Learning from Docked Ligands: Ligand-Based Features Rescue Structure-Based Scoring Functions When Trained On Docked Poses

Journal:	*Journal of Chemical Information and Modeling*
Manuscript ID	ci-2021-000965.R2
Manuscript Type:	Article
Date Submitted by the Author:	20-Jun-2021
Complete List of Authors:	Boyles, Fergus; University of Oxford, Department of Statistics Deane, Charlotte; University of Oxford, Departament of Statistics Morris, Garrett; University of Oxford, Oxford Protein Informatics Group, Department of Statistics

Learning from Docked Ligands: Ligand-Based Features Rescue Structure-Based Scoring Functions When Trained On Docked Poses

Fergus Boyles, Charlotte M. Deane, and Garrett M. Morris*

Department of Statistics, University of Oxford, 24-29 St Giles', Oxford, OX1 3LB, U.K.

E-mail: morris@stats.ox.ac.uk

Abstract

Machine learning scoring functions for protein-ligand binding affinity have been found to consistently outperform classical scoring functions when trained and tested on crystal structures of bound protein-ligand complexes. However, it is less clear how these methods perform when applied to docked poses of complexes.

We explore how the use of docked, rather than crystallographic, poses for both training and testing affects the performance of machine learning scoring functions. Using the PDBbind Core Sets as benchmarks, we show that the performance of a structure-based machine learning scoring function trained and tested on docked poses is lower than that of the same scoring function trained and tested on crystallographic poses. We construct a hybrid scoring function by combining both structure-based and ligand-based features, and show that its ability to predict binding affinity using docked poses is comparable to that of purely structure-based scoring functions trained and tested on crystal poses. We also present a new, freely available validation set for binding affinity prediction using data from DUD-E and ChEMBL. Despite strong performance on docked poses of the PDBbind Core Sets, we find that our hybrid

scoring function sometimes generalises poorly to a protein target not represented in the training set, demonstrating the need for improved scoring functions and additional validation benchmarks.

Introduction

Rapidly prioritising which compounds to make is a key question in early-stage drug discovery[1]. When the structure of the target is known, one commonly-used approach is high-throughput protein-ligand docking,[2,3] which uses scoring functions to rank compounds by their predicted affinity for the target protein. Such scoring functions must make several assumptions that trade biophysical accuracy for speed, and rely predominantly on structure-based features to quantify the various protein-ligand interactions. It is common practice to use X-ray crystallographic structures of bound protein-ligand complexes when training and testing these models to predict protein-ligand binding affinity. This is reasonable, as it isolates the task of binding affinity prediction from structural errors that might be introduced as a result of inaccurate or incorrect ligand pose prediction, or due to the rigid receptor assumption that is often used in docking.[4–6] However, in a real-world drug discovery scenario, it is highly unlikely that a crystal structure of the bound protein-ligand complex will be available for every ligand or class of ligands of interest. Instead, protein-ligand docking is often used to predict the binding mode of each ligand within the binding pocket of a protein target[2,3]. Thus, for scoring functions to be relied upon in prospective screens, they will need to show good predictive performance not only on experimentally-determined binding poses, but also on docked poses. Furthermore, if reliable scoring functions can be found when training on docked ligands, this will greatly expand their applicability and utility in early-stage drug discovery when structural data are less likely to be available.

Although less common than the use of crystal structures, there are a few reported studies of the effect of using docked poses in training and testing scoring functions. Durrant and McCammon[7] used both crystallographic and docked poses in the development and evaluation

of their scoring function, NNScore 2.0, and noted that the optimal choice of docking protocol was highly system-dependent, but did not examine the difference in performance between training using only crystallographic or only docked poses of the same complexes. Zilian and Sotriffer[8] trained $\mathrm{SFCScore_{RF}}$ using crystal poses from the PDBbind database,[9,10] and validated on a combination of crystal poses from the PDBbind database and docked poses from the CSAR–NRC HiQ[11] and the CSAR 2012 benchmarks.[12] $\mathrm{SFCScore_{RF}}$ demonstrated strong performance on crystallographic poses of the PDBbind 2007 Core Set, with a Pearson correlation coefficient between the predicted and experimental binding affinity of 0.779; and on docked poses of the CSAR–NRC HiQ test set, with a Pearson correlation coefficient between predicted and experimental binding affinity of 0.730. Performance on the CSAR 2012 benchmark was highly target-dependent, with the authors attributing poor performance on the kinase targets CHK1 and ERK2 to the poor quality of docked poses. More recently, Jimenez et al.[13] validated $\mathrm{K_{DEEP}}$ using both crystal poses and docked poses but, as in the earlier studies, used docked poses only when crystal structures were unavailable, so the impact of the use of docked poses in place of crystal structures was not determined. A recent study by Morrone et al.[14] combined ligand-based and structure-based graph representations in neural networks for virtual screening, and highlighted the importance of docking pose prediction and the challenges imposed by cross-docking, both of which are important considerations in real-world applications where protein-ligand co-crystal complexes are unavailable. Li et al.[15] investigated how the use of docked poses in place of crystal poses affected the performance of the AutoDock Vina scoring function, RF-Score, and RF-Score v3. They reported that using docked poses in place of crystal poses negatively affected the accuracy of their binding affinity predictions; for example, the Spearman rank correlation coefficient between predicted and experimental binding affinity attained by RF-Score v3 on the PDBbind 2013 Core Set dropped from 0.662 (crystal poses) to 0.633 (docked poses). The authors also reported that this drop in performance was reduced by training RF-Score v3 on docked poses instead of crystal poses, resulting in a Spearman rank correlation coefficient between predicted and ex-

perimental binding affinity of 0.643 on the PDBbind 2013 Core Set, suggesting that training on docked poses can make a scoring function less susceptible to errors when subsequently tested on docked poses.

Here, we describe a new approach to compensate for potential errors introduced into protein-ligand binding affinity prediction through the use of docked (instead of crystallographic) poses. Previously,[16] we reported that the inclusion of pose-independent ligand-based features improved the performance of several machine-learning scoring functions in predicting protein-ligand binding affinity when trained on crystallographic binding modes of the PDBbind database,[9] and tested on the Comparative Assessment of Scoring Functions (CASF) benchmarks.[4–6] We also found that ligand-based features are predictive of the average binding affinity of a ligand for its protein targets in PDBbind[16]. Several other studies have also shown that ligand-based features contribute to improved virtual screening and binding affinity prediction[17–19], however these studies do not directly assess how the contribution of the ligand-based features differs between the use of crystal poses or docked poses. It has also been shown in several recent studies[20–22] that protein-ligand binding affinity may be predicted more accurately when it is possible to exploit some known ligands of a target during model training. This suggests that ligand-based information plays an important role in binding affinity prediction, but also highlights the importance of assessing how model performance is dependent on the similarity between data in training and test sets. Motivated by these studies, and the fact that ligand-based features are not affected by errors introduced by the use of docked poses, we investigated how the inclusion of a rich set of ligand-based features as inputs to a machine learning scoring function affects its performance when trained and tested on docked poses. Firstly, we investigated the effect of training and testing scoring functions using only docked poses generated by Smina,[23] a fork of AutoDock Vina,[24] as opposed to X-ray crystallographic poses, on the resulting scoring functions. We found that pose prediction errors are common when re-docking the ligands in the PDBbind database into their corresponding protein co-crystal structures, and that even when accu-

rate poses were generated, the Smina scoring function often failed to rank these native-like poses higher than non-native ones. We also found that training and testing machine learning scoring functions on docked poses degrades scoring function performance relative to those trained and tested on crystallographic poses even when there is a high level of similarity between proteins in the training and test sets. Augmenting the input features with rapidly computed ligand-based molecular descriptors results in a greater gain in performance when using docked poses, than when using crystal poses for training and testing.

Finally, we construct a new data set consisting of six of the eight protein targets from the DUD-E Diverse Subset,[25] for which ligand binding affinity data were available in the ChEMBL database, version 25,[26]. We find that our models perform well when trained and tested on docked poses of ligands for a single target; but poorly when trained and tested on docked poses for different targets. While there is clearly some way to go in order to create truly generalised machine learning scoring functions, our results suggest that the inclusion of additional ligand-based input features in a scoring function helps to correct for potential errors introduced by docking. This thus greatly expands the set of targets and ligands that can be used to train new scoring functions, and will help to generate better predictions of protein-ligand binding affinities.

Methods

Data

PDBbind Training Set

We used the PDBbind 2018 Refined Set[10] of high-quality protein-ligand structures and binding affinities as our source of training data. This consists of 4,463 crystal structures of protein-ligand complexes from the PDB[27] with experimentally-determined binding affinity values. For our training set we selected all the structures in the PDBbind 2018 Refined

Set, but excluded 481 structures for which either the docking failed or features could not be computed; we also excluded any structures that were also present in the PDBbind Test Set (described below). This resulted in a training set of 3,752 high quality protein-ligand complexes with corresponding binding affinity data, which we refer to as the "PDBbind Training Set".

We used the PDBbind 2018 Refined Set rather than the larger PDBbind 2018 General Set[10] for two reasons. Firstly, structures in the Refined Set are subject to strict quality controls, including a resolution of 2.5Å or better and no missing side chains, reducing the risk of errors in the docking results that might be introduced by the use of inaccurate crystal structures. Secondly, while several authors have reported improved performance on the PDBbind Core Sets when training on the larger General Set instead of the Refined Set,[28,29] we observed previously[16] that this can be attributed to increased representation of the Core Set proteins in the training set: when proteins with high sequence identity to those in the test set were excluded from the training set, there was little difference in performance between training on the General Set and training on the Refined Set.[16]

PDBbind Test Set

To evaluate the performance of our scoring functions, we constructed a test set by combining the structures in the PDBbind 2007, 2013, and 2016 Core Sets, which correspond to the test sets used in the Comparative Assessment of Scoring Functions (CASF) 2009, 2013, and 2016, respectively.[4–6] Combining these three PDBbind Core Sets, and removing 30 structures for which features could not be computed due to RDKit version 2019.09.1[30] failing to sanitise the ligand, resulted in a test set of 525 protein-ligand complexes, which we refer to as the "PDBbind Test Set". Previously, we found that using this combined test set of unique protein-ligand complexes in place of the smaller Core Sets resulted in a better (i.e. narrower) confidence interval for the performance metric.[16]

Updated DUD-E Diverse Subset

We constructed an Updated DUD-E[25] Diverse Subset for six of the eight diverse protein targets for which K_i binding affinity data could be found in version 25 of the ChEMBL database.[26]. These targets were: (1) serine/threonine-protein kinase AKT (AKT1); (2) cytochrome P450 3A4 (CP3A4); (3) glucocorticoid receptor (GCR); (4) HIV-1 protease (HIVPR); (5) HIV-1 reverse transcriptase (HIVRT); and (6) kinesin-like protein 1 (KIF11). Two of the eight targets in the DUD-E Diverse Set, β-lactamase (AMPC) and C-X-C chemokine receptor type 4 (CXCR4), which were the smallest sets in the original DUD-E Diverse Subset, did not have any ligands with recorded K_i measurements in ChEMBL at the time of writing, and so were excluded.

For each target, we queried ChEMBL version 25 for ligands that bind to that target and which had one or more measurement of K_i. We used only K_i data and excluded measurements such as IC_{50} so as not to conflate different types of data. We trained our models using the corresponding pChEMBL value, which is the negative base-10 logarithm of the binding constant reported by ChEMBL, equivalent to the pK value.[31] Duplicate pChEMBL values for a protein-ligand pair were removed; for ligands with different pChEMBL values for the same target, we used the arithmetic mean of the pChEMBL value. The DUD-E abbreviations for the names of the six targets for which K_i data were available in ChEMBL, together with the PDB code of the structure provided by DUD-E and the number of ligands for each target, are shown in Supporting Information Table 1.

The majority of the data for HIVRT and CP3A4 span only four orders of magnitude, with pChEMBL values ranging from 4 to 8. The data for GCR cover a slightly larger range, with pChEMBL values ranging from 4.7 to 10, while the data for AKT1, HIVPT, and KIF11 span at least six orders of magnitude. The binding affinities all lie within the range of values represented in the PDBbind 2018 Refined Set, so a Random Forest (RF) model trained on PDBbind data could be expected to interpolate successfully. The distributions of the pChEMBL values for the six targets are shown in Supporting Information Figure 1.

We performed three different validation experiments using the Updated DUD-E Diverse Subset. First, to explore the effect of using docked poses in place of crystallographic poses, we trained on docked poses (see "Docking Protocol") of the PDBbind Training Set and tested on docked poses of all of the ligands obtained from ChEMBL for each target. Second, we investigated whether our models could learn from docked poses obtained by docking ligands taken from ChEMBL into a single structure of a protein target, rather than using the co-crystal cognate structure of the protein each ligand. To do this, for each target we randomly selected 80% of the ligands obtained from ChEMBL for that target to use as a training set. The resulting model was then tested on the remaining 20% of the ligands for that target (so-called "Intra-target training"). Third, we investigated whether our models could generalise to ligands for a previously-unseen target when trained upon data other than the PDBbind Refined Set. For each target we trained on all ligands known to bind to the other five targets, and tested on the held-out target's set of ligands (so-called "Inter-target training").

Docking Protocol

All docking calculations were performed using Smina[23] (version November 9 2017), a fork of AutoDock Vina.[24] Protein and ligand structures were prepared for docking using the following protocol. For each ligand, an initial conformer was generated using the ETKDG method[32] implemented in RDKit version 2019.09.1.[30] Conformers were not further optimised using a force-field as Riniker and Landrum observed that ETKDG conformers were of similar quality to force-field optimised conformers for PDB-derived data.[32]

For the PDBbind data, generating a new random conformer prior to docking a ligand ensures the docking could not be biased by starting with the crystallographic bond lengths, bond angles, and torsions of the ligand. PDBQT files for both the receptor and ligand were generated using OpenBabel.[33] We used the default parameters of Smina, with the following exceptions: *autobox_add*=8; *exhaustiveness*=20; and *num_modes*=20. For each

protein-ligand complex from PDBbind, the centre of search space was defined by passing the crystallographic binding mode of the ligand using the *autobox_ligand* parameter. For each ligand, up to 20 diverse docked poses were generated by Smina. In addition to docking, we also performed a local AutoDock Vina energy minimisation of the crystallographic binding mode of the ligand provided by PDBbind using Smina to generate a single near-native docked pose. To perform the minimisation, we used Smina's *minimize* option with default parameters. For the Updated DUD-E Diverse Subset, we docked the ligands into the PDB structure provided by DUD-E for each target, using the crystallographic ligand binding pose provided by DUD-E to define the centre of the search space. Docked poses for the ligands in the PDBbind test set and Updated DUD-E Diverse Subset were generated using the same protocol and parameters used for the ligands in the PDBbind Training Set.

The quality of a docked pose was assessed by computing the root-mean-squared deviation (RMSD) of the coordinates of the atoms of the ligand's docked pose with respect to the coordinates of the ligand's atoms in the crystallographic pose. To ensure we correctly accounted for symmetry when computing the RMSD between two conformers, we identified symmetrically equivalent permutations of the atomic indices of a molecule by performing a substructure match of the molecule against itself using RDKit version 2019.09.1. We then applied these permutations to the indices of the atoms and re-computed the RMSD of the docked and crystallographic structures for each permutation. The lowest computed RMSD value was then taken as the RMSD of that docked pose.

Scoring Function Construction

We used the Random Forest (RF)[34] as implemented in Scikit-Learn version 0.22.0[35] as our learning algorithm, as our previous results demonstrated that it consistently outperformed other tested machine learning methods.[16] We built three models to predict the binding affinity of a protein-ligand complex that differed in the types of input features used: a purely ligand-based (LB) model; a traditional structure-based (SB) model using protein-

ligand intermolecular features; and a hybrid (HB) model consisting of both ligand-based and structure-based features.

For the input features of the ligand-based (LB) model, we used a set of pose-independent molecular descriptors computed for each ligand using RDKit version 2019.09.1. These descriptors are conformation-independent and may be categorised as either (computed) experimental bulk properties (such as molar refractivity or logP) or theoretical descriptors derived from a symbolic representation of the molecule. The theoretical descriptors may be further categorised according to the dimensionality of the representation of the molecule from which they are derived. The conformer-independent descriptors we consider are either 1-D compositional properties (such as heavy atom counts, bonds counts, and molecular weight) or 2-D topological properties (such as fragment counts, topological polar surface area, and connectivity index). Any features with zero variance across the training data set, or that were null-valued (*i.e.* infinite or not computable) were excluded. We removed the Ipc index[36] as it produced extreme numerical values for larger molecules that were too large to be represented as 32-bit floats. This resulted in 185 ligand-based features, and the full list of features is given in the Supporting Information, under "RDKit Features".

Our structure-based (SB) model uses the features of the Random Forest-based scoring function RF-Score v3.[37] Six of these features are the same six terms used by the AutoDock Vina scoring function: five empirical force-field-like potentials derived from the interactions between protein and ligand atoms, and the number of rotatable bonds in the ligand. These force-field-like potentials are: two Gaussian potentials (*vina_gauss1* and *vina_gauss2*), a repulsive term (*vina_repulsion*), a hydrophobic term (*vina_hydrophobic*), and a hydrogen bonding term (*vina_hydrogen*).[24] The remaining 36 features are the counts of pairwise interactions between protein and ligand atoms within 12Å of each other, for example, the number of protein carbon-ligand nitrogen pairs. Four elements (C, N, O, and S) are considered in the protein and nine elements (C, N, O, F, P, S, Cl, Br, and I) in the ligand. These features were calculated using the Open Drug Discovery Toolkit version 0.6.[38]

For the hybrid (HB) model, we used the features of both the ligand-based and structure-based models as inputs and again trained Random Forest models using the training set described above. As the number of rotatable bonds are in the features of both the LB model and the SB model, we kept only a single instance of this feature in the HB model to avoid redundancy. Our previous work[16] has shown that the correlations within the set of structure-based features or within the set of ligand-based features are higher than correlations between ligand-based and structure-based features, so additional redundancy is not introduced by combining these feature sets. We did not remove correlated features as the RF algorithm is robust with respect to feature correlations: if two correlated features are available to a tree, it will simply make a split using the most informative of the two features. The parameters for each model (such as the number of trees and the maximum proportion of the total number of features used per split) were the same, as we previously demonstrated that RF models using these features are robust with respect to parameter tuning[16].

In addition to these three models, we consider the features used by two additional structure-based methods — RF-Score v2[39] and PLECScore[29] — to explore how our results generalise to other scoring functions. RF-Score v2 uses the same protein-ligand atomic interaction features as RF-Score v3, but bins the counts in 2Å intervals, resulting in a more fine-grained representation of the interactions in a protein-ligand complex. PLECScore uses the protein-ligand extended connectivity fingerprint, which generalises the extended connectivity fingerprint[40] to protein-ligand interactions, resulting in a sparse interaction fingerprint with 65,536 bits. Both sets of features were computed using ODDT[38]. As the PLEC fingerprint is a sparse, high-dimensional representation of the data, it is not practical to directly combine it with the RDKit features, as this would prevent the use of sparse matrices to handle the large fingerprints in a computationally-efficient manner. To address this, we created an additional feature set by using truncated singular value decomposition (SVD) as implemented in scikit-learn.[35] Truncated SVD computes only the first k singular values and singular vectors of a matrix, and unlike principal component analysis (PCA) is capable

of operating on a matrix without first centring it, allowing it to be applied efficiently to a sparse matrix such as that obtained by computing PLEC fingerprints. We generated 200 components using truncated SVD, resulting in a feature set comparable in dimensionality to the features of RF-Score v2. In all experiments, the SVD is computed using the training set for that experiment, and the computed SVD applied to the test set, as computing the SVD using all data would allow information about the test set to leak into the SVD and, hence, the transformed training set. We refer to the features generated by applying truncated SVD to the PLEC fingerprints as "PLEC-SVD" throughout.

Model Training and Testing

PDBbind Validation

To investigate the effect of using docked or crystallographic ligand binding modes on binding affinity prediction, we performed a five-fold cross-validation on the PDBbind Training Set. Four approaches to training our scoring functions were compared, training on structure-based features derived from: (i) the crystallographic pose of the ligand; (ii) a single docked pose obtained by performing local minimisation of the ligand using Smina; (iii) a single docked pose, ranked highest by Smina; and (iv) multiple docked poses for each ligand, in this case up to 20 diverse poses per ligand generated by Smina. In case (iv), each pose was labelled with the same experimental binding affinity value during training.

We tested three strategies for predicting the binding affinity of a ligand using its docked poses: (i) predicting the binding affinity using the pose ranked highest by Smina; (ii) predicting the binding affinity for each docked pose and taking the highest; and (iii) predicting the binding affinity for each docked pose and taking the arithmetic mean.

Updated DUD-E Diverse Subset

Using the six targets of the Updated DUD-E Diverse Subset, we applied three different approaches to training and validation. First, models were trained on docked poses of the

PDBbind Training Set and tested on docked poses of the Updated DUD-E Diverse Subset. Second, we performed an "inter-target" validation in which each of the six targets in turn was held out as a test set with the remaining five targets forming the training set. Third, we performed an "intra-target" validation in which we randomly selected 20% of the ligands for each target to be a validation set, and trained on the remaining 80% of the ligands for that target. We then repeated both the inter-target and intra-target validation experiments by combining the DUD-E/ChEMBL data with the PDBbind Training Set.

The performance of each scoring function was evaluated by computing the Pearson correlation coefficient between the predicted and experimentally-determined values of the protein-ligand binding affinity, expressed as $\mathrm{p}K_\mathrm{i}$ values.

Model Evaluation

The performance of each model was assessed by computing the Pearson correlation coefficient, ρ_p, between its predictions and the corresponding experimentally-determined protein-ligand binding affinity. In our cross-validation experiments, we computed the mean and standard deviation of the Pearson correlation coefficient across the cross-validation folds. In experiments where a Pearson correlation coefficient is reported on a single validation set, we instead estimated the two-sided 95% confidence interval by taking 10,000 bootstrap samples of the validation set, computing the Pearson correlation coefficient for each sample, and taking the 2.5^th and 97.5^th percentiles of the range of bootstrapped coefficients as the 95% confidence interval. To test for statistical significance, we performed a one-sided permutation test under the null hypothesis that there is no correlation by taking 10,000 random permutations of the predicted $\mathrm{p}K$ values and re-computing the Pearson correlation coefficient between the permuted predicted values and non-permuted experimentally-determined values. The p-value for the one-sided permutation test was then the proportion of permutations for which the Pearson correlation coefficient was at least as large as that obtained using the non-permuted predicted values. The null hypothesis is rejected if $\mathrm{p} < 0.05$.

As the AutoDock Vina scoring function predicts the change in Gibbs free energy upon binding in units of kcal/mol, we converted its predictions to a pK value as follows. The change in Gibbs free energy upon binding, ΔG, is related to the inhibition constant K_i by:

$$\Delta G = -RTlnK_i \tag{1}$$

where $R = 0.001987$ kcal K^{-1} mol^{-1} is the gas constant and T is the temperature in Kelvin. The pK value is then given by:

$$\mathrm{p}K = -\log_{10}(\exp(\Delta G/RT)). \tag{2}$$

When computing a pK_i value from a ΔG value, we assume a standard temperature of 298K.

Results and Discussion

Accuracy of docked poses

We first evaluated the quality of the predicted binding poses generated by Smina for the 4,277 ligands in the PDBbind Training Set (3,752) and PDBbind Test Set (525). For each ligand, up to 20 diverse poses were generated. Only 1,357 ($\approx 32\%$) of the ligands re-docked by Smina had at least one pose with with an RMSD less than 2Å with respect to the crystallographic coordinates. For 1,003 ligands ($\approx 23\%$), every pose generated by Smina had an RMSD greater than 4Å. The distribution of the lowest RMSD docked pose for each complex is shown in Supporting Information Figure 2. The AutoDock Vina scoring function used by Smina often failed to rank a near-native ("good") pose first even when one was generated by Smina: of the 1,357 complexes with at least one pose with an RMSD below 2Å, about half (691) of the top-ranked poses returned for Smina had an RMSD greater than 2Å, and about a quarter (371) of these had an RMSD of greater than 4Å. This agrees with

previous results that have shown that docking scoring functions are not always a reliable way of selecting the best pose from a set of putative binding poses.[41] One well documented source of error in docking is the difficulty of sampling the conformational space of large, highly flexible molecules. However, we found little correlation between either the size or the flexibility of the ligand and the RMSD of the best docked pose (Supporting Information Figure S3). This is in agreement with the findings of Trott and Olson, who found that the accuracy of poses predicted by Vina did not correlate with the flexibility of the ligand.[24] We suspect that, while larger, more flexible ligands are generally more difficult to dock, the limitations imposed by the rigid receptor assumption and the Vina scoring function lead to errors in pose prediction regardless of the size and flexibility of the ligand. Koes et al. and McNutt et al. also discuss the limitations of Vina and Smina, in particular the default scoring function, highlighting the need for further work in docking pose prediction.[23,42]

Effect of the quality and quantity of docked poses on model training and testing

As there was considerable variation in the accuracy of the docked poses generated by Smina, we investigated how the quality of the docked pose, or poses, used to train and test our models affected binding affinity prediction. For this, we trained and tested each model using five-fold cross-validation of the PDBbind Training Set. In our first set of experiments, to determine the influence of the ligand's pose on model performance, we trained the models using a single minimised pose (obtained by minimising the crystallographic binding pose using Smina) of each ligand in the training folds and tested on all of the (up to) 20 docked poses of each ligand in the test fold. We performed six experiments investigating strategies for making a single prediction of binding affinity using multiple docked poses of the test ligand. These involved using our models to predict the binding affinity using: (1.1) the pose ranked first by Smina; (1.2) all (up to) 20 docked poses, taking the highest value as the predicted affinity; (1.3) all (up to) 20 poses and taking the mean score as the predicted

affinity; (1.4) the crystallographic coordinates of each ligand for both training and testing, as a control, and (1.5) the minimised poses of each ligand for both training and testing, as a second control. Minimised crystal poses were used during training for experiments (1.1)-(1.3) to control for possible errors due to poor docking while ensuring that, like the docked poses for the test set, the training poses were optimised using the same Smina scoring function.

Table 1 compares the average Pearson correlation coefficient achieved in each experiment under five-fold cross validation using the three docking-based scoring strategies (experiments 1.1-1.3) and the two controls (experiments 1.4-1.5). The performance of the ligand-based (LB) model is shown for comparison; its performance is necessarily the same regardless of scoring strategy, as its ligand-based features are by definition pose-independent. We also report the performance of the AutoDock Vina scoring function for each experiment. Both the structure-based (SB) model and hybrid (HB) model (a combination of LB and SB features) perform worse when tested on docked ligand poses instead of the experimentally-determined crystallographic pose, regardless of the scoring strategy. Using the mean predicted affinity across a set of docked poses when testing yields particularly poor results when compared to training and testing on crystal poses, with the Pearson correlation coefficient between predicted and experimentally-determined binding affinity dropping from 0.747 ± 0.013 to 0.604 ± 0.012 for the SB model, and from 0.768 ± 0.012 to 0.677 ± 0.011 for the HB model. Testing on the pose ranked best by Smina results in a smaller drop in Pearson correlation coefficient than when taking the mean of the predicted affinities across multiple poses: to 0.659 ± 0.017 for the SB model and to 0.714 ± 0.016 for the HB model. One possible explanation for this drop in performance is that by training only on minimised poses the models only see docked poses that are very close to the crystal pose, and so may not be capable of extrapolating to the less accurate docked poses used for testing. Regardless of the scoring strategy, the LB model, SB model, and HB model all outperform the Vina scoring function, which at best achieves a mean Pearson correlation coefficient of 0.458 ± 0.038 when tested on minimised poses (experiment 1.5). Note that the Vina scoring function performs

worse on crystal poses (0.381 ± 0.055, experiment 1.4) as these poses have not been optimised with respect to the force-field-like potentials used by Vina.

Having found that for structure-based models, training on minimised crystallographic poses and testing on docked poses results in worse binding affinity prediction performance than when training and testing on crystal poses, we next investigated whether training on docked poses instead of minimised crystal poses would enable the models to better generalise to docked poses. We thus repeated the experiments that tested on docked poses (1.1, 1.2, and 1.3), but this time trained on the pose ranked best by Smina, rather than the minimised crystallographic pose. The results are shown in Table 2. In all cases, our SB and HB models perform better when trained on docked poses than on minimised crystallographic poses. This difference is marginal when taking the highest predicted affinity of the docked poses of the test ligands (2.2), but noticeable when using just the docked pose ranked best by Smina (2.1), or when taking the mean predicted of the poses (2.3). This suggests that training on docked poses helps our SB and HB models to generalise to docked poses with varying accuracy, and that when presented with multiple docked poses, the highest score assigned by our models is likely the most accurate, reflective of the actual protein-ligand binding affinity.

Regardless of the types of ligand poses used for training and testing, the LB model consistently outperformed the SB model under cross-validation when tested on docked poses. Even better performance was obtained using the HB model when trained on the docked pose ranked highest by Smina, and using this to predict the affinities across all poses and taking the best binding affinity; indeed, this model's performance (2.2; $\rho_{\mathrm{P}} = 0.744 \pm 0.009$) was comparable to our SB model when trained and tested on crystal poses (1.4; $\rho_{\mathrm{P}} = 0.747 \pm 0.013$), suggesting that augmenting structure-based features with ligand-based features in a machine learning-based scoring function can help to correct for errors that might be introduced by the use of docked poses. Finally, as in experiments 1.1, 1.2, and 1.3, all three RF scoring functions consistently outperformed the Vina scoring function regardless of the pose(s) used for training or testing.

Experiment	Train Pose(s)	Test Pose(s)	LB Model	SB Model	HB Model	Vina
1.1	Minimised crystal	Top docked	0.719 (0.010)	0.659 (0.017)	0.714 (0.016)	0.451 (0.021)
1.2	Minimised crystal	All (max) docked	0.719 (0.010)	0.682 (0.014)	0.732 (0.010)	0.451 (0.021)
1.3	Minimised crystal	All (mean) docked	0.719 (0.010)	0.604 (0.012)	0.677 (0.011)	0.443 (0.027)
1.4	Crystal	Crystal	0.719 (0.010)	0.747 (0.013)	0.768 (0.012)	0.381 (0.055)
1.5	Minimised crystal	Minimised crystal	0.719 (0.010)	0.744 (0.015)	0.766 (0.014)	0.458 (0.038)

Table 1: Mean Pearson correlation coefficient between predicted and experimental binding affinity under five-fold cross-validation on the PDBbind Training Set (standard deviation in parentheses). Random Forest models were trained using ligand-based (LB) features, structure-based (SB) features derived from a single minimised pose, or using both LB and SB features (HB), and tested on up to 20 docked poses, or the crystallographic pose of each ligand. The Pearson correlation coefficient was computed between the experimental and predicted binding affinity values obtained in five experiments: (1.1) scoring the pose ranked best by Smina, (1.2) scoring all poses and taking the maximum score, (1.3) scoring all poses and taking the mean score, (1.4) training and testing using the experimentally-determined binding pose, and (1.5) training and testing using the minimised binding pose. As expected, the LB model is insensitive to the docked ligand pose(s), and there is little difference between using crystal poses and minimised poses for both training and testing. The best performance for each experiment is underlined. In all experiments, every ML SF outperforms the Vina SF. Note that the SB and HB models when tested on docked poses perform worse than when tested on crystallographic or minimised crystallographic poses (which perform almost identically well).

Experiment	Train Pose(s)	Test Pose(s)	LB Model	SB Model	HB Model	Vina
2.1	Top docked	Top docked	0.719 (0.010)	0.676 (0.017)	<u>0.738</u> (0.014)	0.451 (0.021)
2.2	Top docked	All (max) docked	0.719 (0.010)	0.687 (0.013)	<u>0.744</u> (0.009))	0.451 (0.021)
2.3	Top docked	All (mean) docked	0.719 (0.010)	0.643 (0.010)	<u>0.725</u> (0.013))	0.443 (0.027)
1.4	Crystal	Crystal	0.719 (0.010)	0.746 (0.013)	<u>0.768</u> (0.011)	0.381 (0.055)
1.5	Minimised crystal	Minimised crystal	0.719 (0.010)	0.744 (0.015)	<u>0.766</u> (0.014)	0.458 (0.038)

Table 2: Mean Pearson correlation coefficient between predicted and experimental binding affinity under five-fold cross-validation on the PDBbind Training Set (standard deviation in parentheses). Random Forest models were trained using the docked pose ranked highest by Smina and tested on up to 20 docked poses for each ligand. The Pearson correlation coefficient was computed using affinity predictions obtained in five ways: (2.1) scoring the pose ranked best by Smina; (2.2) scoring all poses and taking the maximum; (2.3) scoring all poses and taking the mean; (1.4) training and testing using the crystal pose; and (1.5) training and testing using the minimised binding pose. As expected, the LB model is insensitive to the docked ligand pose(s). The best performance for each scoring strategy is underlined. Once again, every ML SF outperforms Vina, and the hybrid model, HB, consistently performs better than either the LB or SB models.

To examine whether the effect of using docked poses on performance was just an artefact of the RF-Score v3 features, we repeated experiments 1.4 and 2.1 with Random Forests using RF-Score v2 features, a combination of RF-Score v2 features and the features of our LB model, PLEC fingerprints, PLEC-SVD features, and a combination of PLEC-SVD and RDKit features. For RF-Score v2, PLEC, and PLEC-SVD features, we observe a similar drop in cross-validated performance when using docked poses in place of crystal poses (Supporting Information Table S6). We also find that the combination of RF-Score v2 and RDKit features outperforms the performance of RF-Score v2 features when using either docked or crystal poses, and similarly for the PLEC-SVD features (Supporting Information Table S6). This indicates that our observations are not merely a consequence of the choice of RF-Score v3 features: the use of docked poses in place of crystal poses consistently reduced RF scoring function performance, and the addition of ligand-based features helps to recover this lost performance.

Effect of training using multiple poses

As training on a single docked pose for each ligand was more effective than training on minimised crystal poses, we next investigated how training on multiple docked poses for each ligand affected model performance. To do this, we repeated the cross-validation experiment described above, this time in experiments 3.1, 3.2, and 3.3 training on all docked poses for each ligand, using the same binding affinity value for each pose of a ligand. We used our ML models to score all poses of the ligands in the test fold, and took each model's maximum predicted binding affinity as the predicted value for that ligand. To control for the effect of substantially increasing the size of the training set, we repeated this cross-validation, but for each training ligand used the same number of copies of the pose ranked top by Smina as the number of diverse docked poses output for that ligand by Smina (up to 20). Table 3 shows the mean Pearson correlation coefficient achieved by each model over the five cross-validation folds, when trained on multiple diverse or redundant poses. Overall, there is little

difference between training on a single pose and training on multiple poses or redundant poses, with the SB model performing slightly better when trained on all poses instead of one pose ($\rho_P = 0.699 \pm 0.015$ vs $\rho_P = 0.676 \pm 0.013$), and the HB model performing slightly worse ($\rho_P = 0.723 \pm 0.008$ vs. $\rho_P = 0.738 \pm 0.009$). Similarly, training on multiple copies of the pose ranked highest by Smina does not significantly affect performance, indicating that the models are not adversely affected by redundancy in the training data. Finally, as with experiments 1.1-1.3 and 2.1-2.3, the RF scoring functions consistently outperform the Vina scoring function regardless of the poses used for training. These results, together with those in Table 1 and Table 2, suggest that when using Random Forest models for protein-ligand binding affinity prediction, it is important for the model to see examples of docked poses of varying quality, but using a variety of example poses for each complex is not necessary.

Validation on PDBbind Test Set

We next validated our LB, SB, and HB models on the PDBbind Test Set. Each model was trained on the PDBbind Training Set. We used the docked pose ranked highest by Smina for each ligand in the training set and scored all poses for each complex in the PDBbind Test Set, taking the highest predicted affinity as the value for that complex, as this was found to be the most effective strategy under cross-validation (Experiments 1.2 and 2.2).

As the PDBbind database is based on the PDB, its contents will invariably be unbalanced, as some proteins are more highly-represented than others. Further, as the CASF test sets upon which our PDBbind Test Set are based is by construction a set of proteins that are represented in the PDBbind Refined set, it is important to understand how the level of similarity between the training and test data affects model performance. We examined the influence of protein similarity between the training and test sets by excluding from the training set any structure with a protein whose sequence identity was above a given threshold to any protein in the test set. Similarly, we examined the influence of ligand similarity between the training and test sets by excluding from the training set any structure with a

Experiment	Train Pose(s)	Test Pose(s)	LB Model	SB Model	HB Model	Vina
3.1	Top docked	All (max) docked	0.719 (0.010)	0.676 (0.013)	<u>0.738</u> (0.009)	0.451 (0.021)
3.2	Diverse docked	All (max) docked	0.713 (0.006)	0.699 (0.015)	<u>0.723</u> (0.008)	0.451 (0.021)
3.3	Redundant docked	All (max) docked	0.713 (0.006)	0.689 (0.013)	<u>0.745</u> (0.010)	0.451 (0.021)

Table 3: Mean Pearson correlation coefficients between predicted and experimental binding affinity under five-fold cross-validation on the PDBbind Training Set (standard deviation in parentheses). For each ligand, models were trained using either: (3.1) the pose ranked highest by Smina (top pose), or (3.2) all diverse poses, or (3.3) redundant copies of the pose ranked highest by Smina. Predictions were made for the test fold by scoring all docked poses of each ligand and taking the highest score. The best performance for each experiment is underlined.

ligand whose 2048-bit ECFP4 fingerprint Tanimoto similarity was above a given threshold to any ligand in the test set. For each such threshold of protein sequence identity or ligand Tanimoto similarity, we repeated the PDBbind validation experiment, training only on those protein-ligand complexes that were below the similarity threshold to any complex in the test set.

Effect of training and testing using docked poses

Figure 1 shows the Pearson correlation coefficient between the predicted and experimental binding affinity achieved by the LB, SB, and HB models on the PDBbind Test Set, when trained and tested using either crystallographic binding poses or docked poses. Figure 1A shows how their performance varies with the maximum permitted protein sequence identity between the training and test sets, and Figure 1B shows how performance varies with the maximum permitted ligand Tanimoto similarity between the training and test sets. In both figures, the performance of the Vina scoring function is included for comparison. Numerical values for correlation coefficients and their corresponding bootstrapped confidence intervals were also calculated (Supporting Information Tables S2-S5).

Regardless of the maximum level of protein or ligand similarity permitted between the training and test sets, both the SB and HB models perform worse when trained and tested on docked poses (dotted lines) than when trained and tested on crystallographic poses (solid lines; in Figure 1, the dotted lines are always below the corresponding solid lines for the SB and HB models). The LB model performs identically when using docked and crystallographic poses: the RDKit molecular descriptors used by the model are independent of the pose of the ligand, so the model is unaffected by the use of different poses. The LB model actually performs better than the SB model using docked poses when no data are excluded from the training set. The LB model's performance drops below that of the SB model when complexes with 100% protein sequence identity to those in the test set (Figure 1A), or with a ligand whose Tanimoto similarity was greater than 0.6 to those in the test set (Figure 1B), are

excluded from the training set. Regardless of the level of similarity between the complexes in the training and test sets, the LB model always achieves a Pearson correlation coefficient greater than 0.55, comparable to the performance of many classical scoring functions,[4–6] indicating that these ligand-based features capture useful information for affinity prediction.

The HB model, which combines structure-based and ligand-based features, consistently outperforms the SB and LB models when using docked poses (in Figure 1A and 1B, the dotted yellow line is always above the dotted red line and blue line). Furthermore, the HB model trained and tested on docked poses has comparable performance to the SB model when trained and tested on crystal poses (in Figure 1A, the dotted yellow line is above the solid red line; in Figure 1B, the dotted yellow line is nearly identical to the solid red line for a Tanimoto similarity threshold greater than 0.5). Furthermore, the HB model is less deleteriously affected by the use of docked poses in place of crystal poses than the SB model (see, in Figure 1, the dotted yellow line much closer to the solid yellow line than the dotted red line is to the solid red line). This suggests not only that combining ligand-based and structure-based features leads to more accurate binding affinity predictions, but that the addition of ligand-based features to a structure-based scoring function can help to compensate for the errors in affinity prediction that may result from the use of potentially inaccurate docked poses. There was no statistically-significant correlation between the prediction error on the pK value and the RMSD of the docked poses for the SB or HB models on the PDBbind test set. This observation agrees with the results of Li et al.[15], who also found that the RMSD of docked poses had little correlation with their model's prediction errors. This suggests that the models are sensitive to noise in the data, but, having not been trained for the task of pose prediction, are not able to differentiate between 'bad' and 'good' docked poses.

All three models (LB, SB, and HB) are strongly affected by the similarity between the training and test sets, with the exclusion of training set complexes with similar proteins or ligands to those in the test set significantly reducing performance. These results echo our earlier results[16] and more recently those of Su et al.,[43] indicating that even when potentially

less-accurate binding poses are used, it is necessary to consider the effect of biases in the available structural data when training and evaluating models. The inclusion of ligand-based features in structure-based models always improves performance when using docked poses, and only ceases to improve performance when using crystal poses if the maximum fingerprint Tanimoto similarity between ligands in the training and test set is less than or equal to 0.5. Regardless of the level of similarity between the training and test sets, every RF scoring function out-performs the Vina scoring function, which achieved Pearson correlation coefficients of 0.574 on locally-minimised crystal poses, and 0.504 on docked poses. This indicates that training on highly-similar data is not a prerequisite for a machine learning scoring function to outperform a classical scoring function, and that, as with the machine learning scoring functions, the performance of the classical Vina scoring function is also negatively affected by the use of docked poses in place of (locally minimised) crystal poses. Furthermore, the LB, SB, and HB models and the Vina scoring function all exhibit stronger performance on the PDBbind Test Set than in the cross-validation experiments in Tables 1-3. While some difference in performance should be expected due to the size of the training set under five-fold cross-validation being only 80% that of the full PDBbind Training Set, the large difference in performance suggests that the PDBbind Test Set, derived from the PDBbind Core Sets, is not fully representative of the data found in the PDBbind Training Set, which was derived from the PDBbind Refined Set. We previously found that the PDBbind Refined Set contains both a large number of distinct clusters of proteins and a large number of proteins with unique sequences,[16] indicating that the PDBbind Refined Set and hence our PDBbind Training Set is considerably more diverse than our PDBbind Test Set. It is likely that the highly-diverse PDBbind Training Set contains more examples of complexes that are challenging to score than the PDBbind Test set, contributing to the lower performance observed under cross-validation. Sánchez-Cruz et al. observed a similar difference in performance between cross-validation and the PDBbind 2016 Core Set,[44] indicating that this effect is not just a result of our choice of models. Finally, we observe similar results

when using RF-Score v2 features or PLEC-SVD features in place of RF-Score v3 features (Supporting Information Tables S7-S8; S11-S12), and a similar impact when using docked poses in place of crystal poses with PLEC fingerprints (Supporting Information Tables S9-S10), indicating that these results are not just an artefact of the particular structure-based features of RF-Score v3. This suggests that the inclusion of ligand-based features is a robust method of enhancing scoring function performance, particularly when using docked poses in both training and testing.

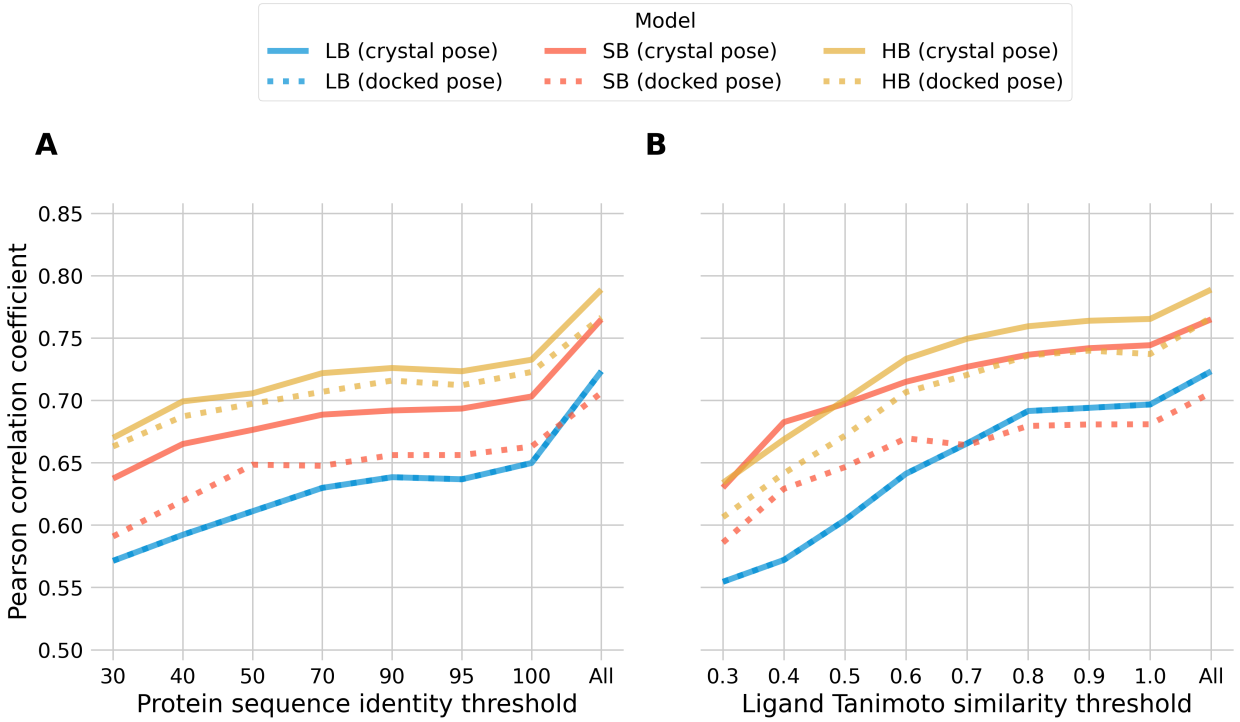

Figure 1: Pearson correlation coefficient between predicted and experimental pK values on the PDBbind Test Set, for varying levels of protein sequence identity or ligand fingerprint similarity permitted between the training and test set. Solid lines show performance when trained and tested using crystallographic binding poses; dotted lines show performance when trained and tested using docked poses. The Vina scoring function achieved a Pearson correlation coefficient of 0.574 when using locally-minimised crystal poses, and 0.504 when using docked poses. The maximum permitted level of protein sequence identity (A) or ligand fingerprint Tanimoto similarity (B) between the training and test set is shown along the horizontal axis. Note: both (A) and (B) have the same vertical axis.

Effect of training using multiple docked poses

We also checked whether training on multiple docked poses for each ligand affected performance on the PDBbind Test Set. For this we focused on the HB model as it consistently outperformed the LB and SB models in our previous experiments. We repeated the above experiment for the HB model, this time using all of the docked poses we generated for each ligand in the PDBbind Training Set, as described earlier. In contrast with the cross-validation experiment, training on multiple diverse docked poses for each ligand significantly reduced the performance of the HB model on the PDBbind Test Set (Supporting Information Figure S4). To understand how training on diverse poses for each ligand affects the HB model, we examined the relative importance of the features used by the RF. We found that when diverse poses for each ligand are used for training (Supporting Information Figure S5B), ligand-based features become more important than when only a single pose is used in training (Supporting Information Figure S5A), suggesting that using diverse poses for each ligand introduces additional noise into the structure-based features that causes the RF to rely more on ligand-based features.

Generalising to unseen proteins remains challenging

Validation on PDBbind target clusters

To understand how the HB model might be expected to perform on an unseen target using docked poses, we clustered the PDBbind Training Set at 90% protein sequence identity and identified the six largest clusters. These correspond to the proteins HIV-1 protease, carbonic anhydrase, trypsin, thrombin, heat shock protein 90α (HSP90α), and coagulation factor X. We also identified all proteins in the PDBbind Training Set that were unique at the 90% sequence identity threshold and grouped these into a seventh cluster of 'singletons', forming a highly diverse set of proteins. For each cluster, we trained the HB model on the remaining complexes in the PDBbind Training Set, using either the pose ranked highest by Smina.

We focused on the HB model as it consistently outperformed the LB and SB models in our previous experiments, but report the performance of both the LB model and the AutoDock Vina scoring function for comparison.

The results of this experiment are shown in Table 4. Overall, performance varies greatly across the seven clusters. The LB model and HB model both perform well on Trypsin ($\rho_P > 0.7$), as does the Vina scoring function. On HSP90α the LB model actually outperforms the HB model($\rho_P = 0.663$ vs $\rho_P = 0.591$, respectively), though each model's performance lies within the 95% confidence interval of the other, while the Vina scoring function performs poorly ($\rho_P = 0.202$, with a permutation test p-value greater than 0.05). On the cluster of singletons, the HB model performs best, achieving a moderate correlation ($\rho_P = 0.500$), which is greater than the 95% confidence interval of either the LB model or the Vina scoring function. Both the LB and HB models achieve moderate correlation on HIV-1 Protease and Carbonic Anhydrase 2 ($\rho_P > 0.4$), while the Vina scoring function performs poorly ($\rho_P = 0.003$, with a permutation test p-value greater than 0.05), and all models perform poorly on Coagulation Factor X $\rho_P < 0.2$, with permutation test p-values greater than 0.05. These results are similar to those of Kramer and Gedeck,[45] who showed that the performance of RF-Score[46] varies greatly when validated on held-out clusters of PDBbind data, with worse overall performance than when validated on the PDBbind 2007 Core Set. This suggests that performance on diverse benchmark sets such as the CASF test sets is not necessarily indicative of a scoring function's ability to rank ligands of a single target accurately within that benchmark when that target is held out of the training set.

Validation on the Updated DUD-E Diverse Subset

Having found that the performance of the HB model varies greatly between different protein targets in the PDBbind Test Set, we next investigated how well the model generalised to novel binding data from outside the PDBbind database. To do this, we created the Updated DUD-E Diverse Subset of targets: serine/threonine-protein kinase AKT (AKT1), cytochrome

Table 4: Pearson correlation coefficients achieved by the HB model on seven target sets of the PDBbind Training Set when trained on the remaining structures in the PDBbind Training Set (95% confidence interval in parentheses). For the training set, the pose ranked highest by Smina was used. For the test sets, all docked poses were scored and the highest value taken as the predicted binding affinity for that ligand. The best performance for each target set is underlined

Target Set	**LB Model**	**HB Model**	**Vina**
HIV-1 Protease	0.414 (0.299, 0.512)	<u>0.450</u> (0.347, 0.538)	0.003 (-0.148, 0.149)
Carbonic Anhydrase 2	<u>0.477</u> (0.373, 0.572)	0.460 (0.354, 0.555)	0.229 (0.114, 0.342)
Trypsin	0.715 (0.589, 0.812)	<u>0.746</u> (0.620, 0.846)	0.740 (0.627, 0.825)
Thrombin	0.270 (0.032, 0.458)	0.316 (0.078, 0.493)	<u>0.358</u> (0.168, 0.533)
HSP90α	<u>0.663</u> (0.472, 0.793)	0.591 (0.373, 0.741)	0.202 (-0.057, 0.435)
Coagulation Factor X	0.111 (-0.154, 0.331)	-0.101 (-0.380, 0.168)	<u>0.195</u> (-0.073, 0.424)
Singletons	0.413 (0.346, 0.477)	<u>0.500</u> (0.473, 0.559)	0.351 (0.283, 0.418)

P450 3A4 (CP3A4), glucocorticoid receptor (GCR), HIV-1 protease (HIVPR), HIV-1 reverse transcriptase (HIVRT), and kinesin-like protein 1 (KIF11), excluding the two targets for which no K_i data could be found.

Tables 5 and 6 show the Pearson correlation coefficients achieved by the HB model and LB model, respectively, for each target when trained on the PDBbind Training Set, and under both intra-target and inter-target validation, as described in the Methods section. In each case, we trained on the pose ranked top by Smina and tested by scoring all diverse docked poses and taking the highest predicted binding affinity for each ligand.

Using the PDBbind Training Set the performance for all six targets is poor, with the HB model achieving a Pearson correlation coefficient below 0.5 for each target and, in the case of HIVRT and KIF11, the permutation test indicated no statistically significant correlation between predicted and experimentally-determined binding affinities ($p > 0.05$). The weak correlation achieved for AKT1 and HIVPR ($\rho_{\mathrm{p}} > 0.45$) suggests a small degree of generalisation, but overall performance was poor. These results are in stark contrast with the strong performance on docked poses of the PDBbind Combined Core Set, indicating that even under strict training and validation conditions, the model generalises poorly to data sourced from outside the PDBbind database. One possible source of this difference in performance is that, as previously discussed, the docked poses for the PDBbind data were generated by

re-docking each ligand into the corresponding crystal structure of its protein target, so any induced fit effects may already be captured in the structure of the protein. Because of this, the docked poses for the PDBbind data may be more accurate than those for the Updated DUD-E Diverse Set. In addition, any uncertainties in the more recent ChEMBL binding affinity data could also contribute to noise and performance degradation when training and validating the model.

Under intra-target training and validation, while performance varies greatly between targets, the overall performance is much better than when the models were trained on PDBbind data. For each target, the HB model achieves a Pearson correlation coefficient, ρ_{p} in the range of 0.5 to 0.8, in contrast to the weaker correlations obtained when training on PDBbind data ($\rho_{\mathrm{p}} < 0.5$). This indicates that the HB model is capable of making accurate predictions using docked poses of ligands for a protein target, provided it has previously seen examples of complexes of ligands for that target.

In contrast with intra-target validation, under inter-target validation the HB model achieves at best a weak correlation between predicted and experimentally-determined pK values, with $\rho_{\mathrm{p}} < 0.4$ for all targets, indicating that the model struggles to generalise to a previously-unseen protein target. Again, for HIVRT and KIF11, the permutation test indicated no statistically significant correlation, indicating that these targets were particularly challenging for the model. With the exception of CP3A4, the predicted pK values also span a much narrower range of values than the experimentally-determined values (Supporting Information Figures S10 and S11), suggesting that under this training regime, the model is simply unable to differentiate reliably between different ligands for the unseen target. The AutoDock Vina scoring function only achieved a statistically-significant correlation (permutation test $p < 0.05$) on two of six targets, namely, AKT1 ($\rho_{\mathrm{p}} = 0.398$) and HIVPR ($\rho_{\mathrm{p}} = 0.420$), comparable to that achieved by the HB model when trained on the PDBbind Training Set. Similar results were obtained using the LB model (see Table 6) indicating that both structure-based and ligand-based RF models will struggle to generalise to novel

targets.

Finally, we repeated the inter-target and intra-target validation experiments, adding the PDBbind Training Set to the Updated DUD-E Diverse Subset data, and found that there was no improvement in performance when combining these two sources of training data, suggesting that the size of the training set does not explain the failure to generalise. We also observed similar performance when using RF-Score v2 or PLEC-SVD features, indicating that these results are not solely a consequence of the RF-Score v3 features. This indicates that a larger, more diverse training set does not necessarily translate into improved performance; we made a similar observation in previous work[16] when we showed that the improved performance of several scoring functions on the CASF test sets when training on the PDBbind General Set instead of the PDBbind Refined Set could be attributed not to the increased size and diversity of the General Set, but to its increased representation of the protein targets present in the test sets.

One possible explanation for these results is that docking the ligands for each DUD-E target into a single protein structure provided by DUD-E for that target does not adequately account for induced fit effects, in contrast with docking a ligand into its cognate protein structure and concomitant induced fit, such as when re-docking the PDBbind complexes. This would result in the protein-ligand interaction features failing to capture important interactions between the protein and the ligand, as an accurate docked pose with respect to the correct conformation of the protein's binding pocket is not available. This highlights the need for new experiments to look at the influence of induced fit on docking and subsequent binding affinity predictions when using machine learning models. Further, each of the targets in the Updated DUD-E Diverse Subset is very different from the others, and so these results echo those obtained under Leave-Cluster-Out cross-validation on the PDBbind training set (Table 4), demonstrating that more, and more stringent, benchmarks are required to reliably assess the performance of machine learning scoring functions in real-world drug-discovery scenarios.

Table 5: Pearson correlation coefficients achieved by the HB model between predicted and experimental pK of ligands for six protein targets in the Updated DUD-E Diverse Subset. Three different validation regimes were used. Under 'PDBbind Training', each model was trained on the PDBbind Training Set and tested on all ligands for each target. Under 'Intra-target validation', for each target 80% of the ligands were randomly selected to be used as a training set, and the model tested on the remaining 20%. Under 'Inter-target validation', for each target in turn, all ligands for that target were held out as a test set, with the model trained on all ligands for the remaining five targets. The performance of the AutoDock Vina scoring function on the set of all ligands for each target is shown for comparison. The best performance for each target is underlined.

Target	**PDBbind Training**	**Intra-target Training**	**Inter-target Training**	**Vina**
AKT1	0.418 (0.330, 0.495)	<u>0.628</u> (0.374, 0.810)	0.223 (0.110, 0.336)	0.398 (0.296, 0.495)
CP3A4	0.320 (0.193, 0.445)	<u>0.527</u> (0.233, 0.733)	0.307 (0.184, 0.430)	0.157 (0.048, 0.394)
GCR	0.317 (0.239, 0.391)	<u>0.809</u> (0.740, 0.863)	0.235 (0.172, 0.299)	-0.003 (-0.051, 0.046)
HIVPR	0.456 (0.415, 0.494)	<u>0.739</u> (0.677, 0.792)	0.185 (0.145, 0.225)	0.420 (0.380, 0.458)
HIVRT	-0.101 (-0.322, 0.097)	<u>0.572</u> (0.052, 0.908)	0.069 (-0.136, 0.304)	-0.100 (-0.272, 0.106)
KIF11	-0.177 (-0.375, 0.049)	<u>0.742</u> (0.623, 0.857)	-0.013 (-0.205, 0.180)	0.048 (-0.126, 0.228)

Table 6: Pearson correlation coefficients achieved by the LB model between predicted and experimental pK of ligands for six protein targets in the Updated DUD-E Diverse Subset. Three different validation regimes were used. Under 'PDBbind Training', each model was trained on the PDBbind Training Set and tested on all ligands for each target. Under 'Intra-target validation', for each target 80% of the ligands were randomly selected to be used as a training set, and the model tested on the remaining 20%. Under 'Inter-target validation', for each target in turn, all ligands for that target were held out as a test set, with the model trained on all ligands for the remaining five targets. The performance of the AutoDock Vina scoring function on the set of all ligands for each target is shown for comparison. The best performance for each target is underlined.

Target	**PDBbind Training**	**Intra-target Training**	**Inter-target Training**	**Vina**
AKT1	0.377 (0.296, 0.447)	<u>0.635</u> (0.388, 0.817)	0.211 (0.099, 0.322)	0.398 (0.296, 0.495)
CP3A4	0.337 (0.221, 0.443)	<u>0.522</u> (0.214, 0.745)	0.335 (0.219, 0.454)	0.157 (0.048, 0.394)
GCR	0.504 (0.445, 0.558)	<u>0.813</u> (0.746, 0.866)	0.253 (0.197, 0.309)	-0.003 (-0.051, 0.046)
HIVPR	0.457 (0.420, 0.494)	<u>0.729</u> (0.664, 0.785)	0.072 (0.034, 0.111)	0.420 (0.380, 0.458)
HIVRT	-0.111 (-0.305, 0.069)	<u>0.518</u> (-0.022, 0.905)	0.142 (-0.064, 0.368)	-0.100 (-0.272, 0.106)
KIF11	0.029 (-0.174, 0.246)	<u>0.738</u> (0.595, 0.870)	-0.114 (-0.286, 0.066)	0.048 (-0.126, 0.228)

Conclusions

We have investigated how the use of docked poses in place of X-ray crystallographic binding poses of ligands affects the ability of Random Forest-based scoring functions to predict protein-ligand binding affinity; how best to make use of docked poses when multiple poses are available; and how scoring functions trained on docked poses generalise to novel sets of ligands.

Like Li *et al.*[15], we found that the use of Smina docked poses in place of X-ray crystallographic binding modes for training and validation reduces the performance of a structure-based scoring function. However, a hybrid model that combines structure-based and ligand-based features is less deleteriously affected by the use of docked poses in training and validation than a purely structure-based model. Furthermore, hybrid models trained on docked poses can achieve binding affinity prediction performance comparable to that of a structure-based model that was trained and validated using crystallographic binding modes. Excluding proteins and ligands from the training set that are similar to those in the test set negatively affected scoring function performance, but ligand-based features still improved binding affinity predictions when removing test-set-similar data from the training set, in agreement with our previous work using crystallographic binding poses[16]. hese observations align with those of other recent studies[47,48] on the effect of protein and ligand similarity on machine learning scoring function performance, highlighting the need to consider carefully whether performance on a single benchmark may be influenced by the construction of the benchmark and availability of training data.

We also investigated the effect of training on multiple diverse docked poses. Under cross-validation on the PDBbind Refined Set, there was little difference between training on a single pose per ligand and training on multiple poses. However, when we trained our models using multiple poses for each ligand in the PDBbind Training Set and tested on our held-out PDBbind Test Set, performance was substantially worse than when training using a single docked pose for each ligand. To understand how training on multiple poses per ligand

affected the model, we examined the relative importance of the features in each model, and found that when both structure-based and ligand-based features are used, ligand-based features become more important when training on multiple poses, suggesting errors in the structure-based features from the additional docked poses degrade their relative utility.

Using binding affinity data obtained from the ChEMBL database (version 25) for six of the eight protein targets in the DUD-E Diverse Subset, we investigated how our models performed on unseen data. Under an intra-target validation where a random selection of 80% of the data for a target was used to train the model and 20% of the data was held out as a test set, our model achieved positive Pearson correlation coefficients between the predicted and experimental binding affinites ranging from 0.541 to 0.810, indicating that our hybrid model is capable of accurate predictions on the data obtained from ChEMBL when trained in a target-specific manner. However, under an inter-target validation scenario, where data for five of the six targets were used to train the model, and the remaining target's data held out as a test set, our hybrid model failed to achieve any meaningful correlation between predicted and experimental binding affinity. Combining the ChEMBL data with PDBbind data to form a larger, more diverse training set did not improve performance.

Our results indicate that, on a benchmark such as the PDBbind 2007, 2013, and 2016 Core Sets, the use of docked poses for training and validation decreases the performance of a structure-based scoring function relative to those trained and tested on X-ray crystal structures, in agreement with the findings of Li *et al.*.[15] Similarities between proteins and ligands in the training and test sets has a strong influence on scoring function performance, in agreement with our previous work[16] and the more recent work by Su *et al.*.[43] Our results once again suggest that the inclusion of ligand-based features in the scoring function helps to counteract this effect. However, our results also suggest that a model trained on PDBbind data tends to generalise poorly to external data sets, indicating that additional training, validation, and benchmarking sets are needed for scoring function development, although this will depend on the level of similarity between the training complexes and the unseen proteins

or ligand. Given the challenge posed by generalising to novel targets and sets of ligands, it is clear that further work is needed to understand to what extent machine learning models are learning interpretable and generalisable protein-ligand interactions. Furthermore, a recent study by Sieg et al. highlighted the effect of similarity bias between training and validation data in the context of active-inactive classification,[49] indicating that this problem is not limited to the task of affinity prediction. Another avenue for improvement when working with docked poses is the development of pose classification models capable of identifying accurate docked poses, and affinity prediction models capable of handling poses of varying quality. However, training and validating truly generalisable machine learning models for binding affinity prediction may remain challenging until sufficient quantities of high-quality, diverse, and non-redundant structural and binding data become more readily available.

Overall, despite potential problems generalising to novel proteins and ligands, the inclusion of ligand-based features in a structure-based scoring function can help to compensate for errors in binding affinity prediction due to the use of docked poses. The information captured by ligand-based features is orthogonal to that captured by the more commonly-used structure-based features, and our results show that this information becomes more important as noise is introduced into the structure-based features through docking errors. While the use of ligand-based features increases the risk of over-fitting to known ligands, our results show that ligand-based features can improve scoring function performance even when ligands similar to those in the test set are excluded from the training set. By opening the door to the use of docked poses, our hybrid approach expands the utility of machine learning scoring functions for the discovery of novel small molecules for novel targets.

Data and Software Availability

All features and code necessary to reproduce our results are available from:
https://github.com/oxpig/learning-from-docked-poses

Docked poses for the ligands in the PDBbind Training Set, PDBbind Test Set, and Updated DUD-E Diverse Subset generated using the protocol and parameters described in this paper are available from Figshare:
https://doi.org/10.6084/m9.figshare.13713226.v1

Acknowledgement

This work was supported by funding from the Engineering and Physical Sciences Research Council (EPSRC) [grant numbers EP/G03706X/1; EP/S024093/1; EP/L016044/1].

Supporting Information Available

Lists of feature names for the LB, SB, and HB models, and additional figures and tables showing additional results and estimated confidence intervals, are available in the Supporting Information.

References

(1) Stumpfe, D.; Bajorath, J. Current Trends, Overlooked Issues, and Unmet Challenges in Virtual Screening. *J. Chem. Inf. Model.* **2020**, *60*, 4112—-4115.

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

(3) Gorgulla, C.; Boeszoermenyi, A.; Wang, Z.-F.; Fischer, P. D.; Coote, P. W.; Das, K. M. P.; Malets, Y. S.; Radchenko, D. S.; Moroz, Y. S.; Scott, D. A., et al. An Open-Source Drug Discovery Platform Enables Ultra-Large Virtual Screens. *Nature* **2020**, *580*, 663–668.

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

(5) Li, Y.; Han, L.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results. *J. Chem. Inf. Model.* **2014**, *54*, 1717–36.

(6) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.* **2018**, *59*, 895–913.

(7) Durrant, J. D.; McCammon, J. A. NNScore 2.0: A Neural-Network Receptor–Ligand Scoring Function. *J. Chem. Inf. Model.* **2011**, *51*, 2897–2903.

(8) Zilian, D.; Sotriffer, C. A. SFCscore RF: A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein–Ligand Complexes. *J. Chem. Inf. Model.* **2013**, *53*, 1923–1933.

(9) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collections of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47*, 2977–2980.

(10) Liu, Z.; Su, M.; Han, L.; Liu, J.; Yang, Q.; Li, Y.; Wang, R. Forging the Basis for Developing Protein–Ligand Interaction Scoring Functions. *Acc. Chem. Res.* **2017**, *50*, 302–309.

(11) Dunbar, J. B.; Smith, R. D.; Yang, C.-Y.; Ung, P. M.-U.; Lexa, K. W.; Khazanov, N. A.; Stuckey, J. A.; Wang, S.; Carlson, H. A. CSAR Benchmark Exercise of 2010: Selection of the Protein–Ligand Complexes. *J. Chem. Inf. Model.* **2011**, *51*, 2036–2046.

(12) Damm-Ganamet, K. L.; Smith, R. D.; Dunbar, J. B.; Stuckey, J. A.; Carlson, H. A. CSAR Benchmark Exercise 2011–2012: Evaluation of Results from Docking and Relative Ranking of Blinded Congeneric Series. *J. Chem. Inf. Model.* **2013**, *53*, 1853–1870.

(13) Jiménez, J.; Skalic, M.; Martinez-Rosell, G.; De Fabritiis, G. K_{DEEP}: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* **2018**, *58*, 287–296.

(14) Morrone, J. A.; Weber, J. K.; Huynh, T.; Luo, H.; Cornell, W. D. Combining Docking Pose Rank and Structure with Deep Learning Improves Protein–Ligand Binding Mode Prediction over a Baseline Docking Approach. *J. Chem. Inf. Model.* **2020**, *60*, 4170–4179.

(15) Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P. J. Correcting the Impact of Docking Pose Generation Error on Binding Affinity Prediction. *BMC Bioinf.* **2016**, *17*, 308.

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

(17) Durrant, J. D.; McCammon, J. A. NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein–Ligand Complexes. *J. Chem. Inf. Model.* **2010**, *50*, 1865–1871.

(18) Wang, C.; Zhang, Y. Improving Scoring-Docking-Screening Powers of Protein–Ligand Scoring Functions Using Random Forest. *J. Comp. Chem.* **2017**, *38*, 169–177.

(19) Kumar, S.; Kim, M.-h. SMPLIP-Score: Predicting Ligand Binding Affinity from Simple and Interpretable On-the-fly Interaction Fingerprint Pattern Descriptors. *J. Cheminf.* **2021**, *13*, 1–17.

(20) Jiménez-Luna, J.; Pérez-Benito, L.; Martínez-Rosell, G.; Sciabola, S.; Torella, R.; Tresadern, G.; De Fabritiis, G. DeltaDelta Neural Networks for Lead Optimization of Small Molecule Potency. *Chem. Sci.* **2019**, *10*, 10911–10918.

(21) Nguyen, D. D.; Cang, Z.; Wu, K.; Wang, M.; Cao, Y.; Wei, G.-W. Mathematical

Deep Learning for Pose and Binding Affinity Prediction and Ranking in D3R Grand Challenges. *J. Comput.-Aided Mol. Des.* **2019**, *33*, 71–82.

(22) Li, H.; Sze, K.-H.; Lu, G.; Ballester, P. J. Machine-Learning Scoring Functions for Structure-Based Drug Lead Optimization. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2020**, *10*, e1465.

(23) Koes, D. R.; Baumgartner, M. P.; Camacho, C. J. Lessons Learned in Empirical Scoring with Smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* **2013**, *53*, 1893–1904.

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

(25) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594.

(26) 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*, D945–D954.

(27) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(28) Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P. Low-Quality Structural and Interaction Data Improves Binding Affinity Prediction via Random Forest. *Molecules* **2015**, *20*, 10947–10962.

(29) Wójcikowski, M.; Kukiełka, M.; Stepniewska-Dziubinska, M. M.; Siedlecki, P. Development of a Protein–Ligand Extended Connectivity (PLEC) Fingerprint and Its Application for Binding Affinity Predictions. *Bioinformatics* **2018**, *35*, 1334–1341.

(30) Landrum, G. RDKit: Open-Source Cheminformatics. `http://www.rdkit.org`, Accessed 20/07/2018.

(31) ChEMBL Data Questions. `https://chembl.gitbook.io/chembl-interface-documentation/frequently-asked-questions/chembl-data-questions#what-is-pchembl`, Accessed 04/01/2020.

(32) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.

(33) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 33.

(34) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

(35) 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, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(36) Bonchev, D.; Trinajstić, N. Information Theory, Distance Matrix, and Molecular Branching. *J. Chem. Phys.* **1977**, *67*, 4517–4533.

(37) 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. *Mol. Inf.* **2015**, *34*, 115–126.

(38) Wójcikowski, M.; Zielenkiewicz, P.; Siedlecki, P. Open Drug Discovery Toolkit (ODDT): A New Open-Source Player in the Drug Discovery Field. *J. Cheminf.* **2015**, *7*, 26.

(39) Ballester, P. J.; Schreyer, A.; Blundell, T. L. Does a more precise chemical description of protein–ligand complexes lead to more accurate prediction of binding affinity? *J. Chem. Inf. Model.* **2014**, *54*, 944–955.

(40) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(41) Ramírez, D.; Caballero, J. Is It Reliable to Take the Molecular Docking Top Scoring Position as the Best Solution without Considering Available Structural Data? *Molecules* **2018**, *23*, 1038.

(42) McNutt, A.; Francoeur, P.; Aggarwal, R.; Masuda, T.; Meli, R.; Ragoza, M.; Sunseri, J.; Koes, D. GNINA 1.0: Molecular Docking with Deep Learning. 2021; https://chemrxiv.org/articles/preprint/GNINA_1_0_Molecular_Docking_with_Deep_Learning/13578140/1.

(43) Su, M.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Tapping on the Black Box: How Is the Scoring Power of a Machine-Learning Scoring Function Dependent on the Training Set? *J. Chem. Inf. Model.* **2020**, *60*, 1122–1136.

(44) Sánchez-Cruz, N.; Medina-Franco, J. L.; Mestres, J.; Barril, X. Extended Connectivity Interaction Features: Improving Binding Affinity Prediction Through Chemical Description. *Bioinformatics* **2020**, *37*, 1376–1382.

(45) 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*, 1961–1969.

(46) 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*, 1169–1175.

(47) 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*, 12.

(48) Li, H.; Peng, J.; Sidorov, P.; Leung, Y.; Leung, K.-S.; Wong, M.-H.; Lu, G.; Ballester, P. J. Classical Scoring Functions for Docking are Unable to Exploit Large Volumes of Structural and Interaction Data. *Bioinformatics* **2019**, *35*, 3989–3995.

(49) Sieg, J.; Flachsenberg, F.; Rarey, M. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2019**, *59*, 947–961.

Graphical TOC Entry