Bioinformatics, 36(3), 2020, 758–764
doi: 10.1093/bioinformatics/btz665
Advance Access Publication Date: 26 August 2019
Original Paper

Structural bioinformatics

Learning from the ligand: using ligand-based features to improve binding affinity prediction

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

Department of Statistics, University of Oxford, Oxford OX1 3LB, UK

*To whom correspondence should be addressed.

Associate Editor: Arne Elofsson

Received on May 22, 2019; revised on August 14, 2019; editorial decision on August 16, 2019; accepted on August 21, 2019

Abstract

Motivation: Machine learning scoring functions for protein–ligand binding affinity prediction have been found to consistently outperform classical scoring functions. Structure-based scoring functions for universal affinity prediction typically use features describing interactions derived from the protein–ligand complex, with limited information about the chemical or topological properties of the ligand itself.

Results: We demonstrate that the performance of machine learning scoring functions are consistently improved by the inclusion of diverse ligand-based features. For example, a Random Forest (RF) combining the features of RF-Score v3 with RDKit molecular descriptors achieved Pearson correlation coefficients of up to 0.836, 0.780 and 0.821 on the PDBbind 2007, 2013 and 2016 core sets, respectively, compared to 0.790, 0.746 and 0.814 when using the features of RF-Score v3 alone. Excluding proteins and/or ligands that are similar to those in the test sets from the training set has a significant effect on scoring function performance, but does not remove the predictive power of ligand-based features. Furthermore a RF using only ligand-based features is predictive at a level similar to classical scoring functions and it appears to be predicting the mean binding affinity of a ligand for its protein targets.

Availability and implementation: Data and code to reproduce all the results are freely available at http://opig.stats.ox.ac.uk/resources.

Contact: morris@stats.ox.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Structure-based virtual screening (SBVS) uses the 3D structure of a target protein to screen large compound libraries for small molecules likely to bind. 'Explicit' SBVS uses protein–ligand docking to predict the binding mode of each compound within the active site, and a scoring function to predict the strength of binding (Gilson and Zhou, 2007; Huang *et al.*, 2010; Ripphausen *et al.*, 2012; Sousa *et al.*, 2006, 2013). While it is possible to compute the binding affinity of a compound using more rigorous methods such as free energy perturbation (Aldeghi *et al.*, 2016), their computational cost makes them impractical for screening libraries of millions of compounds (Perez *et al.*, 2016). To save time, in SBVS, more approximate scoring functions designed to estimate the binding affinity are used. Scoring functions are typically employed for four tasks in SBVS: correctly identifying the binding mode of a ligand (pose prediction or 'docking'); classifying molecules as either active or inactive ('virtual screening'); ranking ligands in order of their binding affinity for a given protein target ('ranking'); and scoring the binding affinity of a protein–ligand complex ('screening'). Popular protein–ligand docking packages, such as GOLD (Jones *et al.*, 1997), Glide (Friesner *et al.*, 2004; Halgren *et al.*, 2004), ICM (Abagyan *et al.*, 1994), FlexX (Rarey *et al.*, 1996), Surflex (Jain, 2003) and the AutoDock family (Morris *et al.*, 2009; Ravindranath *et al.*, 2015; Trott and Olson, 2010), rely on a single scoring function to perform all three tasks simultaneously. These scoring functions make use of molecular force fields, statistical potentials or linear combinations of empirical terms to assign a score to a receptor–ligand complex, and are often collectively referred to as 'classical' scoring functions. While docking methods have been used successfully to predict binding modes, and for virtual screening, correctly ranking ligands by their binding affinity for a protein remains extremely challenging (Li *et al.*, 2014b,c).

Recently, however, the application of machine learning techniques has led to the development of new scoring functions that outperform classical scoring functions in terms of ranking compounds by binding affinity (Ballester and Mitchell, 2010; Durrant and McCammon, 2011; Li *et al.*, 2014a; Wójcikowski *et al.*, 2018; Zilian and Sotriffer, 2013). Like classical scoring functions, these methods use the 3D structure of the protein and an automatically generated binding mode of the ligand to compute structure-based interaction features between the protein and ligand. The binding mode-derived features are then used as inputs to a machine learning algorithm to predict the binding affinity. The features used by such

machine learning scoring functions typically focus on capturing interactions between the ligand and the protein, but make limited use of the bulk physical, chemical and topological properties of the protein and/or the ligand alone. Ligand-based features are widely used to select potential binders in ligand-based virtual screening, and have proven to be very effective in understanding polypharmacological relationships between proteins (Lin *et al.*, 2013). Although scoring functions, such as SFCscore (Sotriffer *et al.*, 2008; Zilian and Sotriffer, 2013), NNScore 2.0 (Durrant and McCammon, 2011), and those used in AutoDock 4 (Morris *et al.*, 2009) and AutoDock Vina (Trott and Olson, 2010) include some features of the ligand, the use of detailed information about the ligand to predict cognate protein–ligand binding affinity remains limited. Perhaps the closest is the field of proteochemometric modelling (van Westen *et al.*, 2011), in which classification models are constructed using a combination of features describing the protein and the small molecule. This approach has proven to be capable of predicting ligand selectivity (Ain *et al.*, 2014) and capturing polypharmacology (Paricharak *et al.*, 2015).

Motivated by the utility of ligand-based features in virtual screening and proteochemometric modelling, we investigated whether a more detailed representation of the ligand can improve the ability of a scoring function to predict its binding affinity. Using the cheminformatics toolkit RDKit (https://www.rdkit.org/, accessed May 17, 2019), we computed a diverse set of 1D and 2D ligand molecular descriptors and combined these with the structure-based features used by the machine learning scoring functions RF-Score (Ballester and Mitchell, 2010), RF-Score v3 (Li *et al.*, 2015a), NNScore 2.0 (Durrant and McCammon, 2011), as well as the empirical scoring function of AutoDock Vina (Trott and Olson, 2010). We show that a Random Forest (RF) (Breiman, 2001) regression model using both structure-based and ligand-based features consistently outperforms a model using only structure-based features when benchmarked on three versions of the PDBbind (Liu *et al.*, 2017) core set, corresponding to the scoring power test of the three Comparative Assessment of Scoring Functions (CASF) (Cheng *et al.*, 2009; Li *et al.*, 2014b,c; Su *et al.*, 2018).

We find that removal of test set similar proteins and ligands from the training sets degrades performance but does not abrogate the predictive power of ligand-based features. Furthermore, we show that a model using only ligand-based features appears to be predictive of the mean affinity of a ligand for its binding partners when trained and tested on PDBbind data. We computed the relative importance of the features used by each model and show that when structure-based and ligand-based features are combined, both structure-based and ligand-based features are among the top-ranked features. These results suggest that quickly-computed ligand-based features should be used to improve the ability of a machine learning scoring function to predict protein–ligand binding affinity.

2 Materials and methods

2.1 Training and test sets

In this work we focused on the task of 'scoring': the prediction of protein–ligand binding affinity given the binding mode of the ligand. To accomplish this, we restricted our data to protein–ligand complexes for which a crystal structure of the bound complex and an experimentally determined value of the binding affinity was available. The PDBbind database (Liu *et al.*, 2017) is a curated set of bound macromolecule structures drawn from the Protein Data Bank (PDB) (Berman *et al.*, 2000), each with an experimentally measured binding affinity for its binding partner. Each release of PDBbind includes a 'general set', which contains all the protein–ligand structures in the database; and a 'refined set', a subset of protein–ligand complexes satisfying strict criteria concerning structure quality, affinity data reliability and the nature of the complex. The 2018 release of PDBbind contains 16 151 protein-ligand complexes in the general set, with 4463 complexes in the refined set. We used the refined set as our primary source of training data; however, it has been reported that including the lower-quality data comprising the remainder of the general set can still improve the performance of machine learning scoring functions (Li *et al.*, 2015b), so we repeated our analysis using the general set as our source of training data.

To validate our models, we used a subset of the PDBbind refined set referred to as the 'core set'. This is obtained by clustering the proteins in the refined set at 90% sequence identity and selecting three or five (depending on the version of PDBbind) representatives of each cluster for which the corresponding ligands have a broad range of binding affinity values, resulting in a diverse, non-redundant set of protein–ligand complexes. We repeated our tests using the core sets from the 2007, 2013 and 2016 releases of PDBbind. These versions of the core set were used as the 'scoring power' benchmark in the CASF exercises: CASF2009 (Cheng *et al.*, 2009), CASF2013 (Li *et al.*, 2014b,c) and CASF2016 (Su *et al.*, 2018), respectively, allowing our results to be compared directly to previously published scoring function benchmarks. We used the test sets corresponding to all three CASF benchmarks since the contents of each test set are substantially different to the others and so each set offers a different challenge for a scoring function. Excluding proteins and ligands that could not be parsed by OpenBabel or RDKit resulted in 2007, 2013 and 2016 core sets containing 196, 180 and 276 structures, respectively. The full list of structures omitted from the core sets because of parsing failures is included in the Supplementary Material. These test sets are relatively small, making it difficult to identify statistically significant differences in results generated by different models. To partially overcome this shortcoming without deviating from widely used benchmarks, we combined the structures from each core set into a fourth test set with duplicate strictures removed. This 'combined core set' numbered 525 structures, almost twice the size of the PDBbind 2016 core set, and was used as our primary test set.

PDBbind provides for each complex an experimentally determined value of the inhibition constant K_i, the dissociation constant K_d, or the half-maximal inhibitory concentration IC_{50}, in decreasing order of preference (e.g. if both K_i and K_d values are available, PDBbind reports the measurement of K_i). The refined set includes only measurements of K_i and K_d, while the general set also includes data for which only IC_{50} measurements were available. For our purposes, these values are used interchangeably and are collectively denoted by the binding constant, K. We used the negative base-10 logarithm of K, commonly denoted as pK:

$$pK = -\log_{10} K$$

We evaluated each scoring function by computing the Pearson correlation coefficient, ρ_p, between its predictions $\{\hat{y}\}$ and the experimental values $\{y\}$ of pK for the complexes in the test set. Two-tailed confidence intervals for ρ_p were estimated via bootstrapping and a one-sided permutation test was performed to assess the possibility that correlations arose by random chance. For details of both methods, see 'Confidence Intervals and Permutation Tests', Supplementary Material, p. 2. The Mann–Whitney U-test was used to compare the distribution of bootstrapped ρ_p values to assess the significance of the differences in correlation coefficient between two scoring functions.

2.2 Ligand-based features

To represent the ligand in our models, we used a diverse set of molecular descriptors computed using the cheminformatics toolkit RDKit. Using the *Descriptors* module of the Python RDKit package version 2018.03, we computed a set of 200 molecular descriptors for each ligand. These descriptors are conformation independent and may be categorized as either (computed) experimental properties (e.g. molar refractivity, log P) or theoretical descriptors derived from a symbolic representation of the molecule. The theoretical descriptors may be further categorized according to the dimensionality of the representation of the molecule from which they are derived. The conformer independent descriptors we consider are either 1D compositional properties (e.g. heavy atom counts, bonds counts and molecular weight) or 2D topological properties [e.g. fragment counts, topological polar surface area (TPSA) and connectivity index]. Any features with zero variance across the dataset, or that were null valued (i.e. infinite or not computable) within the dataset were excluded. We removed the Ipc index (an information theory-derived descriptor) as it produced extreme numerical values

for larger molecules (too large to be represented as 32-bit floats). In total, 185 RDKit descriptors were retained. We refer to this set of ligand-based features as 'RDKit descriptors' throughout this work.

2.3 Structure-based features

To investigate the effects of augmentation with ligand molecular descriptors, we considered the features of several publicly available machine learning scoring functions, namely RF-Score (Ballester and Mitchell, 2010), RF-Score v3 (Li *et al.*, 2015a) and NNScore 2.0 (Durrant and McCammon, 2011). Both RF-Score v3 and NNScore 2.0 include the six terms used by the AutoDock Vina scoring function (Trott and Olson, 2010). We therefore considered the AutoDock Vina terms separately to examine the effect of combining ligand molecular descriptors with just the terms used by a classical empirical scoring function. We computed the features of each of these scoring functions using the implementations provided by the Open Drug Discovery Toolkit (ODDT) version 0.6 (Wójcikowski *et al.*, 2015).

2.4 Varying training set size and composition

Previous works by numerous authors have demonstrated that both the size of the training set (Li *et al.*, 2015b), and similarity between training and test set structures (Kramer and Gedeck, 2010; Li and Yang, 2017; Li *et al.*, 2018), can influence scoring function performance using the PDBbind core set. We investigated the effect of three factors in training set composition on the performance of our models: training set size; similarity of ligands between training and test examples; and similarity of proteins between training examples.

To examine the effect of training set size, we simulated the effect of adding more structural and affinity training data over time by restricting the training set to annual releases of the PDBbind database from 2013 to 2018. Each release contains more data than the previous releases, so this results in six training sets of increasing size. By training separately on the general and refined sets of each year, we explore two different scenarios: a larger dataset of varying quality, and a smaller dataset with strict quality controls, giving a total of 12 distinct training scenarios.

To investigate the effect of including similar ligands in both the training and test sets, we used RDKit to compute the Tanimoto similarity between the Morgan fingerprints (radius 2 and 2048 bits) of each pair of ligands. We then constructed a new training set by removing from the available training data that any structure whose ligand had a Tanimoto similarity of $\geq$0.9 to any ligand in the test set.

To study the effect of including similar proteins in the training and test sets, for each version of PDBbind we constructed a series of training sets by removing from the original training set any structures with a protein sequence identity to any protein in the test set above a threshold of sequence identity. Clustering of the entire PDB using BLASTclust at sequence identity values from 30% to 100% computed by BLASTclust were downloaded from the PDB website (http://www.rcsb.org/pdb/statistics/clusterStatistics.do, accessed May 13, 2019). Finally, we construct an additional series of training sets by removing both structures with protein sequence identity above the cut-off value, and structures with ligand Tanimoto similarity greater than or equal to 0.9 to any structure in the test set.

When excluding similar test set complexes from the training data, we treated each test set separately. For example, when testing on the PDBbind 2016 core set, only proteins similar to those found in the 2016 core set were excluded from the training set. A detailed mathematical description of the training set construction is included in the Supplementary Material. Regardless of the choice of training and test set composition, all core set structures were always excluded from the training set.

2.5 Scoring function construction

For each of the four scoring functions considered (AutoDock Vina, RF-Score, RF-Score v3 and NNScore 2.0), we constructed two sets of features. The first used only the original features of the scoring functions, while the second used the original features of the scoring function, plus the 183 RDKit descriptors of the ligand. Since AutoDock Vina (and hence RF-Score v3 and NNScore 2.0) already use the number of rotatable bonds of the ligand, we dropped this from the set of RDKit descriptors added to avoid including the same feature twice. Finally, to examine whether there is any signal in the RDKit descriptors independent of the structure-based features, we constructed models using only the RDKit descriptors as a separate feature set, resulting in a total of nine different sets of features.

For each set of features and each training set, we built a scoring function by using (Breiman, 2001) RF regression to fit an estimator for the pK of a protein–ligand complex. We used the implementation of RF in the Python machine learning library scikit-learn (Pedregosa *et al.*, 2011). Although RF is generally robust with respect to hyperparameter choice, we tested the effect of varying the number of trees in the forest (*n_estimators*) and the maximum number of features considered at each split (*max_features*). We chose to set *n_estimators* = *500* and *max_features* = *0.33* as these values yielded optimal out-of-bag performance (Supplementary Figs S1 and S2). We tested several other ML methods, including regularized linear regression, single hidden-layer neural networks, AdaBoost and XGBoost, but found that RF consistently achieved the best cross-validation scores (Supplementary Fig. S3).

2.6 Investigating affinity predictions for ligands found in multiple structures

We investigated the affinity predictions for ligands that are found in multiple structures in the PDBbind database. We clustered the structures of the PDBbind 2018 general set by the three-character chemical ID of the ligand as specified in the PDB, and selected all ligands found in at least three structures. Holding out each ligand in turn as a test case, we trained a RF regression model using only the RDKit descriptors on all structures not containing that ligand, and used the resulting model to predict the affinity of that ligand.

We repeated the above process for each ligand by removing from the training set all ligands with a Tanimoto similarity to any test set ligand above a defined threshold. We then trained a new RF regression model and predicted the affinity of the test ligand. The Tanimoto similarity threshold in this process was reduced in steps of 0.1 from 0.9 to 0.1 inclusive.

3 Results and discussion

3.1 Ligand-based features improve ML scoring functions

Figure 1 shows the Pearson correlation coefficient between predicted and experimental pK achieved by each scoring function on 5-fold cross-validation, the PDBbind 2007, 2013 and 2016 core sets, and the combined core set, when trained on all data from the PDBbind 2018 general set not also found in any of the core sets. In all cases, a scoring function combining structure-based protein–ligand features with the RDKit molecular descriptors outperforms the corresponding scoring function using protein–ligand features alone. The difference between the Pearson correlation coefficient in each such case was found to be significant at 95% confidence (Mann–Whitney U-test $P < 0.05$), however difference in performance on the 2016 core set for scoring functions using the RF-Score v3 and NNScore 2.0 features is marginal (0.814 versus 0.821 and 0.819 versus 0.826). Our best-performing models are competitive with state-of-the-art ML scoring functions reported in the literature, such as PLECScore [up to $\rho_p = 0.83$ on the 2016 core set (Wójcikowski *et al.*, 2018)], K_{DEEP} [$\rho_p = 0.82$ on the 2016 core set (Jiménez *et al.*, 2018)] and RF-Score v3 [$\rho_p = 0.803$ on the 2007 core set (Li *et al.*, 2015a)]. In particular, the combination of AutoDock Vina terms and RDKit molecular descriptors is competitive with more sophisticated models, achieving $\rho_p = 0.840$, $\rho_p = 0.749$ and $\rho_p = 0.792$, respectively on the 2007, 2013 and 2016 core sets. Cross-validation performance of all models is lower than observed on the core sets. There are two factors that might contribute to this. First, the effective training set size during 5-fold cross-validation is 80% of the size of the full training set. Second, the PDBbind general set is far more

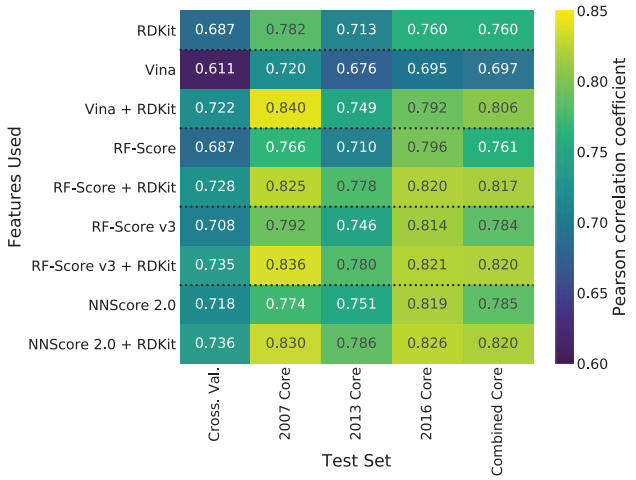

Fig. 1. Pearson correlation coefficient of predicted versus experimental binding affinity achieved by RF scoring functions when trained on the PDBbind 2018 general set and tested on different validation sets. In all cases, the RF scoring function combining a predominantly structure-based feature set with RDKit molecular descriptors consistently outperforms the corresponding RF using the structure-based feature set alone. This improvement is marginal on the 2016 core set when using the RF-Score v3 or NNScore 2.0 features; however, the AutoDock Vina features and RF-Score features still benefit from the addition of the RDKit molecular descriptors in this case. The cross-validation correlation coefficient is the mean value across a 5-fold cross-validation on the PDBbind 2018 general set. In all cases, the training set was identical and contained none of the core set complexes

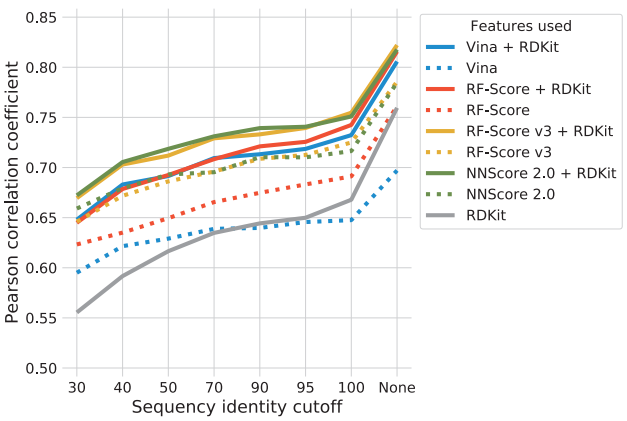

Fig. 2. Pearson correlation coefficient of predicted versus experimental binding affinity achieved by RF scoring functions when trained on the PDBbind 2018 general set and tested on the Combined Core set (Section 2.1). The sequence identity cut-off value above which proteins similar to those found in the combined core set were excluded from the training set is shown on the horizontal axis. Solid coloured lines denote RF scoring functions augmented with the RDKit molecular descriptors; dotted coloured lines denote RF scoring functions not augmented with the RDKit descriptors. The solid grey line corresponds to an RF using only the RDKit descriptors. Regardless of training set composition, the solid line is always above the dotted line of the same colour, indicating that models augmented with the RDKit ligand-based features outperform the corresponding (primarily) structure-based models without them

diverse than the core sets, including many examples of unique protein structures. Thus, under cross-validation, the test set contains a greater variety of structures many of which are not represented in the training data, and so we should expect performance to be lower. We obtained similar results when using XGBoost in place of RF (Supplementary Fig. S4).

Figure 2 shows how the Pearson correlation coefficient between experimental and predicted pK on the combined core set for our nine different scoring functions varies with the level of protein sequence identity allowed between the training and test sets. Regardless of training set construction, the addition of ligand-based features to a structure-based RF scoring function improves performance (in Fig. 2, for each colour, the solid line showing ligand-based plus structure-based features is consistently above the corresponding dotted line showing structure-based features alone). The trend toward a ligand-feature augmented scoring function outperforming the corresponding scoring function, using only structure-based features alone, is exemplified by the combination of AutoDock Vina terms and RDKit descriptors. This combination results in predictive performance comparable to that of the RF-Score, RF-Score v3 and NNScore 2.0 features, suggesting that a more complex and detailed set of features describing protein–ligand interactions is not necessarily more predictive than a comparatively simple set of force-field-like terms (AutoDock Vina terms) and molecular descriptors of the ligand (RDKit descriptors).

Exclusion from the training set of test-set similar proteins results in significantly reduced scoring function performance. There is a significant drop in the performance of all scoring functions even when a sequence identity cut-off of 100% is imposed, i.e. when only proteins with identical sequence to those in the test set are excluded from the training set. Reducing the sequence identity cut-off from 90% to 50% has a smaller impact on performance than the initial imposition of a 100% cut-off. Further reducing the cut-off from 50% to 30% has a more apparent effect. We found that the largest drop in training set size occurs when proteins with 100% sequence identity to those in the test set are excluded from the training set, with only a small decrease in training set size occurring when proteins with sequence identity of 90% or more are also excluded (Supplementary Fig. S5). This suggests that 'homology bias' due to the presence of similar proteins in the training and test sets is predominantly caused by the presence of many complexes of the same protein (100% sequence identity), rather than the presence of large numbers of similar or nearly identical proteins. This would explain the sharp drop in performance when excluding proteins from the training set at 100% sequence identity, since this eliminates almost all of the most similar structures to those in the test set, with subsequent, stricter sequence identity thresholds removing fewer less-similar structures.

There was no consistent improvement in performance when larger, more recent versions of the general set or refined set are used (Supplementary Figs S6 and S7). This is contrary to the results of Li *et al.* (2015b) who found that a larger training set resulted in improved performance. However, for a given release of PDBbind, a scoring function trained on the general set outperforms the same scoring function trained on the refined set (Supplementary Figs S8 and S9), consistent with the findings of Li *et al.* This difference in performance vanishes for all scoring functions when structures with test set similar ligands are excluded from the training set, and is greatly reduced when structures with 90% protein sequence identity to those in the test set are excluded. This suggests that the increase in performance when training on the general set can be attributed to increased representation of the core set proteins and ligands in the training data.

To better understand this result, we also investigated the performance of the scoring functions on bootstrapped samples of each version of the refined set when trained on the rest of the data from the same version of the refined set, and found that the mean correlation coefficient achieved by each scoring function did not vary with the version of the refined set used (Supplementary Figs S10 and S11). We verified that the number of distinct clusters of structures at different sequence identity thresholds increases with each release of PDBbind (Supplementary Fig. S12). The number of distinct clusters of ligands also increases with each release of PDBbind (Supplementary Fig. S13). This suggests that by using a larger, more diverse training set, the domain of applicability of the scoring function grows, allowing it to generalize to a more diverse test set without affecting performance. This is consistent with the fact that performance does not change when training on more recent versions of PDBbind.

The exclusion of structures containing test set similar ligands has a consistently deleterious effect on the performance of all scoring functions, regardless of whether they make use of the RDKit

descriptors, probably because many structures containing similar ligands also contain similar proteins. The bootstrapped 95% confidence intervals for the value of ρ_p when training on the PDBbind 2018 general set decrease when using a larger test set (Supplementary Fig. S14). For all scoring functions under all training and test scenarios, we reject the null hypothesis that the correlation between predicted and experimental affinity was due to random chance (permutation test $P < 0.05$).

Using a RF with only RDKit descriptors consistently results in a scoring function with greater performance than a RF with the AutoDock Vina terms, and is often close to or even exceeds the performance of RF-Score, even when test set similar ligands are excluded from the training data. This is surprising, since we would not expect to be able to predict protein–ligand binding affinity across a diverse set of protein–ligand complexes without knowing which protein the ligand was assayed with. We discuss the possible source and interpretation of this signal next, but this observation may suggest that the redundancy between the PDBbind core set and the remainder of the database makes this particular approach to training and test set construction inappropriate for validating and benchmarking machine learning scoring functions for predicting protein–ligand binding affinity.

3.2 Ligand-based features are predictive of mean binding affinity

Given the success of the RDKit descriptors alone (Fig. 2), we examined the affinity predictions for ligands that bind to proteins in multiple structures in the PDBbind database. Our RDKit RF model will produce only one value so it must be 'incorrect' for many of the protein–ligand complexes. When the RDKit RF model was tested on a previously unseen ligand, having been trained on all other data in the PDBbind 2018 general set, the score was found to be strongly correlated with the mean experimental pK of that ligand for its targets across the PDBbind 2018 general set ($\rho_p = 0.72$, Fig. 3). For the most common ligands in the PDBbind 2018 general set, the reported affinity values can span several orders of magnitude, so the RDKit RF model is not simply predicting a single 'correct' value for a ligand that happens to have many similar affinity measurements. Figure 4 shows the predictions of the RDKit RF model for the most common

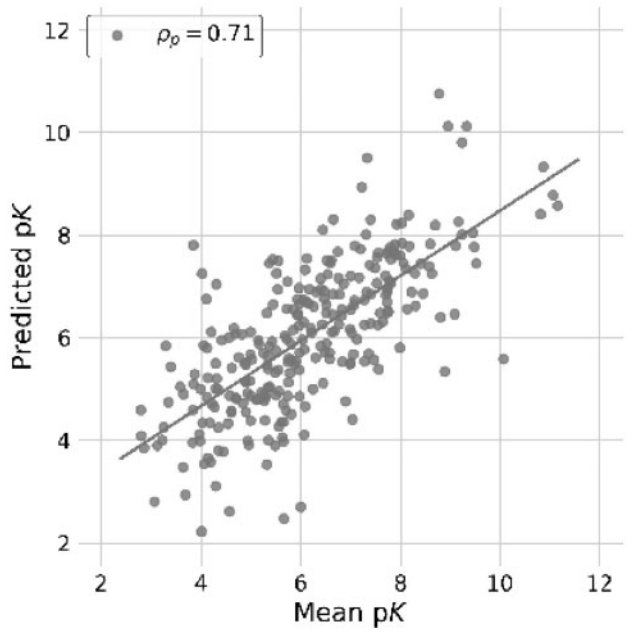

Fig. 3. Correlation of ligand-based affinity prediction with mean ligand affinity. Ligands with identical chemical ID were excluded from the training data. Pearson correlation coefficient $\rho_p = 0.71$. Each point represents one ligand. For each ligand, a new RF regression was trained

ligands in the PDBbind 2018 general set (markers) against the pK values reported by PDBbind (swarm plots). With the clear exception of biotin (BTN), the marker is often close to the centre of the swarm plot, indicating that the RDKit RF model appears to be predicting the mean affinity of the ligand even when the experimental data have a range of several pK units. We verified that many of the structures for biotin were biotin–streptavidin or biotin–avidin complexes, explaining the incredibly high experimental affinity values. Since ensemble-based methods such as RF cannot extrapolate beyond the range of values seen during training, it is unsurprising that the model cannot predict such a high average affinity when no such examples are included in the training set.

Further, we found that when nearly-identical ligands (Tanimoto similarity > 0.9) were excluded from the training data, this correlation is actually stronger ($\rho_p = 0.74$; Supplementary Fig. S15). Reducing the Tanimoto similarity threshold above which ligands are excluded from the training set results in gradually weakening correlation (Supplementary Fig. S15), suggesting that the RDKit RF model is not simply overfitting to the average affinity of highly similar training ligands.

Structure-based scoring functions should be able to differentiate between different complexes featuring the same ligand. To test this, we repeated the above experiment with each set of structure-based features, and computed the correlation coefficient between the predicted and experimental pK values for the structures featuring each ligand (Supplementary Fig. S16). In most cases, there is no correlation between the predicted and experimental pK, and there is no consistent trend toward a certain model performing well on certain sets of ligands. As an example of this behaviour, models using the RF-Score v3 and the RF-Score v3 + RDKit features were unable to accurately score the complexes featuring ADP (Supplementary Fig. S17).

3.3 Both structure-based and ligand-based features are important

The relative importance of the 20 highest-ranked ligand-based features for the RDKit RF model trained on the PDBbind 2018 general set is shown in Figure 5. The bulk properties such as molar refractivity (MolMR) and the logarithm of the octanol–water partition coefficient (MolLogP) are ranked highest. Molar refractivity captures the total polarizability of the molecule and log P captures its solubility; we might therefore expect these features to capture useful information about the ability of a small molecule to bind to a charged, buried active site. Both these properties are also used to characterize drugs (Ghose *et al.*, 1999) and log P is also used to predict bioavailability (Lipinski, 2004). It is possible that the predictive power of these features can in part be attributed to systematic bias in favour of crystallizing complexes featuring high-affinity engineered compounds. However, we found that there was no trivial correlation between either of these features and the pK of a compound ($\rho_p = 0.26$ and $\rho_p = 0.16$, respectively, across the PDBbind 2018 general set), suggesting that their contribution to the scoring function is only in concert with other features.

Perhaps easier to explain are features capturing molecular weight and charge (ExactMolWt, MaxAbsPartialCharge, MolWt, MinAbsPartialCharge, MaxPartialCharge) as the size and charge of the molecule will impose constraints on both its ability to fit within a binding pocket, its electrostatic complementarity and the number of interactions it has the ability to form. Similarly, TPSA is an approximation of a molecule's polar surface area computed using its 2D chemical graph, and may provide information about its hydrophobicity and ability to fit within a binding pocket. Van der Waals surface area contributions, captured by the PEOE_VSA descriptors, likewise characterize the molecular surface and hence potential interactions. More complicated are the 2D descriptors capturing molecular connectivity (Chi) and graph complexity (BertzCT), whose contribution to the model might also be through capturing the shape and surface area of the molecule or some aspect of conformational entropy.

We also found that when ligand-based features were combined with structure-based features in our other models, both ligand-based

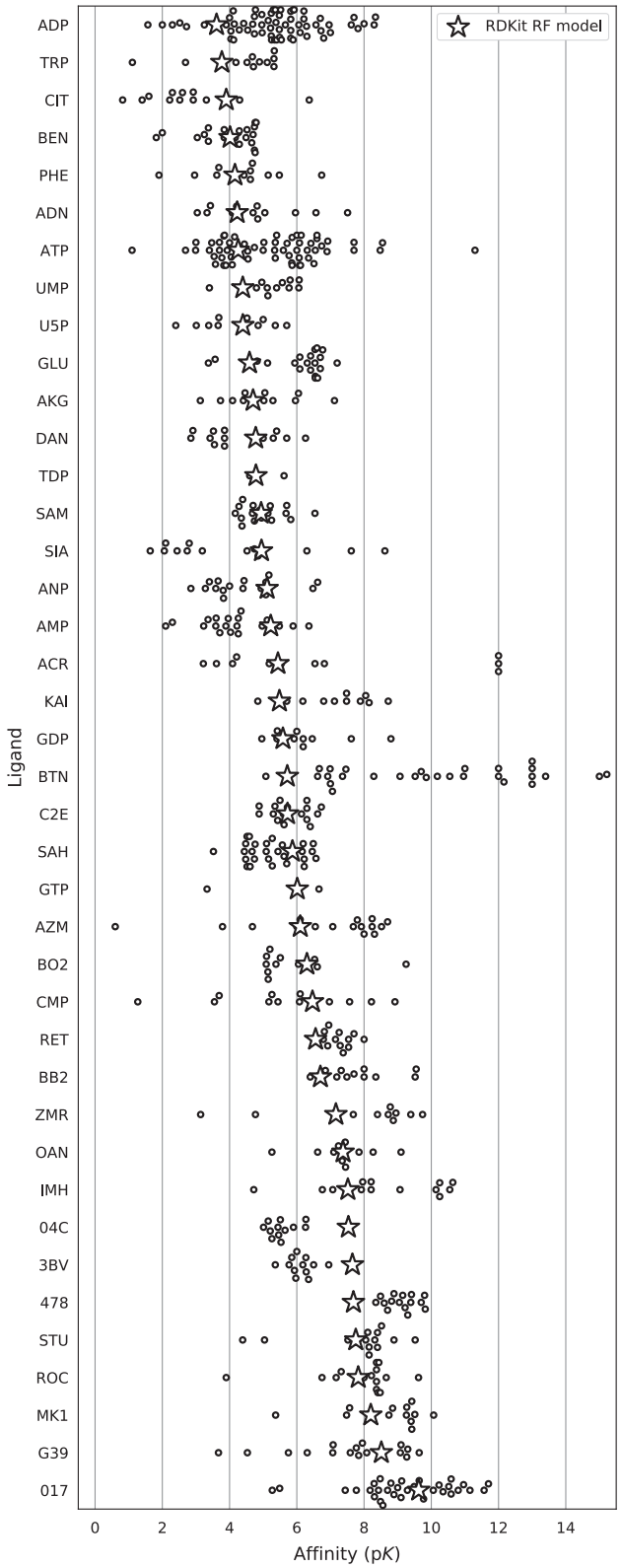

Fig. 4. Ligand feature-based affinity predictions (white star) against experimental values (dots) for the most common ligands in the PDBbind 2018 general set. For each ligand, a RF using only RDKit descriptors was trained on all complexes from the PDBbind 2018 general set that did not contain that ligand

and structure-based features were ranked highly, and that the same ligand-based features were consistently found to be important regardless of which structure-based features were used

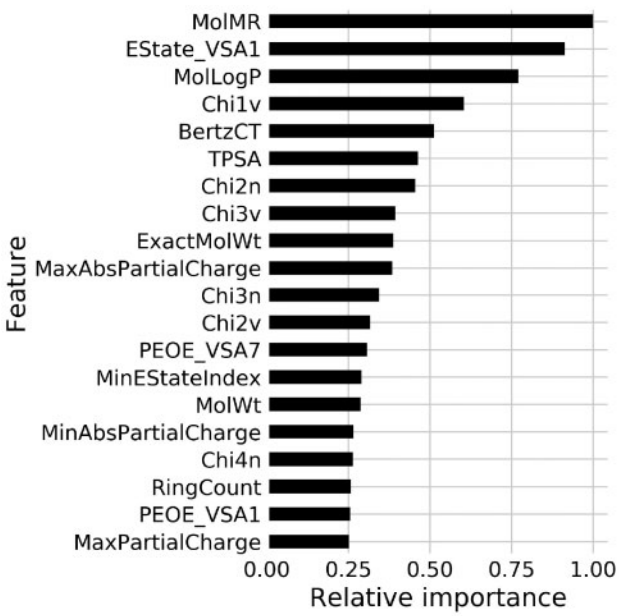

Fig. 5. Relative importance of features in the RDKit RF model trained on the PDBbind 2018 general set. A description of each feature is provided in the RDKit documentation (https://www.rdkit.org/docs/GettingStartedInPython.html, accessed May 17, 2019)

(Supplementary Figs S18 and S19). Finally, we investigated the correlation between structure-based and ligand-based features for the structures in the PDBbind 2018 general set (Supplementary Fig. S20) and found that the strongest correlations occur between pairs of structure-based features or pairs of ligand-based features, while weaker correlations exist between some pairs of structure-based and ligand-based features. This suggests that the ligand-based features are consistently capturing useful information that is not present in the structure-based features, beyond the number of rotatable bonds in the ligand. These results suggest that when using ML to predict protein–ligand binding affinity, it is better to use a richer description of the ligand in the model.

4 Conclusion

We have shown that the inclusion of a diverse set of readily-computed ligand-based features in machine learning scoring functions consistently improves their ability to rank ligands by their protein–ligand binding affinity.

Varying the composition of the training set chronologically, by restricting to data available only up until a particular year, had little effect on affinity predictions. This suggests an element of learning saturation for the targets tested with the data currently available. We showed that, in contrast, excluding proteins from the training set that are sequence similar to those in the test set has a deleterious effect on affinity predictions and that even excluding only those proteins with identical sequence to those in the test set leads to significantly reduced scoring function performance. We also showed that even when ligands with high Tanimoto similarity to those in the test set were excluded from the training set, the predictive power of the scoring functions was still increased by including ligand-based features.

Given the power of the ligand-based features, we investigated their predictive ability for ligands that bind to multiple targets and found that the predicted binding affinity of a model using only ligand-based features was strongly correlated with the mean of the experimental protein–ligand binding affinity of a ligand for its binding partners. This correlation remained strong when ligands with a Tanimoto similarity of greater than 0.9 to the test ligand were excluded from the training data. This correlation gradually

weakened when progressively less similar ligands were also excluded, suggesting that while the model's predictions are not reliant on overfitting to previously seen highly similar ligands, it does not extrapolate well to completely novel ligands.

Finally, we analysed the relative importance of the features of each scoring function. We found that when structure-based and ligand-based features are combined, both structure-based and ligand-based features were ranked highly, and that the same ligand-based features are ranked highly regardless of which structure-based features they were combined with. This suggests that the same information is consistently extracted from the ligand-based features and that this information is not redundant with that provided by structure-based features. Our results suggest that even under stringent validation, the addition of a diverse, quickly computed set of ligand-based features to a scoring function yields improved predictions of binding affinity.

Funding

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [EP/G03706X/1].

Conflict of Interest: none declared.

References

Abagyan,R. *et al*. (1994) Icm – a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem*., **15**, 488–506.

Ain,Q.U. *et al*. (2014) Modelling ligand selectivity of serine proteases using integrative proteochemometric approaches improves model performance and allows the multi-target dependent interpretation of features. *Integr. Biol*., **6**, 1023–1033.

Aldeghi,M. *et al*. (2016) Accurate calculation of the absolute free energy of binding for drug molecules. *Chem. Sci*., **7**, 207–218.

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**, 1169–1175.

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.

Breiman,L. (2001) Random forests. *Mach. Learn*., **45**, 5–32.

Cheng,T. *et al*. (2009) Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inform. Model*., **49**, 1079–1093.

Durrant,J.D. and McCammon,J.A. (2011) Nnscore 2.0: a neural-network receptor–ligand scoring function. *J. Chem. Inform. Model*., **51**, 2897–2903.

Friesner,R.A. *et al*. (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J. Med. Chem*., **47**, 1739–1749.

Ghose,A.K. *et al*. (1999) A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Combinat. Chem*., **1**, 55–68.

Gilson,M.K. and Zhou,H.X. (2007) Calculation of protein-ligand binding affinities. *Ann. Rev. Biophys. Biomol. Struct*., **36**, 21–42.

Halgren,T.A. *et al*. (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem*., **47**, 1750–1759.

Huang,S.-Y. *et al*. (2010) Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys*., **12**, 12899–12908.

Jain,A.N. (2003) Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem*., **46**, 499–511.

Jiménez,J. *et al*. (2018) K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *J. Chem. Inform. Model*., **58**, 287–296.

Jones,G. *et al*. (1997) Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol*., **267**, 727–748.

Kramer,C. and Gedeck,P. (2010) Leave-cluster-out cross-validation is appropriate for scoring functions derived from diverse protein data sets. *J. Chem. Inform. Model*., **50**, 1961–1969.

Li,H. *et al*. (2014a) Substituting random forest for multiple linear regression improves binding affinity prediction of scoring functions: cyscore as a case study. *BMC Bioinformatics*, **15**, 291.

Li,H. *et al*. (2015a) Improving autodock vina using random forest: the growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. *Mol. Informatics*, **34**, 115–126.

Li,H. *et al*. (2015b) Low-quality structural and interaction data improves binding affinity prediction via random forest. *Molecules*, **20**, 10947–10962.

Li,H. *et al*. (2018) The impact of protein structure and sequence similarity on the accuracy of machine-learning scoring functions for binding affinity prediction. *Biomolecules*, **8**, 12.

Li,Y. and Yang,J. (2017) Structural and sequence similarity makes a significant impact on machine-learning-based scoring functions for protein–ligand interactions. *J. Chem. Inform Model*., **57**, 1007–1012.

Li,Y. *et al*. (2014b) Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set. *J. Chem. Inform. Model*., **54**, 1700–1716.

Li,Y. *et al*. (2014c) Comparative assessment of scoring functions on an updated benchmark: 2. Evaluation methods and general results. *J. Chem. Inform. Model*., **54**, 1717–1736.

Lin,H. *et al*. (2013) A pharmacological organization of g protein–coupled receptors. *Nat. Methods*, **10**, 140.

Lipinski,C.A. (2004) Lead-and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol*., **1**, 337–341.

Liu,Z. *et al*. (2017) Forging the basis for developing protein–ligand interaction scoring functions. *Acc. Chem. Res*., **50**, 302–309.

Morris,G.M. *et al*. (2009) Autodock4 and autodocktools4: automated docking with selective receptor flexibility. *J. Comput. Chem*., **30**, 2785–2791.

Paricharak,S. *et al*. (2015) Proteochemometric modelling coupled to in silico target prediction: an integrated approach for the simultaneous prediction of polypharmacology and binding affinity/potency of small molecules. *J. Chemoinformatics*, **7**, 15.

Pedregosa,F. *et al*. (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res*., **12**, 2825–2830.

Perez,A. *et al*. (2016) Advances in free-energy-based simulations of protein folding and ligand binding. *Curr. Opin. Struct. Biol*., **36**, 25–31.

Rarey,M. *et al*. (1996) A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol*., **261**, 470–489.

Ravindranath,P.A. *et al*. (2015) Autodockfr: advances in protein-ligand docking with explicitly specified binding site flexibility. *PLoS Comput. Biol*., **11**, e1004586.

Ripphausen,P. *et al*. (2012) Analysis of structure-based virtual screening studies and characterization of identified active compounds. *Future Med. Chem*., **4**, 603–613.

Sotriffer,C.A. *et al*. (2008) Sfcscore: scoring functions for affinity prediction of protein–ligand complexes. *Proteins Struct. Funct. Bioinform*., **73**, 395–419.

Sousa,S.F. *et al*. (2006) Protein-ligand docking: current status and future challenges. *Proteins*, **65**, 15–26.

Sousa,S.F. *et al*. (2013) Protein-ligand docking in the new millennium – a retrospective of 10 years in the field. *Curr. Med. Chem*., **20**, 2296–2314.

Su,M. *et al*. (2018) Comparative assessment of scoring functions: the casf-2016 update. *Journal of Chemical Information and Modeling*,

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. *J. Comput. Chem*., **31**, 455–461.

van Westen,G.J. *et al*. (2011) Which compound to select in lead optimization? Prospectively validated proteochemometric models guide preclinical development. *PLoS One*, **6**, e27518.

Wójcikowski,M. *et al*. (2015) Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. *J. Cheminform*., **7**, 26.

Wójcikowski,M. *et al*. (2018) Development of a protein–ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics*, **35**, 1334–1341.

Zilian,D. and Sotriffer,C.A. (2013) Sfcscore rf: a random forest-based scoring function for improved affinity prediction of protein–ligand complexes. *J. Chem. Inform. Model*., **53**, 1923–1933.