers in Bioinformatics*, **2**.
- Mobley, D. L. and Gilson, M. K. (2017). Predicting Binding Free Energies: Frontiers and Benchmarks. *Annual review of biophysics*, **46**, 531–558.
- Moon, S., Zhung, W., Yang, S., Lim, J., and Kim, W. Y. (2022). PIGNet: a physics-informed deep learning model toward generalized drug–target interaction predictions. *Chemical Science*, **13**, 3661–3673.
- Morris, G. M., Ruth, H., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., and Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *Journal of computational chemistry*, **30**, 2785.
- O’Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of cheminformatics*, **3**(1), 1–14.
- Pletnev, I., Erin, A., McNaught, A., Blinov, K., Tchekhovskoi, D., and Heller, S. (2012). InChIKey collision resistance: an experimental testing. *Journal of cheminformatics*, **4**, 1–9.
- Riniker, S. and Landrum, G. A. (2015). Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *Journal of Chemical Information and Modeling*, **55**(12), 2562–2574. PMID: 26575315.
- Scantlebury, J., Vost, L., Carbery, A., Hadfield, T. E., Turnbull, O. M., Brown, N., Chenthamarakshan, V., Das, P., Grosjean, H., von Delft, F., and Deane, C. M. (2023). A Step Towards Generalisability: Training a Machine Learning Scoring Function for Structure-Based Virtual Screening. *bioRxiv*, page 2022.10.28.511712.
- Scardino, V., Filippo, J. I. D., and Cavasotto, C. N. (2023). How good are AlphaFold models for docking-based virtual screening? *iScience*, **26**, 105920.
- Shen, C., Hu, X., Gao, J., Zhang, X., Zhong, H., Wang, Z., Xu, L., Kang, Y., Cao, D., and Hou, T. (2021). The impact of cross-docked poses on performance of machine learning classifier for protein–ligand binding pose prediction. *Journal of Cheminformatics*, **13**, 1–18.
- Stepniewska-Dziubinska, M. M., Zielenkiewicz, P., and Siedlecki, P. (2018). Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, **34**, 3666.
- Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., and Wang, R. (2019). Comparative Assessment of Scoring Functions: The CASF-2016 Update. *Journal of Chemical Information and Modeling*, **59**, 895–913.
- Trott, O. and Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *Journal of computational chemistry*, **31**, 455.
- Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., and Taylor, R. D. (2003). Improved protein-ligand docking using GOLD. *Proteins*, **52**, 609–623.
- Volkov, M., Turk, J. A., Drizard, N., Martin, N., Hoffmann, B., Gaston-Mathé, Y., and Rognan, D. (2022). On the Frustration to Predict Binding Affinities from Protein-Ligand Structures with Deep Neural Networks. *Journal of Medicinal Chemistry*, **65**, 7946–7958.
- Wang, C. and Zhang, Y. (2017). Improving Scoring-Docking-Screening Powers of Protein-Ligand Scoring Functions using Random Forest. *Journal of computational chemistry*, **38**, 169.
- Wang, C., Wu, Q., Weimer, M., and Zhu, E. (2021a). FLAML: A Fast and Lightweight AutoML Library. In A. Smola, A. Dimakis, and I. Stoica, editors, *Proceedings of Machine Learning and Systems*, volume 3, pages 434–447.
- Wang, R., Fang, X., Lu, Y., and Wang, S. (2004). The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry*, **47**, 2977–2980.
- Wang, Z., Zheng, L., Liu, Y., Qu, Y., Li, Y. Q., Zhao, M., Mu, Y., and Li, W. (2021b). OnionNet-2: A Convolutional Neural Network Model for Predicting Protein-Ligand Binding Affinity Based on Residue-Atom Contacting Shells. *Frontiers in Chemistry*, **9**, 913.
- Wong, F., Krishnan, A., Zheng, E. J., Ark, H. S., Manson, A. L., Earl, A. M., Jaakkola, T., and Collins, J. J. (2022). Benchmarking AlphaFold-enabled molecular docking predictions for antibiotic discovery. *Molecular Systems Biology*, **18**, e11081.
- Zhu, H., Yang, J., and Huang, N. (2022). Assessment of the Generalization Abilities of Machine-Learning Scoring Functions for Structure-Based Virtual Screening. *Journal of Chemical Information and Modeling*, **62**, 5485–5502.
- Zilian, D. and Sottriffer, C. A. (2013). SFCscore(RF): a random forest-based scoring function for improved affinity prediction of protein-ligand complexes. *Journal of chemical information and modeling*, **53**, 1923–1933.

Supplementary Information: Robustly interrogating machine learning-based scoring functions: what are they learning?

Guy Durant, Fergus Boyles, Kristian Birchall, Brian Marsden and Charlotte M. Deane

October 26, 2023

1 Complex Type

1.1 Protein structure selection methodology

For the Redocked set, ligands were simply redocked using Smina back into their cognate structure. As CASF 2016 consists of 5 hand-picked different complex structures with different ligands bound for the same protein type, e.g. HIV protease, for 57 different protein types, there is variability within the conformation for each protein allowing cross-docking into alternative structures within the test set. To generate two classes of cross-docked structures, we aligned the pocket files, provided by PDBBind, using TM-Align (Zhang and Skolnick, 2005) in each set of 5 conformations and calculated TM-Score (Zhang and Skolnick, 2004) for each alignment. The highest TM-Score between a pocket and another pocket within the set was considered the best quality structure for cross-docking (Cross-docked (Best)), and the lowest TM-Score was chosen as the worst quality structure for cross-docking (Cross-docked (Worst)). The original ligand was docked at the site of the cognate ligand for the "best" and "worst" structures.

Apo structures for each cluster were identified by hand from the PDB. The PDB ID for the apo structure and their corresponding CASF 2016 PDB IDs are listed below. Only proteins that had no ligand bound in the active site and had 100% sequence identity for one of the five conformations in the set of the 57 in CASF 2016 were considered. As the conformations in each set in CASF 2016 are not 100% sequence identical, we could not guarantee 100% sequence identical apo structure for each structure in CASF 2016. Of the 57 sets in CASF 2016, only 46 had a suitable apo structure so the final test set was only 230 complexes in size. For each complex, we aligned the apo structure to the original structure and then docked the ligand into the pocket of the apo structure.

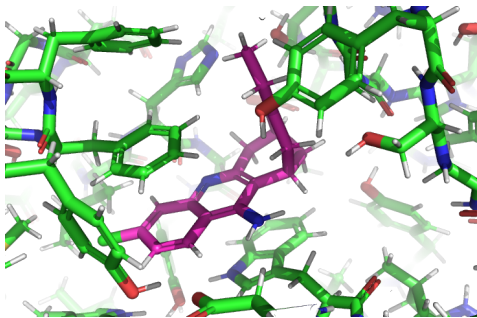
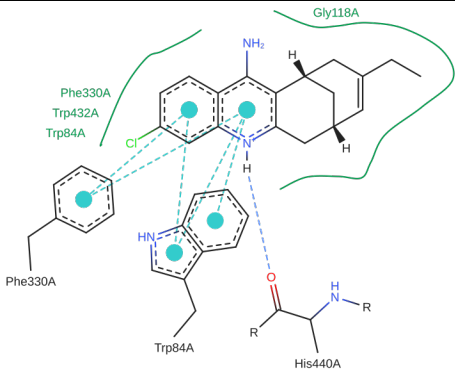
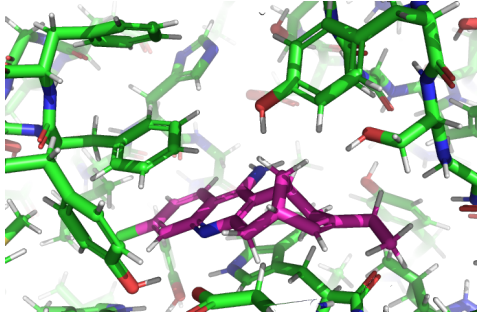
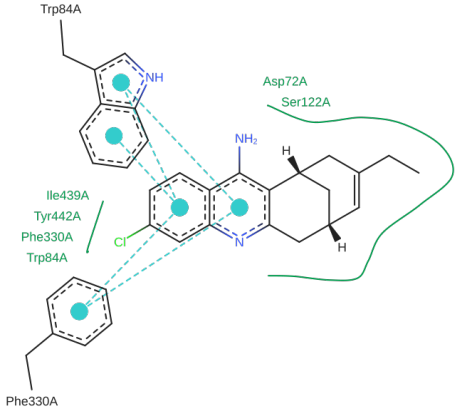
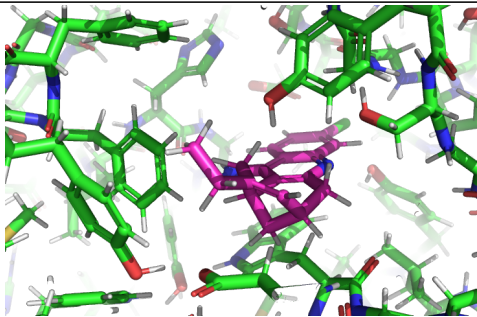
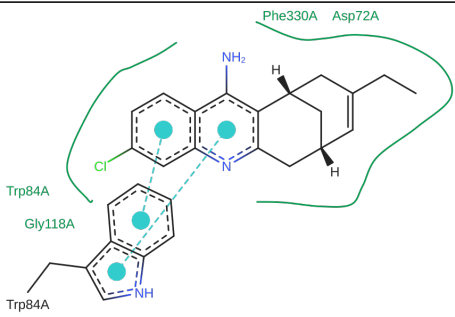
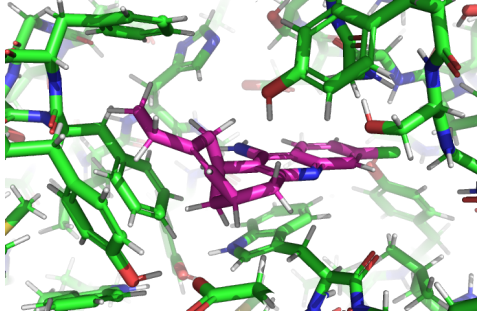
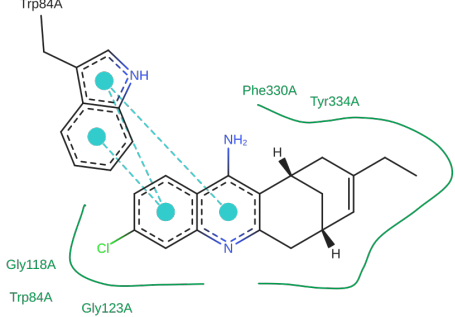
For the AlphaFold 2 version of CASF 2016, we predicted structures using AlphaFold2 (Jumper *et al.*, 2021) for monomers, and AlphaFoldMultimer v2.1 (Evans *et al.*, 2022) for proteins consisting of multiple polypeptides. Predictions were run as described in the original publication (Jumper *et al.*, 2021), including sequences from UniRef (Suzek *et al.*, 2007) as well as BFD (Jumper *et al.*, 2021) and Mgnify (Mitchell *et al.*, 2019). To emulate a realistic blind prediction scenario, we did not include any templates in the prediction, although of course a notable fraction of the targets will have been part of AlphaFold 2's training set. The ligand was then docked into the aligned AlphaFold structure. All CASF 2016 structures could be successfully predicted, except one: PDB:1YDR. Finally, for the Wrong Protein set, the ligand was docked into a randomly chosen protein not from the same set but still from the CASF 2016 set. We also include visualisations in 3D and 2D how these different complex types impact the docking pose and so the interactions formed for the PDB:1E66.

1.2 PDB IDs of CASF 2016 proteins and their respective Apo PDB ID

Holo					Apo
2xb8	3n76	3n7a	3n86	4ciw	2dhq
1nc1	1nc3	1y6r	4f2w	4f3c	1z5p
3u8k	3u8n	3zdg	4qac	3wtj	3sq9
1e66	1gpk	1gpn	1h22	1h23	7b38
2wvt	2xii	4j28	4jfs	4pcs	4j27
1ps3	3d4z	3dx1	3dx2	3ejr	3bub
1z95	3b5r	3b65	3b68	3g0w	None
3qqs	3r88	3twp	4gkm	4owm	3qr9
2fxs	2iwx	2vw5	2wer	2yge	1ah6
2cbv	2cet	2j78	2j7h	2wbg	1od0
2r9w	3gr2	3gv9	4jxs	4kz6	6t3d
3g2z	3g31	4de1	4de2	4de3	2p74
3nq9	3ueu	3uev	3uew	3uex	1b8e
2vkm	3rsx	3udh	4djv	4gid	2zhv
1kli	1o3f	1uto	3gy4	4abg	5mnz
3p5o	3u5j	4lzs	4ogj	4wiv	4lyi
3ui7	3uu0	4llx	5c28	5c2h	2oup
1q8t	1q8u	1ydr	1ydt	3ag9	None
2weg	3dd0	3kwa	3ryj	4jsz	3ks3
3nw9	3oe4	3oe5	3ozs	3ozt	4pym
1pxn	2fvd	2xnb	3pxf	4eor	4ek3
4agn	4agp	4agq	5a7b	5aba	6shz
3arp	3arq	3aru	3arv	3ary	3b8s
1lpg	1mq6	1z6e	2xbv	2y5h	1hcg
4cr9	4cra	4crc	4ty7	4x6p	None
2zcc	2zcr	2zy1	3acw	4ea2	2zco
2v00	3prs	3pww	3uri	3wz8	5rdh
1qkt	2p15	2pog	2qe4	4mgd	None

Holo					Apo
1p1n	1p1q	1syi	2a15	4u4s	1fto
1vso	3fv1	3fv2	3gbb	4dld	None
3ebp	3g2n	3l7b	3syr	4eky	3e3l
1yc1	2xdl	2yki	3b27	3rlr	5j2v
3ao4	3zso	3zsx	3zt2	4cig	None
3ehy	3lka	3nx7	3tsk	4gr0	1os9
2zb1	3e92	3e93	4dli	4f9w	4e5b
2vvn	2w4x	2w66	2wca	2xj7	4ais
3coy	3coz	3ivg	4ddh	4ddk	3cov
2p4y	2yfe	3b1m	3fur	3u9q	6l8b
1a30	1eby	2qnq	3o9i	1g2k	3phv
1r5y	1s38	3gc5	3ge7	3rr4	4pun
1o0h	1u1b	1w4o	3d6q	3dxg	6etk
3cj4	3gnw	4eo8	4ih5	4ih7	1nb4
2wtv	3e5a	3myg	3uo4	3up2	6cpe
1nvq	2br1	2brb	3jvr	3jvs	1ia8
2c3i	3bgz	3jya	4k18	5dwr	1xqz
2wn9	2wnc	2x00	2xys	2ymd	2byn
3kr8	4j21	4j3l	4kzq	4kzu	3kr7
1qf1	1z9g	3fcq	4tmn	5tmn	2g4z
1bcu	1oyt	2zda	3bv9	3utu	4nzq
4bkt	4w9c	4w9h	4w9i	4w9l	3zrf
3f3a	3f3c	3f3d	3f3e	4mme	5jae
2v7a	3k5v	3mss	3pyy	4twp	None
3qgy	4m0y	4m0z	4qd6	4rfm	None
4e5w	4ivb	4ivc	4ivd	4k77	None
4e6q	4f09	4gfm	4hge	4jia	None
1bzc	2hb1	2qbp	2qbr	2qbr	5k9v
1c5z	1o5b	1owh	1sqa	3kgp	4dw2

1.3 Visualisation of different complex types for PDB:1E66 and their interactions

Complex Type	3D Structure	Interactions
Crystal		
Redocked		
Crossdocked (Best)		
Crossdocked (Worst)		

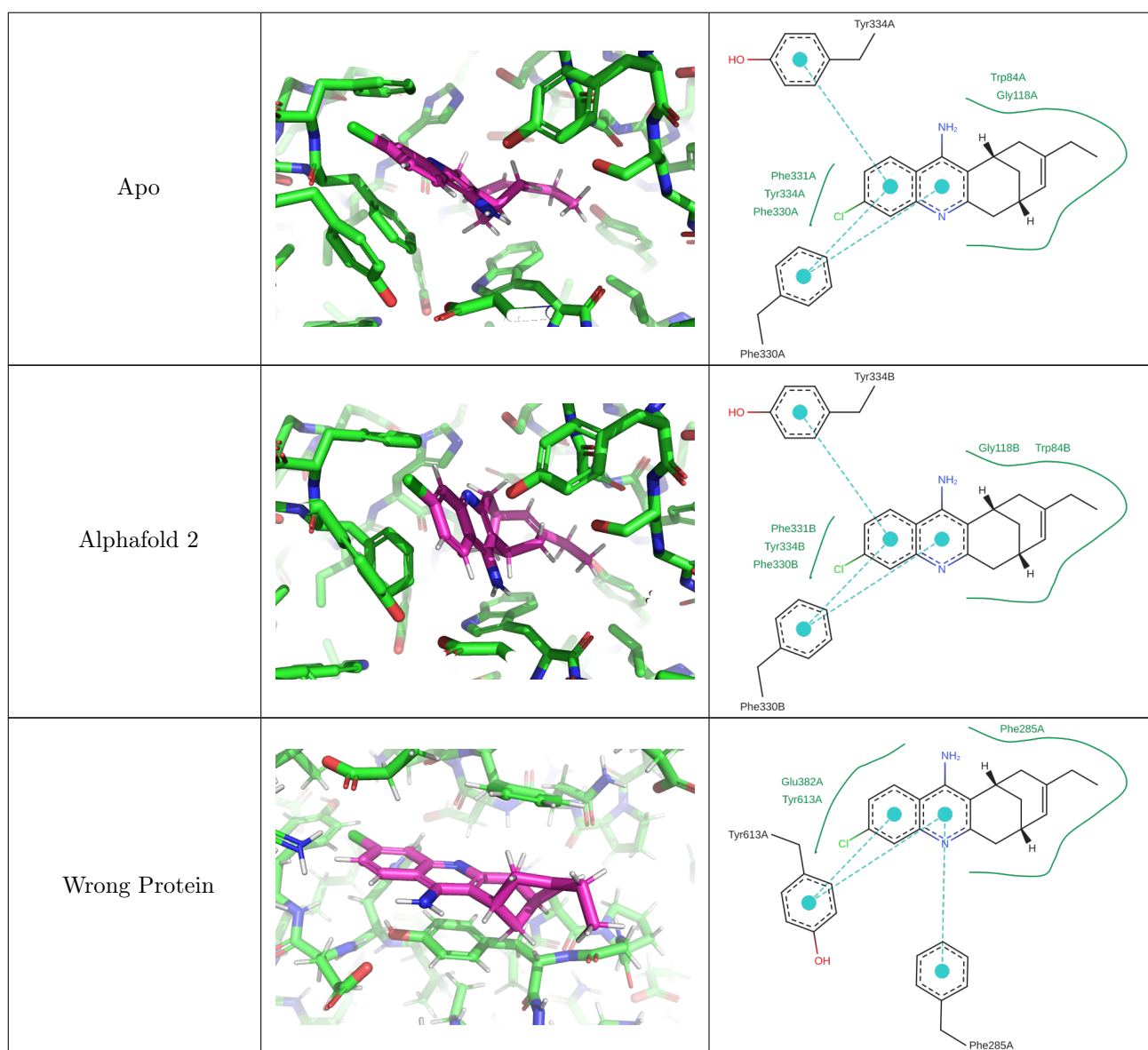


Table 3: 3D and 2D depictions of the protein-ligand complex PDB:1E66 for the different complex types of the Complex Type test sets. The 3D visualisations are at a fixed, consistent orientation and were made using PyMOL [Schrödinger, LLC \(2015\)](#). The 2D depiction of the ligand and the interactions with side chains were generated using PoseEdit [Stierand and Rarey \(2010\)](#) These demonstrate the effect of docking into increasingly noisy protein structures on complex accuracy and the interactions.

2 Model implementation

2.1 Machine learning-based scoring function (MLBSF) modified implementations

MLBSF	Modification
RFScore (Ballester and Mitchell, 2010)	We reimplemented it with RandomForestRegressor, using the original hyperparameters, from the SciKit package (Pedregosa <i>et al.</i> , 2011) with features taken from the Open Drug Discovery Toolkit (ODDT) (Wójcikowski <i>et al.</i> , 2015) package. Although the hyperparameters could not be perfectly matched due to differences in the Random Forest implementations, it is unlikely this affected performance.
PointVS (Scantlebury <i>et al.</i> , 2023)	Required a validation set for early stopping; we used a random sample of 1000 data points from the training set.
Pafnucy (Stepniewska-Dziubinska <i>et al.</i> , 2018)	The original implementation used charges assigned by ChimeraX (Goddard <i>et al.</i> , 2018), however, due to its complexity as a dependency, we chose to set all charges to zero.
SIGN (Li <i>et al.</i> , 2021)	Like Pafnucy, also uses charges for the protein pocket from ChimeraX, so also had these features set to zero. It also required a validation set for early stopping; we used a random sample of 1000 data points from the training set.
OnionNet-2 (Wang <i>et al.</i> , 2021)	In the original implementation, they used a loss function that combined Pearson's R and root mean squared error (RMSE), PCC-RMSE. However, we found that this loss function often produced NaN losses during training so we used the RMSE loss function provided in their codebase instead. It also required a validation set for early stopping; we used a random sample of 1000 data points from the training set.

2.2 Hyperparameters of baseline models

LigandBias	ProteinBias	BothBias
<u>Random Forest Regressor</u> <ul style="list-style-type: none"> • <code>n_estimators</code> = 362 • <code>max_features</code> = 0.51 • <code>max_leaf_nodes</code> = 2038 	<u>Random Forest Regressor</u> <ul style="list-style-type: none"> • <code>n_estimators</code> = 1447 • <code>max_features</code> = 0.32 • <code>max_leaf_nodes</code> = 5460 	<u>LGBMRegressor</u> <ul style="list-style-type: none"> • <code>n_estimators</code> = 205 • <code>num_leaves</code> = 291 • <code>min_child_samples</code> = 2 • <code>learning_rate</code> = 0.03 • <code>log_max_bin</code> = 9 • <code>colsample_bytree</code> = 0.6 • <code>reg_alpha</code> = 0.01 • <code>reg_lambda</code> = 0.01

2.3 Impact of protein pocket distance cutoff on accuracy for baseline models

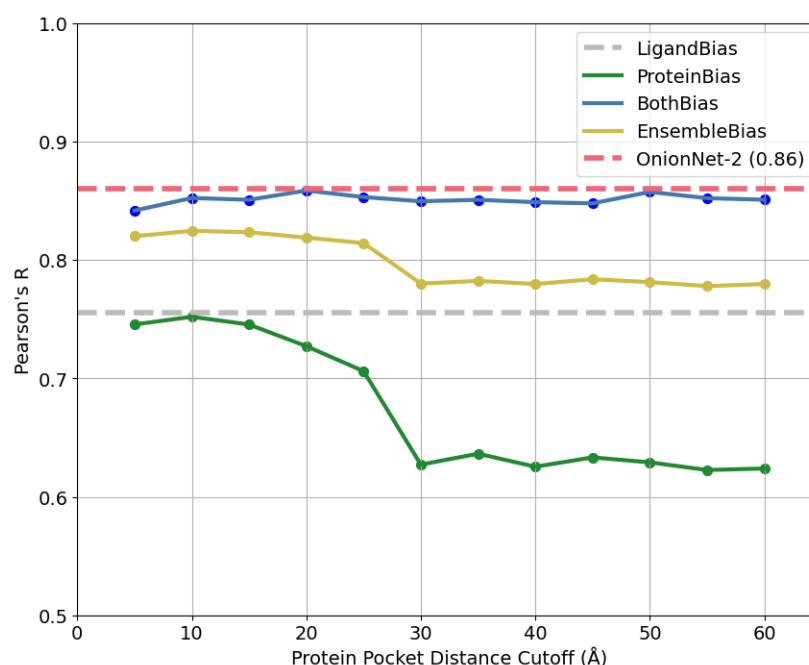


Figure 1: Accuracy of baseline models on the CASF 2016 benchmark when changing the distance threshold, from any protein atom to any ligand atom, to include residues as being part of the protein pocket for featurisation. State-of-the-art performance is included to demonstrate how close the BothBias baseline model is getting to matching it (Pearson's $R = 0.86$, OnionNet-2). The EnsembleBias model is the mean prediction of the LigandBias and ProteinBias models. The LigandBias model does not change with distance as it does not use any features derived from the protein of the complex.

3 Accuracy of scoring functions and baseline models on protein family hold-out clusters

All the benchmarks investigated in the paper do not reflect a realistic drug discovery scenario as they measure accuracy across many different protein families at once instead of the typical process of screening against a single protein target. We also created the Protein Family Out Benchmarks test sets to measure scoring function accuracy on specific protein families. We clustered the PDBBind dataset, using 90% sequence identity clusters from the PDB (RCBS, 2023) and took any cluster that had more than 100 data points as separate test sets to simulate the screening of a single protein target. The 100 data point size limit ensured there were sufficient data points to evaluate scoring function accuracy. We then trained a new version of each scoring function without each cluster and tested them on the held-out cluster to simulate the scenario of these scoring functions being used on a novel protein target in a virtual screen.

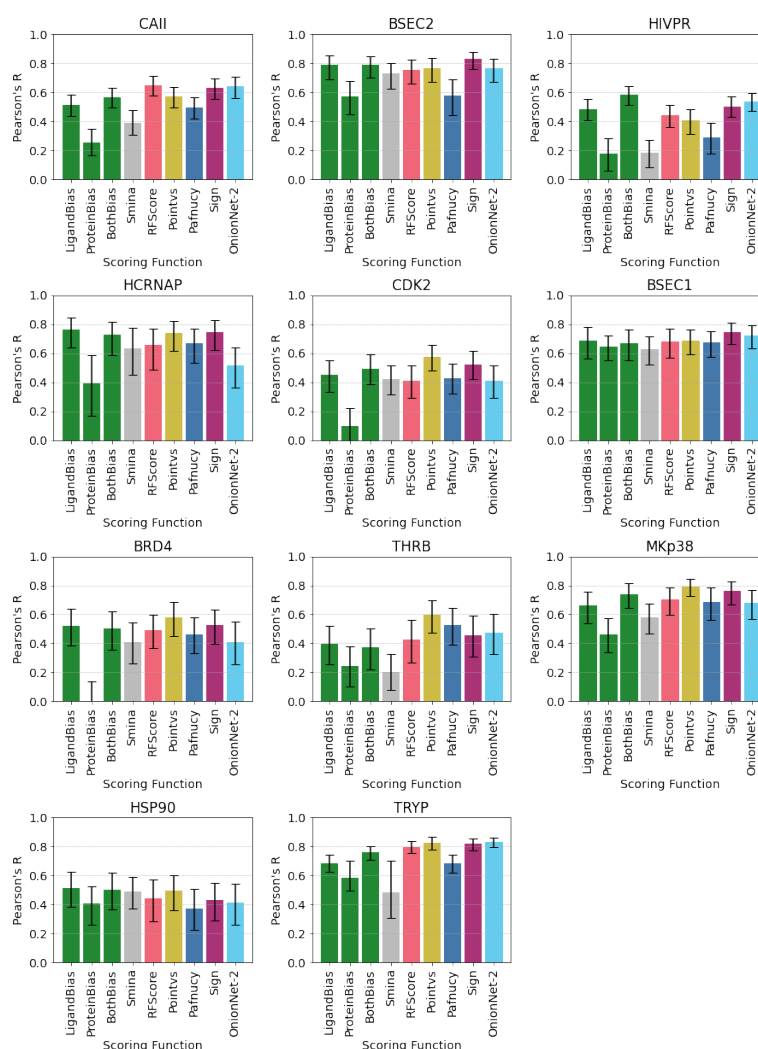


Figure 2: Pearson's R between predicted and true pK values for protein-ligand complexes for our baseline models (Ligand Bias, Protein Bias and Both Bias), a non-ML-based scoring function (Smina) and five commonly used MLBSFs (RFScore, PointVS, Pafnucy, SIGN and OnionNet-2) for eleven protein family hold-out clusters. These eleven families are Carbonic Anhydrase II (CAII), Beta-secretase (BSEC2), HIV protease (HIVPR), Hepatitis C Virus RNA-polymerase (HCRNAP), Cyclin-dependent kinase 2 (CKD2), Beta-secretase (BSEC1), Bromodomain-containing protein 4 (BRD4), Thrombin (THRB), MAP Kinase p28 (MKp38), Heat Shock Protein 90 (HSP90) and Trypsin (TRYP). Error bars represent the 95% confidence intervals from bootstrapped Pearson's R (N=10000).

These hold-out tests demonstrate the overall trend that the scoring functions do not vary greatly in the accuracy they achieve, with larger differences in the average ability between protein families than between model accuracy on the same family. The baseline models also follow this trend suggesting that the reason for these differences is probably due to the protein families having different similarities to the training dataset rather than any deeper insight into the scoring functions' performance. However, some scoring functions can outperform the baseline models in certain families such as CAII, where RFScore and OnionNet-2 (0.64) outperform the BothBias model (0.57). Due to the small differences between the baseline models and the scoring functions and their consistent performance across methods, it is likely that a large amount of accuracy for all methods is due to learning the same bias in the dataset.

4 Further scoring functions and baseline models on differing docking accuracy versions of CASF 2016

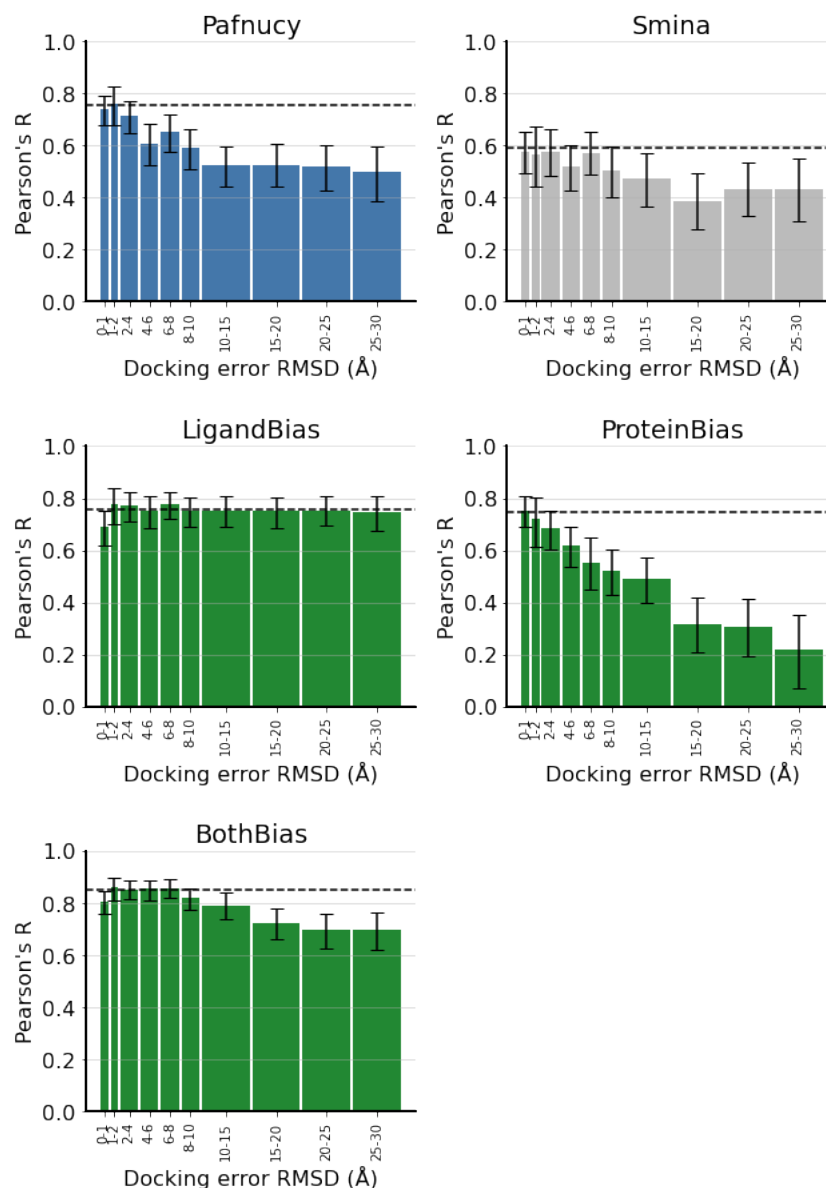


Figure 3: Pearson's R between predicted and true pK values for protein-ligand complexes for our baseline models (LigandBias, ProteinBias and BothBias), Pafnucy and Smina on different accuracy poses of CASF 2016 complexes. Accuracy on the crystal structures of CASF 2016 is shown as a dashed black line. Errors are the 95% confidence intervals from the bootstrapped Pearson's R (N=10000).

5 Accuracy of scoring functions and baseline models on differing docking accuracy versions of 2019 Holdout

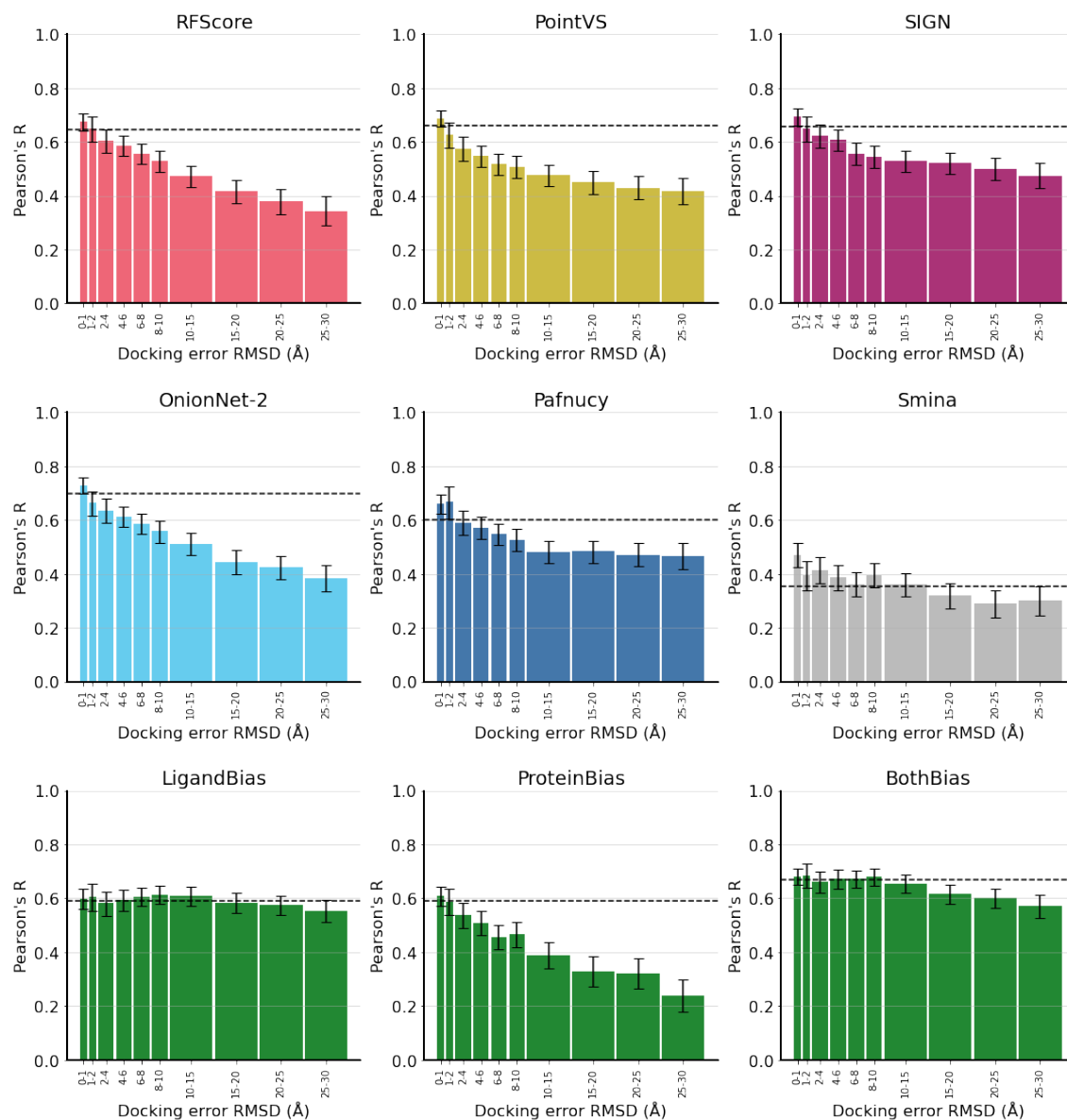


Figure 4: Pearson's R between predicted and true pK values for protein-ligand complexes for our baseline models (LigandBias, ProteinBias and BothBias), a non-ML-based scoring function (Smina) and five commonly used MLBSFs (RFScore, PointVS, Pafnucy, SIGN and OnionNet-2), on different accuracy poses of 2019 Holdout complexes. Accuracy on the crystal structures of 2019 Holdout is shown as a dashed black line. Errors are the 95% confidence intervals from the bootstrapped Pearson's R (N=10000).

6 Accuracy of scoring functions and baseline models on differing docking accuracy versions of 0 Ligand Bias

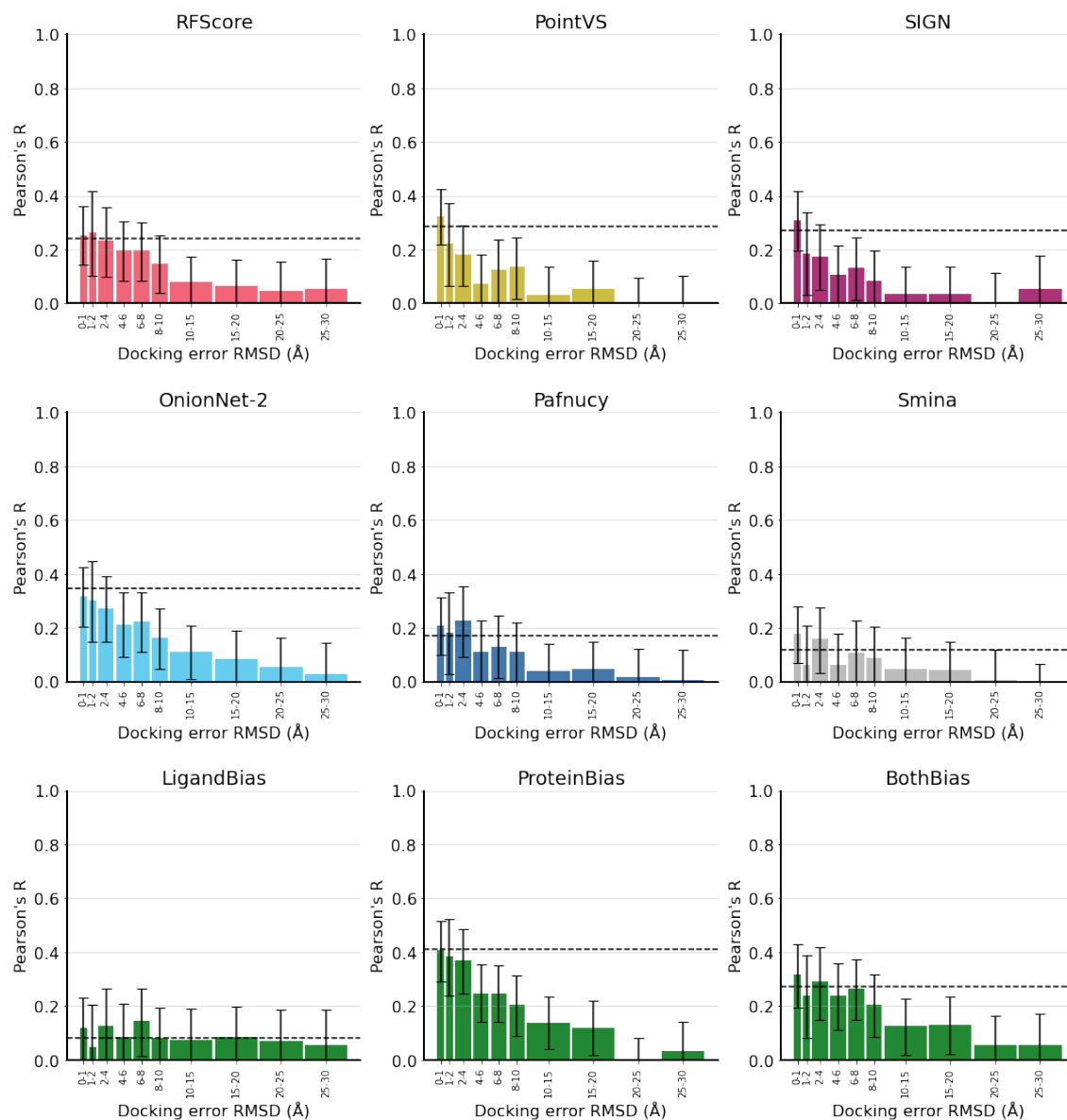


Figure 5: Pearson's R between predicted and true pK values for protein-ligand complexes for our baseline models (LigandBias, ProteinBias and BothBias), a non-ML-based scoring function (Smina) and five commonly used MLBSFs (RFScore, PointVS, Pafnucy, SIGN and OnionNet-2), on different accuracy poses of 0 Ligand Bias complexes. Accuracy on the crystal structures of 0 Ligand Bias is shown as a dashed black line. Errors are the 95% confidence intervals from the bootstrapped Pearson's R (N=10000).

7 Accuracy of scoring functions and baseline models on progressively displaced ligands of 2019 Holdout

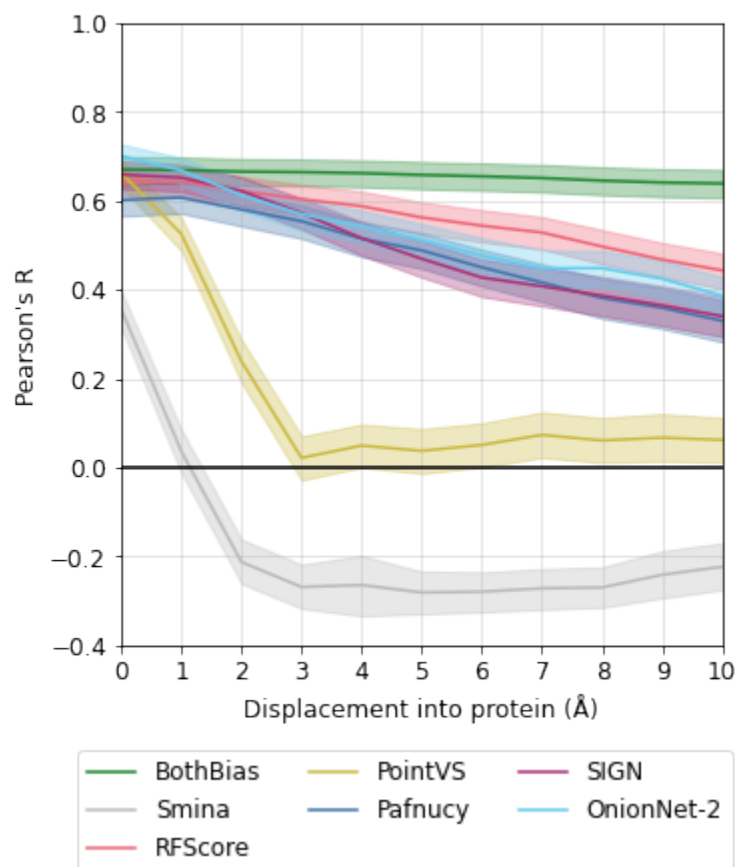


Figure 6: Pearson's R between predicted and true pK values for protein-ligand complexes for our baseline models (LigandBias, ProteinBias and BothBias), a non-ML-based scoring function (Smina) and five commonly used MLBSFs (RFScore, PointVS, Pafnucy, SIGN and OnionNet-2), on progressively displaced ligands into the protein originally from 2019 Holdout crystal structures. Errors are the 95% confidence intervals from the bootstrapped Pearson's R (N=10000).

8 Accuracy of scoring functions and baseline models on progressively displaced ligands of 0 Ligand Bias

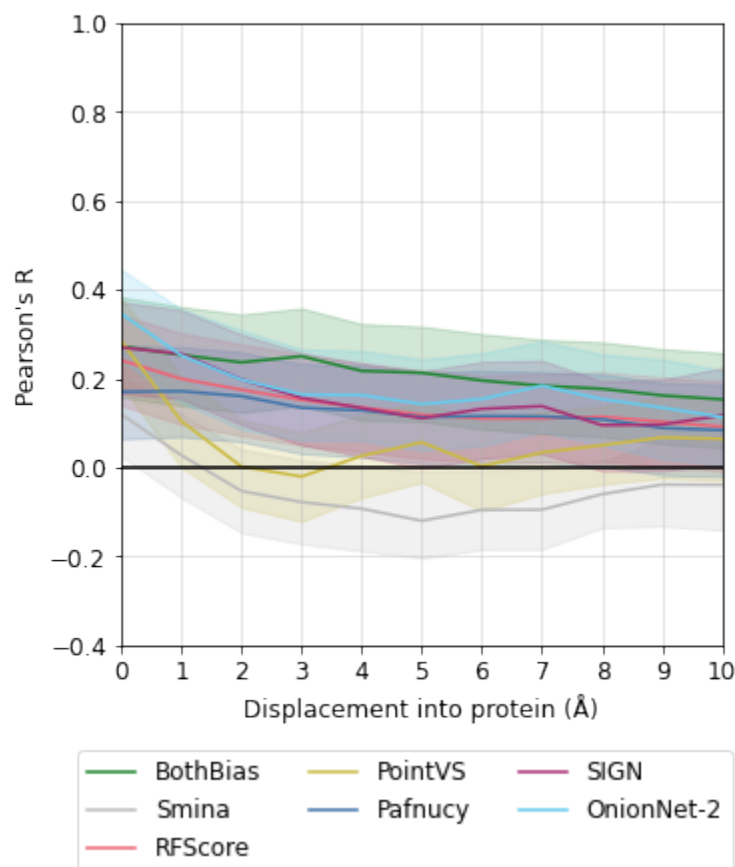


Figure 7: Pearson's R between predicted and true pK values for protein-ligand complexes for our baseline models (LigandBias, ProteinBias and BothBias), a non-ML-based scoring function (Smina) and five commonly used MLBSFs (RFScore, PointVS, Pafnucy, SIGN and OnionNet-2), on progressively displaced ligands into the protein originally from 0 Ligand Bias crystal structures. Errors are the 95% confidence intervals from the bootstrapped Pearson's R (N=10000).

References

- Ballester, P. J. and Mitchell, J. B. (2010). A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*, **26**(9), 1169–1175.
- Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J., and Hassabis, D. (2022). Protein complex prediction with AlphaFold-Multimer.
- Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H., and Ferrin, T. E. (2018). UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Science*, **27**(1), 14–25.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., Hassabis, D., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583.
- Li, S., Zhou, J., Xu, T., Huang, L., Wang, F., Xiong, H., Huang, W., Dou, D., and Xiong, H. (2021). Structure-aware Interactive Graph Neural Networks for the Prediction of Protein-Ligand Binding Affinity; Structure-aware Interactive Graph Neural Networks for the Prediction of Protein-Ligand Binding Affinity. *KDD*, **21**.
- Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M. R., Kale, V., Potter, S. C., Richardson, L. J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A., and Finn, R. D. (2019). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research*, **48**(D1), D570–D578.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, **12**(Oct), 2825–2830.
- RCBS (2023). RCSB PDB: Programmatic Access - File Download Services - Sequence Clusters Data. Accessed on June 28, 2023.
- Scantlebury, J., Vost, L., Carbery, A., Hadfield, T. E., Turnbull, O. M., Brown, N., Chenthamarakshan, V., Das, P., Grosjean, H., von Delft, F., and Deane, C. M. (2023). A Step Towards Generalisability: Training a Machine Learning Scoring Function for Structure-Based Virtual Screening. *bioRxiv*, page 2022.10.28.511712.
- Schrödinger, LLC (2015). The PyMOL molecular graphics system, version 1.8.
- Stepniewska-Dziubinska, M. M., Zielenkiewicz, P., and Siedlecki, P. (2018). Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, **34**, 3666.
- Stierand, K. and Rarey, M. (2010). Drawing the PDB: ProteinLigand Complexes in Two Dimensions. *ACS Medicinal Chemistry Letters*, **1**(9), 540–545.
- Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**(10), 1282–1288.
- Wang, Z., Zheng, L., Liu, Y., Qu, Y., Li, Y. Q., Zhao, M., Mu, Y., and Li, W. (2021). OnionNet-2: A Convolutional Neural Network Model for Predicting Protein-Ligand Binding Affinity Based on Residue-Atom Contacting Shells. *Frontiers in Chemistry*, **9**, 913.
- Wójcikowski, M., Zielenkiewicz, P., and Siedlecki, P. (2015). Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. *Journal of cheminformatics*, **7**(1), 1–6.
- Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, **57**(4), 702–710.
- Zhang, Y. and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, **33**(7), 2302–2309.