Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity?

Journal:	*Journal of Chemical Information and Modeling*
Manuscript ID:	ci-2013-00703t.R1
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Ballester, Pedro; European Bioinformatics Institute, Schreyer, Adrian; University of Cambridge, Blundell, Tom L.; University of Cambridge, Department of Biochemistry

Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity?

Pedro J. Ballester[1,*], Adrian Schreyer[2] and Tom L. Blundell[2]

[1] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton - CB10 1SD, U.K.

[2] Dept. of Biochemistry, University of Cambridge, 80 Tennis Court Rd, Cambridge - CB2 1GA, U.K.

* Corresponding author: pedro.ballester@ebi.ac.uk

Abstract: Predicting the binding affinities of large sets of diverse molecules against a range of macromolecular targets is an extremely challenging task. The scoring functions that attempt such computational prediction are essential for exploiting and analysing the outputs of molecular docking, which is in turn an important tool for structural biology, chemical biology and drug design. Classical scoring functions assume a predetermined theory-inspired functional form for the relationship between the variables that characterise the X-ray crystal structure of the complex and its binding affinity. The inherent problem of this approach is in the difficulty of explicitly modelling the various contributions of intermolecular interactions to binding affinity.

New scoring functions based on machine-learning regression models, which are able to exploit effectively much larger amounts of experimental data and circumvent the need for a predetermined functional form, have already been shown to outperform a broad range of state-of-the-art scoring functions in a widely-used benchmark. Here we investigate the impact of the chemical description of the complex on the predictive power of the resulting scoring function using a systematic battery of numerical experiments. The latter resulted in the most accurate scoring function to date on the benchmark. Strikingly, we also found that a more precise chemical description of the protein-ligand complex does not generally lead to more accurate prediction of binding affinity. We discuss four factors that may contribute to this result: modelling assumptions; co-dependence of representation and regression; data restricted to the bound state; and conformational heterogeneity in data.

INTRODUCTION

Molecular docking plays a key role in addressing a number of important problems such as protein-function prediction[1,2] or drug-lead identification and optimisation[3,4]. Docking is a two-stage process: starting with the determination of the position, orientation and conformation of a molecule as docked to the target's binding site (pose generation) and followed by the prediction of how strongly the docked pose of such putative ligand binds to the target (scoring). Whereas there are many relatively robust and accurate algorithms for pose generation, the inaccuracies of current scoring functions continue to be the major limiting factor for the reliability of docking[5,6,7]. Indeed, despite intensive research over more than two decades, accurate prediction of the binding affinities of large sets of diverse protein-ligand complexes remains one of the most important open problems in computational bioscience.

Scoring functions are traditionally classified into three groups: force field[8,9], knowledge-based[10,11,12,13,14] and empirical[15,16,17,18,19]. For the sake of efficiency, scoring functions do not fully account for certain physical processes that are important for molecular recognition, which in turn limits their ability to rank-order and select small molecules by computed binding affinities. Two major sources of error in scoring functions arise from their limited description of protein flexibility and the implicit treatment of solvent. Instead of scoring functions, other computational methodologies based on molecular dynamics or Monte Carlo simulations can be used to model protein flexibility and desolvation upon binding. In principle, a more accurate prediction of binding affinity than that from scoring functions is obtained in those cases amenable to these techniques[20,21]. However, such expensive free energy calculations remain impractical for the evaluation of large numbers of protein–ligand complexes and their application is generally limited to predicting binding affinity in series of congeneric molecules binding to a single target[22].

In addition to these two enabling simplifications, there is an important computational issue in scoring function development that has received little attention until recently[23]. Each scoring function assumes a predetermined theory-inspired functional form for the relationship between the variables that characterise the complex, which also includes a set of parameters that are fitted to experimental or simulation data, and its predicted binding affinity. Such a relationship can take the form of a sum of weighted physicochemical contributions to binding in the case of empirical scoring functions or a reverse Boltzmann methodology in the case of knowledge-based scoring functions. The inherent drawback of this rigid approach is that it leads to poor predictivity in those complexes that do not conform to the modelling assumptions.

As an alternative to these classical scoring functions, nonparametric machine learning can be used to capture implicitly the binding interactions that are hard to model explicitly. By not imposing a particular functional form for the scoring function, the collective effect of intermolecular interactions in binding can be directly inferred from experimental data, which should lead to scoring functions with greater generality and prediction accuracy. This unconstrained approach should result in performance improvement, as it is well known that the strong assumption of a predetermined functional form for a scoring function constitutes an additional source of error (e.g. imposing an additive form for the considered energetic contributions[24]). On the other hand, recent experimental results have resulted in novelties in the definition of molecular interactions such as the hydrogen bond[25] and the hydrophobic interaction[26], implying that previously proposed expressions for these energetic contributions might need to be revised accordingly.

While a few classifiers exploiting X-ray crystal structural data for discriminating between binders and non-binders of a protein target have been presented[27,28], it is only recently that machine learning for nonlinear regression has been shown[23,29] to be a particularly powerful approach to build generic scoring functions. This approach has been highlighted[30,31,32,33] as very promising for the improvement of scoring functions. Indeed, a growing number of studies showing the benefits of machine learning scoring functions have been presented[23,29,34,35,36,37]. However, these initial models are relatively coarse in the description of the complex and thus the question remains as to whether the incorporation of additional chemical information relevant for binding would improve performance further.

Here we investigate the impact of a more precise chemical characterisation of the protein-ligand complex on the predictive power of the resulting scoring function. This includes the use of structural interaction fingerprints[38], using atom and interaction type definitions from the CREDO structural interactomics database[39]. We show that the new version of RF-Score performs much better than classical scoring functions on the same test set. The RF-Score performed best when describing a complex using a 12Å distance cutoff between atom pairs, suggesting that there is a minor contribution from long-range atom pairs. In the light of the improved performance obtained and considering the uncertainty introduced by the static nature of crystal structures, we discuss the role of interatomic distance cutoffs and binning as well as protonation states in binding affinity prediction. As a by-product of this systematic battery of numerical experiments, the most accurate scoring function to date on a widely used pre-existing benchmark is presented. An important conclusion of our study is that a more chemically precise description of the protein-ligand complex does not generally lead to more accurate prediction of binding affinity. We discuss four factors that may contribute to this result:

modeling assumptions; co-dependence of representation and regression; data restricted to the bound state; and conformational heterogeneity in data.

METHODS

Defining descriptors

Each complex was described by a vector of integer-valued descriptors or features. Three description schemes were implemented: the *Element* scheme uses the combination of the element symbols of the interacting atoms to classify the interaction, e.g. C-C or N-O. The fingerprint of this scheme has a position for each pairwise combination of element symbols and the directionality is preserved, i.e. N-O is distinct from O-N. Here all the heavy atoms commonly observed in PDB complexes (C, N, O, F, P, S, Cl, Br, I) are considered.

The *Sybyl* scheme uses SYBYL atom types instead of the element symbols to define the range of considered protein-ligand atom pairs. These atom types permit deconvoluting the element into hybridisation state and bonding environment. For instance, instead of having a single C element atom type, the *Sybyl* scheme considers the following subtypes: C+, C1, C2, C3, Cac and Car (a description of SYBYL atom types can be found at http://www.tripos.com/mol2/atom_types.html). The latter leads to 36 distinct C-C descriptors in the *Sybyl* scheme in contrast to a single C-C descriptor in the *Element* scheme.

The credo scheme uses *Structural Interaction Fingerprints* (SIFts)[38] to encode protein-ligand interactions. Here, interatomic pairs are categorised as interactions if particular geometrical and atom type constraints are satisfied. Atom types were defined through a set of SMARTS patterns that are completely customisable through a configuration file. The atom types of atoms belonging to standard amino acids as well as non-standard binding site residues that occurred in the used test sets were pre-calculated because determining them “on the fly” was not feasible with the Open Babel toolkit. These atom types were stored in a separate configuration file and can therefore be easily changed by the user. The determination of standard and weak hydrogen bonds required the protonation state to be known and there the complexes to re-score must be already protonated. 12 different contact types are encoded in the SIFt of which four are solely distance-based. The latter are covalent bonds, van der Waals clashes & contacts and finally proximal interactions. The other eight “feature” contact types are hydrogen bonds, weak hydrogen bonds, halogen bonds, ionic, metal complexes, aromatic, hydrophobic and carbonyl. The definition of these including the source of the SMARTS definitions and other constraints are described in the original CREDO publication[39] with the following exception: the carbonyl interaction type has since then been implemented based on the *ab initio* molecular-orbital calculations by Allen *et al.*[40] who have shown that carbonyl-carbonyl interactions can have similar strengths to those from hydrogen bonds. Appendix A3 in the Supplementary Information summarises the exact classification scheme for all interactions that was used for *SIFt* descriptor.

Once the scheme is selected, descriptors are generated by counting interatomic pairs between protein binding sites and their ligand molecules. These possibly interacting atoms were assigned a specific interaction type if their distance was within a distance threshold and if a combination of possible atom type and geometry constraints were satisfied. The software to calculate these three description schemes uses the open source chemistry toolkit Open Babel[41] in version 2.3.0 (through Python bindings) and the SciPy library[42]. The molecular structures of protein targets and their ligands comprising the used data sets from PDBbind are read-in separately. Interatomic contacts for a given distance cutoff are determined using the KDTree structure in the scipy.spatial module with the atomic coordinates as input (the KDTree is an efficient space-partitioning algorithm that limits the search space for inter-atomic contacts in order to prevent expensive all-by-all searches). The interacting atoms are then analysed depending on the descriptor and the appropriate position on the fingerprint incremented.

This descriptor-generating software is also capable of binning the identified interatomic pairs into arbitrary distance ranges. For each normal feature on a descriptor, a number of columns equal to the number of distance bins are created. Using a distance cutoff of 6Å and a bin size of 1Å for example would create six bins for each feature: from 0-1Å, 1-2Å and so on. The correct bin for each interatomic pair that has to be incremented is determined using the numpy.digitize function in the NumPy package (http://numpy.scipy.org).

The complete source code of the software that was used to generate the described results was released at https://bitbucket.org/aschreyer/rfscore under the MIT license.

Regression model

RF-Score uses Random Forest[42] (RF) as the regression model. A RF is an ensemble of many different decision trees randomly generated from the same training data. RF trains its constituent trees using the CART algorithm. As the learning ability of an ensemble of trees improves with the diversity of the trees, RF promotes diverse trees by introducing the following modifications in tree training. First, instead of using the same data, RF grows each tree without pruning from a bootstrap sample of the training data (i.e. a new set of N complexes is randomly selected with replacement from the N training complexes, so that each tree grows to learn a closely related but slightly different version of the training data). Second, instead of using all features, RF selects the best split at each node of the tree from a typically small number (m_{try}) of randomly chosen features.

This subset changes at each node, but the same value of m_{try} is used for every node of each of the P trees in the ensemble. RF performance does not vary significantly with P beyond a certain threshold and thus P=500 was set as a sufficiently large number of trees. In contrast, m_{try} has some influence on performance and thus constitutes the only tuning parameter of the RF algorithm. In regression problems, the RF prediction is given by arithmetic mean of all the individual tree predictions in the forest. The performance of each tree on predicting Out-Of-Bag (OOB) data, that is complexes not selected in the bootstrap sample and thus not used to grow that tree, gives an internal validation of RF. OOB is a fast resampling strategy carried out in parallel to RF training that yields estimates of prediction accuracy that are very similar to those derived from more computationally expensive k-fold cross-validations. The RF-Score software is available at http://pedroballester.com/software.

Training and Test Data

The PDBbind benchmark[41] is an excellent choice for validating generic scoring functions. It is based on the 2007 version of the PDBbind database, which contains a particularly diverse collection of protein-ligand complexes, assembled through a systematic mining of the entire Protein Data Bank. The first construction step was to identify all the crystal structures formed exclusively by protein and ligand molecules. This excluded protein-protein and protein-nucleic acid complexes, but not oligopeptide ligands as they do not normally form stable secondary structures by themselves and therefore may be considered as common organic molecules. Secondly, Wang et al. collected binding affinity data for these complexes from the literature. Emphasis was placed on reliability, as the PDBbind curators manually reviewed all binding affinities from the corresponding primary journal reference in the PDB.

In order to generate a refined set suitable for validating scoring functions, the following data requirements were additionally imposed. First, only complete and binary complex structures with a resolution of 2.5Å or better were considered. Second, complexes were required to be non-covalently bound and without serious steric clashes. Third, only high quality binding data were included. In particular, only complexes with known K_d or K_i were considered, leaving those complexes with assay-dependent IC_{50} measurements out of the refined set. Also, because not all molecular modelling software can handle ligands with uncommon elements, only complexes with ligand molecules containing just the common heavy atoms (C, N, O, F, P, S, Cl, Br, I) were considered. In the 2007 PDBbind release, this process led to a refined set of 1300 protein-ligand complexes with their corresponding binding affinities. Still, the refined set contains a higher proportion of complexes belonging to protein families that are overrepresented in the PDB. This was considered detrimental to the goal of identifying those generic scoring functions that will perform best over all known protein families. To minimise this bias, a core set was generated by clustering the refined set according to BLAST sequence similarity (a total of 65 clusters were obtained using a 90% similarity cutoff). For each cluster, the three complexes with the highest, median and lowest binding affinity were selected, so that the resulting set had a broad and fairly uniform binding affinity coverage. By construction, this core set is a large, diverse, reliable and high quality set of protein-ligand complexes suitable for validating scoring functions. The PDBbind benchmark essentially consists of testing the predictions of scoring functions on the 2007 core set, which comprises 195 diverse complexes with measured binding affinities spanning more than 12 orders of magnitude. The protonation states of both proteins and ligands were already calculated by PDBbind developers.

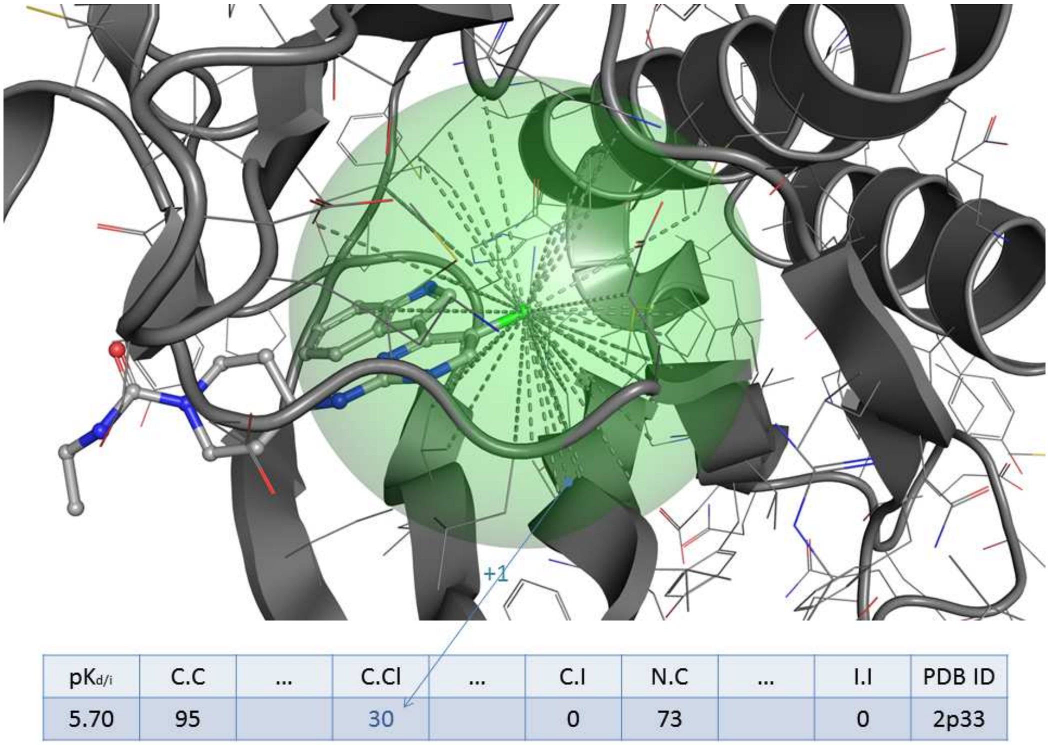
Lastly, just as in Cheng et al.[41], the 1105 complexes in the PDBbind 2007 refined set that are not in the core set will be used as the training set, whereas the core set of 195 complexes will be used as the independent test set. In this way, a set of protein-ligand complexes with measured binding affinity can be processed to give two non-overlapping data sets, where each complex is represented by its feature vector $\vec{x}^{(n)}$ and its binding affinity $y^{(n)}$:

$$D_{train} = \left\{\left(y^{(n)}, \vec{x}^{(n)}\right)\right\}_{n=1}^{1105}; D_{test} = \left\{\left(y^{(n)}, \vec{x}^{(n)}\right)\right\}_{n=1106}^{1300}; y \equiv -\log_{10} K$$

RESULTS

Preamble

We start this section with a concise description of the approach to predicting binding affinity using machine learning[23]. This process starts with the characterisation of each protein-ligand complex as a set of intermolecular features or descriptors relevant to binding affinity prediction. The sketch in Figure 1 shows an example of how descriptors are generated from the X-ray crystal structure of a complex (PDB:2p33). Each descriptor is given by the occurrence count of a particular protein-ligand atom pair within a predetermined distance range of each other. For example, a descriptor could be defined as the number of times that protein nitrogen and ligand oxygen are separated by less than a distance cutoff (d_{cutoff}). As in previous studies[23], nine atom types commonly observed in PDB complexes were selected by considering atomic number only (C, N, O, F, P, S, Cl, Br, I) to give rise to 81 protein-ligand atom pairs which are considered as descriptors (this descriptor scheme is named here ‘element’).

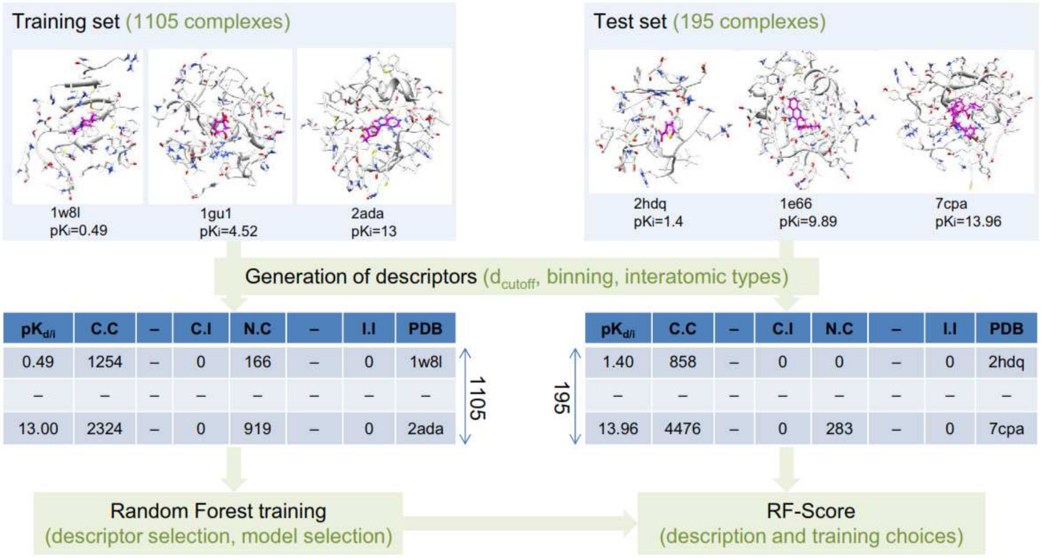


pK$_{d/i}$	C.C	...	C.Cl	...	C.I	N.C	...	I.I	PDB ID
5.70	95		30		0	73		0	2p33

Figure 1: Sketch of the process of characterising a protein-ligand complex (PDB: 2p33) as a set of structure-derived descriptors (C.C to I.I). The discontinuous green lines connect the ligand chlorine atom with all protein carbon atoms within the distance cutoff represented by the green sphere, with the number of these pairs giving value to the C.Cl descriptor. The rest of the descriptors are calculated in an analogous manner.

Once the descriptor scheme (i.e. considered atom types, binning and cutoff) is chosen, the next step is to select a source of curated structural and interaction data suitable for training and testing scoring functions. The PDBbind database[43] is an excellent choice for this purpose, with the additional advantage that a large number of scoring functions have already been benchmarked on a common PDBbind test set[44], which permits comparing new developments against the state of the art. Moreover, some scoring functions[23,41] have not only been tested on this common dataset but also calibrated on the same training set (further details can be found in the Methods section). This is important to avoid the often large bias introduced by using a different training set for each scoring function (such bias makes comparisons among scoring functions unreliable, even if compared on the same test set[29]). Therefore, we will be focusing here on these common training and test sets.

Lastly, a regression model is needed to predict binding affinity of test set complexes from the structural and interaction data in the training set. Here we build upon RF-Score[23], a machine learning scoring function using Random Forest (RF)[45] for regression. RF is typically tuned using a single control parameter (m_{try}, which controls the number of features that are considered for the split at each tree node) and may be subjected to a feature selection strategy intended to remove descriptors with low information content as a way to improve performance (that is, in addition to the common practice of removing all those descriptors that have zero values across training complexes). As usual, predictive performance is measured as the difference between predicted and measured binding affinity across test set complexes. Figure 2 summarises the process of training and testing machine learning scoring functions. Full details on descriptors, data and regression protocols can be found in the Methods section.


Training set (1105 complexes)
1w8l
pKi=0.49
1gu1
pKi=4.52
2ada
pKi=13
Test set (195 complexes)
2hdq
pKi=1.4
1e66
pKi=9.89
7cpa
pKi=13.96
Generation of descriptors (d_{cutoff}, binning, interatomic types)
1105
195
Random Forest training
(descriptor selection, model selection)
RF-Score
(description and training choices)

$pK_{d/i}$	C.C	–	C.I	N.C	–	I.I	PDB
0.49	1254	–	0	166	–	0	1w8l
–	–	–	–	–	–	–	–
13.00	2324	–	0	919	–	0	2ada

$pK_{d/i}$	C.C	–	C.I	N.C	–	I.I	PDB
1.40	858	–	0	0	–	0	2hdq
–	–	–	–	–	–	–	–
13.96	4476	–	0	283	–	0	7cpa

Figure 2: Training and testing RF-Score workflow. Top: descriptors are generated from two non-overlapping data sets with 1105 and 195 complexes for training and testing respectively. Bottom: training Random Forest to learn the non-linear relationship

between this atomic-level description of the complex and its binding affinity (pK_d or pK_i; $pK_{d/i}$ denotes both without distinction). The resulting scoring function (RF-Score) is used to predict the binding affinity of the test set.

We have seen that not only can a protein-ligand complex be described in various ways, but also the process of building a predictive model from these descriptors involves a number of choices. In a way, the overall process of training a scoring function can be regarded as a quest for finding an optimal combination of these design variables. An exhaustive evaluation of all possible combinations is impractical, as this would involve a prohibitively large number of RF training runs (2^m runs, one for each possible subset of m features, even if we fix m_{try} to its recommended value). Thus, we were forced to assume the independency of these design variables and searched for the optimal value of each variable in a sequential manner as explained in the next subsections.

Optimal interatomic distance cutoff

The first question we addressed was which distance cutoff leads to the best performance on the independent test set (henceforth referred to as simply the test set). Different distance cutoffs have been previously used in the literature, some as large as 12Å (PMF[12]) and 15Å (Fresno[46]). Here we addressed this question empirically by generating element descriptors for four cutoffs (6Å, 9Å, 12Å and 15Å), which gave rise to four different numerical characterisations of the training set with their corresponding test set counterparts. Thereafter, RF was calibrated on each of these training set versions and the resulting model used to predict the binding affinity of the corresponding test set complexes. As scoring function calibration is a stochastic process, a slightly different model is obtained with a different random seed. To assess the variability in RF prediction due to this factor, we repeated the training 10 times for each cutoff, each time with a different random seed. Such assessment is needed to establish whether the improvement in prediction is due to using another distance cutoff or just comes from variability in model calibration (this procedure is more accurate than basing the analysis on a single model calibration as has been the case so far). Lastly, we considered three commonly used metrics for quantifying the difference between predicted and measured binding affinity across the test set of protein-ligand complexes: R_p (Pearson's correlation coefficient), R_s (Spearman's correlation coefficient) and SD (Standard Deviation in log $K_{d/i}$ units). Figure 3 illustrates the results of this numerical experiment.

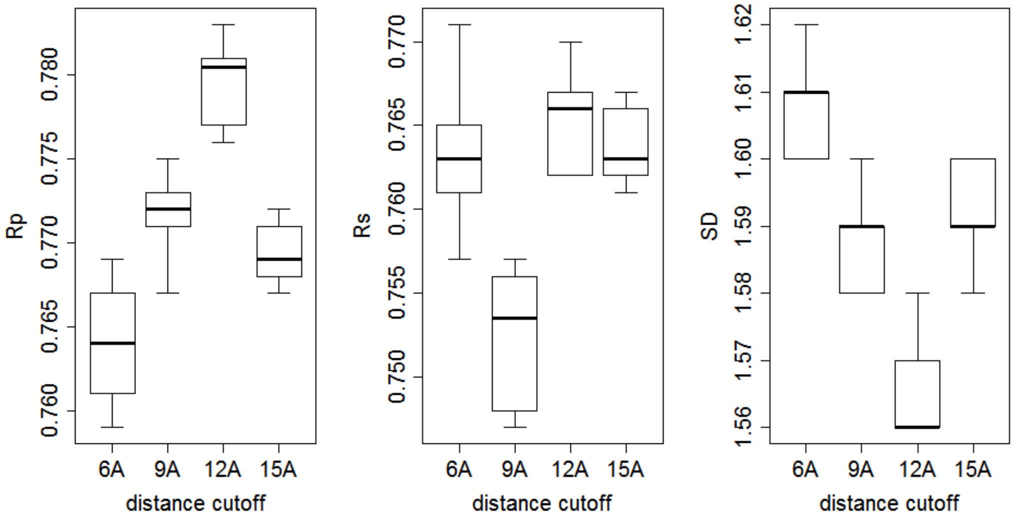


Figure 3: Test set performance of RF-Score with element descriptors for each of the four distance cutoffs (6Å, 9Å, 12Å and 15Å). Performance is measured as the difference between observed and predicted binding affinity in the test set using three metrics: Pearson's correlation coefficient (Rp; left plot), Spearman rank-correlation coefficient (Rs; middle plot) and standard deviation (SD; right plot). Ten models are built from each of these four versions of the training data sets (6Å, 9Å, 12Å and 15Å), each time using a different random seed (the boxplot summarises the performance on the test set achieved by each of the 10 models). Results showed that the best median performance, i.e. that with the highest correlations and lowest standard deviation, is obtained with the 12Å cutoff in all three performance metrics. It is worth noting that optimising the distance cutoff only led to a modest performance improvement (+0.017 in median Rp and -0.05 log K units in median SD).

Role of protonation states and bonding neighbourhood

The element descriptors used in the previous experiment constitute a coarse representation of the complex. Distinguishing between atoms of the same element in different local environments leads to a more chemically precise characterisation (e.g. deconvoluting the occurrence counts of a carbon-carbon intermolecular pair into pairs that specify the hybridisation state of both atoms) and thus the resulting model would in principle be expected to perform better. To test this hypothesis we used Sybyl atom types with these characteristics. Further, one could also incorporate additional information into the descriptors such as the angle between hydrogen bond donors, acceptors and hydrogen atoms as well as covalent and van der Waals radius. These are the Credo descriptors, which measure the abundance of a range of intermolecular interactions such as hydrogen bonds, hydrophobic interactions or van der Waals clashes. Using the same training and test set, each description scheme gives rise to a different set of features that are used to characterise every complex. The performance of each of these three description schemes is presented in Figure 4.

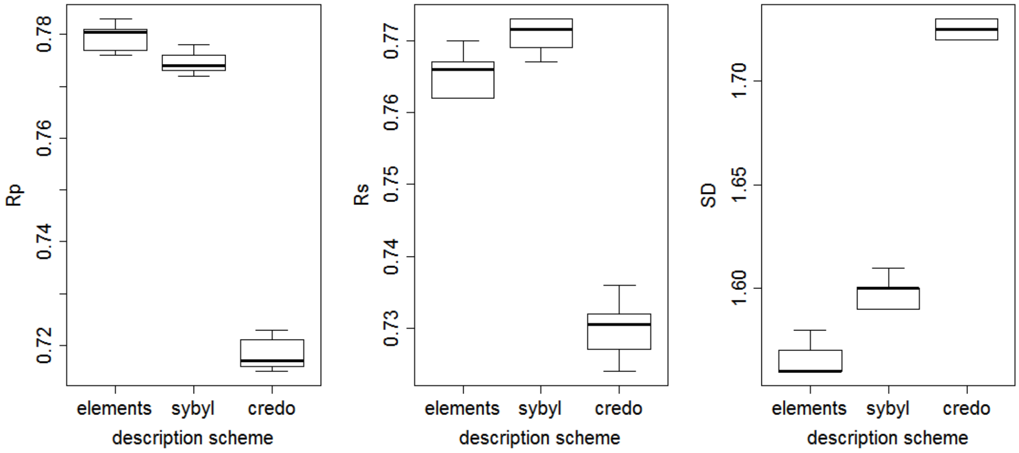


Figure 4: Test set performance of RF-Score using the optimal 12Å interatomic distance cutoff with element, Sybyl and credo descriptors. Interestingly, the model based on Credo descriptors obtained much lower performance than that using Sybyl and element descriptors. Element descriptors led to a small improvement over Sybyl descriptors. These results hint a trade-off between the predictability and interpretability of the model, which we will discuss later in this paper.

Incorporating interatomic distance

The strength of the interatomic interactions that collectively form the noncovalent intermolecular bond depends on the separation between the interacting atoms. Therefore, it is reasonable to think that partitioning the descriptors into a number of interatomic distance bins should lead to a model with more predictivity. Consequently, we generated element descriptors with 12Å cutoff, i.e. using all the optimal values, for six bin sizes (a 12Å bin size with a 12Å cutoff simply corresponds to the case without binning, which was previously shown in the first boxplot in Fig. 4 and the third boxplot in Fig. 3). Figure 5 shows the results for each bin size, where the best median performance is achieved by models with lower bin sizes (1Å, 2Å and 3Å), representing a moderate improvement with respect to the model based on single-binned descriptors (12Å). The experiments were repeated using the same bin sizes but now with Sybyl and Credo descriptors instead of elements descriptors (the maximum cutoff for a Credo interaction type is 4.5Å, all other atom pairs in this description scheme are labelled as “proximal”). It was observed that the performance was not as high as that with element descriptors (the best median performance for Sybyl was Rp=0.779, Rs=0.771 and SD=1.59, whereas that for Credo was Rp=0.739, Rs=0.742 and SD=1.68).

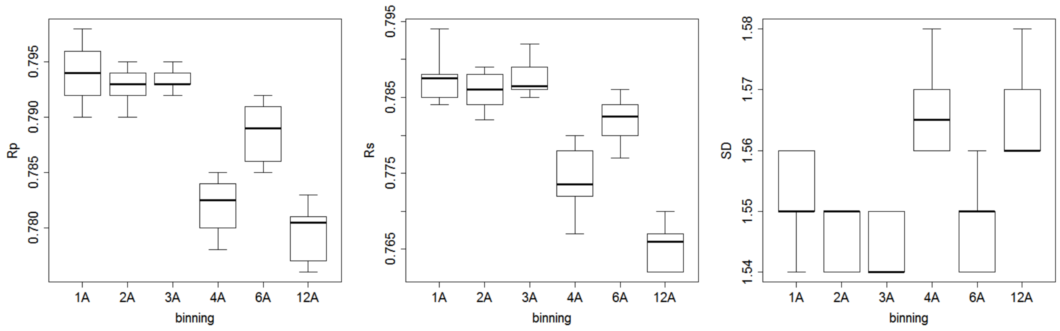


Figure 5: Test set performance of RF-Score with element descriptors and 12Å interatomic distance cutoff using six different bin sizes (1Å, 2Å, 3Å, 4Å, 6Å and 12Å). The best median performance is achieved by models with lower bin sizes (1Å, 2Å and 3Å), representing a modest improvement with respect to the model based on single-binned descriptors (12Å).

Feature selection

In addition to exploring the impact of different ways to describe the complexes, we also applied basic feature selection strategies intended to remove sparse features that increased the complexity of the model without improving performance. Here the sparsity (spr) of a descriptor is defined as the average number of occurrence counts per training complex. In previous versions of RF-Score, only features with sparsity higher than the zero threshold (spr=0), i.e. those that are non-zero for at least one training complex, were considered. Here we also considered two additional sparsity thresholds (spr=1 and spr=2). We conducted this experiment for the three best bin sizes in Figure 5 (1Å, 2Å and 3Å), which had spr=1 as the optimal value on all three bin sizes. Figure 6 presents the results for the best bin size across the three spr values (2Å).

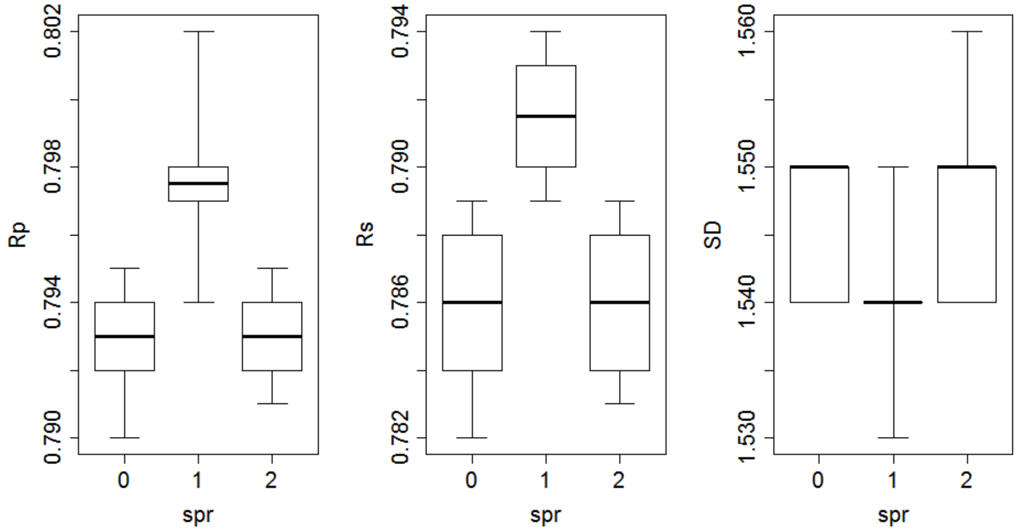


Figure 6: Test set performance of RF-Score with element descriptors, 12Å cutoff and 2Å bin size using three values of the feature selection threshold (spr). Best median

performance is obtained by spr=1, which corresponds to only considering descriptors that have an average of at least one atom-atom pair in the considered distance range per training complex. The latter represents a moderate improvement with respect to the models using descriptors with spr=0 and spr=2.

Model selection

Lastly, we have been using the recommended value for the RF m_{try} parameter so far. However, interval validation strategies can be used to select an optimal value for m_{try}. One of these strategies is called Out-Of-Bag (OOB) validation and essentially consists in training the model for each possible value of m_{try} and select the model that best performs on the internal validation set (a subset of the training set, as further explained in the Methods section). Figure 7 shows that this model selection strategy carries a small improvement in performance at the cost of much higher computational expense in model selection (one RF training run per considered m_{try} value).

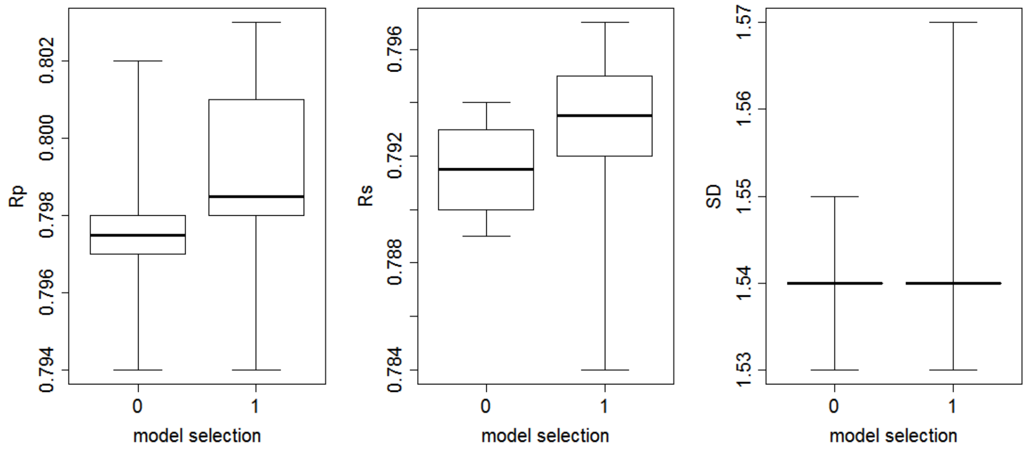


Figure 7: Test set performance of RF-Score with element descriptors, 12Å cutoff, 2Å bin size and feature selection threshold (spr=1) for the recommended m_{try} (model selection=0) and m_{try} selected by OOB validation (model selection=1), which requires 123 times more training than just using the recommended setting (as many RF training runs as features were selected to describe each complex).

Predictive performance

This systematic battery of numerical experiments led to the new scoring function RF-Score::Elem(c12,b2)_spr1_oob (RF-Score::Elem-v2 for short). As we have seen, the descriptors come from partitioning occurrence counts of each element atom type pairs into six interatomic distance bins of 2Å size and the RF model is built with the 123 descriptors that are sufficiently dense (spr=1) using the internally validated m_{try} value (m_{try}=14 in this case). Figure 8 shows the predictive power of RF-Score::Elem-v2 on the test set.

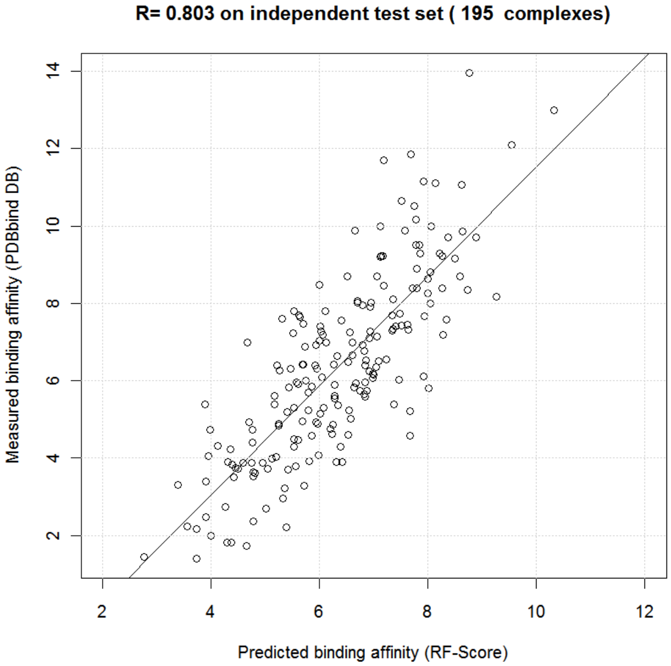


Figure 8: RF-Score::Elem-v2 predicted versus measured binding affinity on the independent test set (195 complexes). Pearson's correlation coefficient R_p=0.803, Spearman's correlation coefficient R_s=0.797, standard deviation SD =1.54 log $K_{d/i}$ units and Root Mean Square Error RMSE =1.53 log $K_{d/i}$ units. This plot can be visually compared to those for the best performing scoring functions in Cheng et al.[41]'s figure 6. Performance comparisons on the same test set are presented here in figure 9.

In terms of efficiency, RF-Score::Elem-v2 scored all 195 protein-ligand complexes in 0.01 seconds (all the computation in this study was carried out with a single processing core Intel Core i7-2920XM at 2.50GHz with 16GB RAM). In addition, the time to generate these features for the 195 complexes was 8 seconds and hence this is the most expensive part of the calculation. Therefore, the average time to score one protein-ligand complex is about 0.04 seconds if the features have not been calculated before, which makes RF-Score suitable to re-score a high number of docking poses in virtual screening applications.

The predictive power of RF-Score::Elem-v2 was also compared against that of a wide selection of scoring functions on the PDBbind benchmark[41]. By using a pre-existing benchmark, the danger of constructing a benchmark complementary to the presented scoring function is avoided. It also has the advantage of ensuring that previously tested scoring functions were provided with optimal settings by their authors. Several of the scoring functions tested in the PDBbind benchmark have different versions or multiple options. However, for the sake of practicality, only the version/option of each scoring function that performs best on the PDBbind benchmark was reported by Cheng et al.[41]. In addition to these 16 scoring functions, we also tested a more recent function called IMP::RankScore[47]. Figure 9 reports the performance of all scoring functions on the test set, with RF-Score::Elem-v2 obtaining the best performance with R_p=0.803 (the performance of the original version of RF-Score[23] is also included). In contrast, classical scoring functions tested on the same test set obtained a lower R_p spanning from 0.216 to 0.644. This trend was also observed with the other two performance measures (R_s, SD). It is worth noting that the root-mean-square error of the free energy of binding on such a diverse test set is just 2.1 kcal/mol.

scoring function	R_p	R_s	SD
RF-Score::Elem-v2	0.803	0.797	1.54
RF-Score::Elem-v1	0.776	0.762	1.58
X-Score::HMScore	0.644	0.705	1.83
DrugScore^{CSD}	0.569	0.627	1.96
SYBYL::ChemScore	0.555	0.585	1.98
DS::PLP1	0.545	0.588	2.00
GOLD::ASP	0.534	0.577	2.02
SYBYL::G-Score	0.492	0.536	2.08
DS::LUDI3	0.487	0.478	2.09
DS::LigScore2	0.464	0.507	2.12
GlideScore-XP	0.457	0.435	2.14
DS::PMF	0.445	0.448	2.14
GOLD::ChemScore	0.441	0.452	2.15
by NHA	0.431	0.517	2.15
SYBYL::D-Score	0.392	0.447	2.19
IMP::RankScore	0.322	0.348	2.25
DS::Jain	0.316	0.346	2.24
GOLD::GoldScore	0.295	0.322	2.29
SYBYL::PMF-Score	0.268	0.273	2.29
SYBYL::F-Score	0.216	0.243	2.35

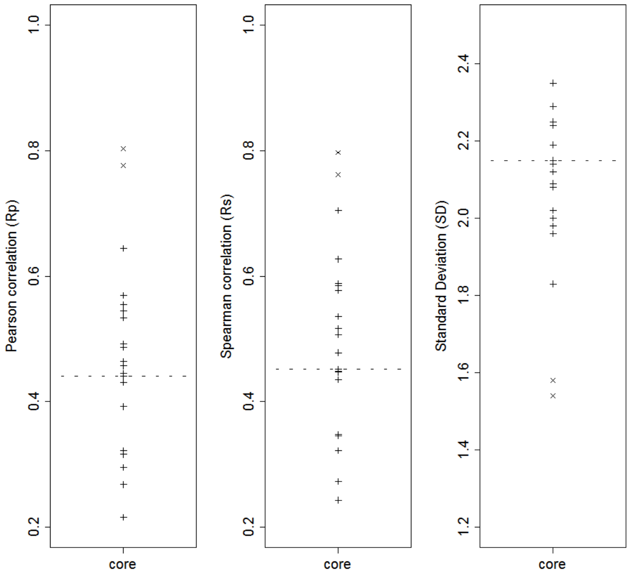


Figure 9: Performance of 18 scoring functions on the PDBbind benchmark as measured by Pearson's correlation coefficient (R_p), Spearman's correlation coefficient (R_s) and standard deviation of the difference between predicted and measured binding affinity (SD). The three plots on the right visually shows the relative predictive power of RF-Score ('x' signs) against that of the other 17 scoring functions ('+' signs). NHA is the performance of a linear regression model with the number of heavy atoms of the ligand as only variable (this baseline is shown as a horizontal discontinuous line in the plots).

When introducing a scoring function, only the scoring function built with the random seed that provides the best performance is generally reported. We have followed here a more precise way to assess performance differences between scoring functions by comparing median performances from a set of independent trials. Moreover, in order to address the question of how significant is the reported improvement over the original version of RF-Score, we have trained and tested the original RF-Score using 10 different random seeds. Thereafter, we have repeated the process, using the same random seeds, with the new version of RF-Score. The resulting boxplots are compared in Figure 10. Lastly, we carried out a two-sample t-test for each performance measure to find out that all differences are statistically significant (Rp p-value= 6.0x10^{-12}, Rs p-value=1.5x10^{-11} and SD p-value=3.8x10^{-4}).

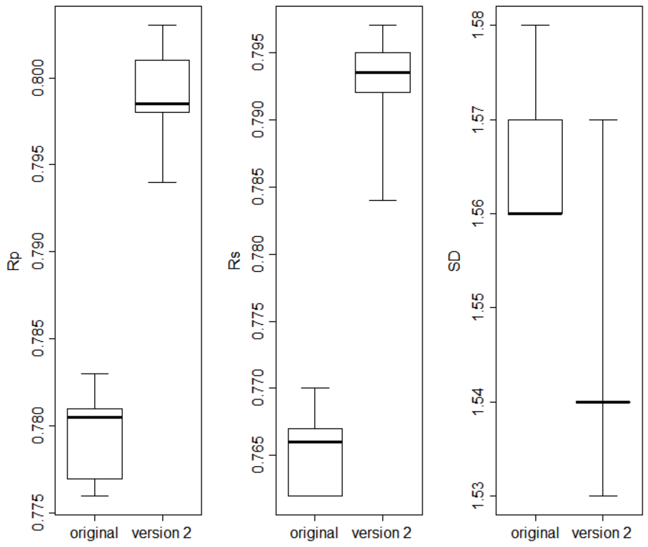


Figure 10: Performance comparison between the original version and the new version of RF-Score.

DISCUSSION

The new version of RF-Score performs much better than classical scoring functions on the same test set. In fact, this performance gain must be actually larger in most cases since only RF-Score and X-Score, among all scoring functions represented in Figure 9, use training sets that do not overlap with the test set. Having training complexes in the test set artificially enhances the performance of a scoring function, as it is not exclusively predicting unseen complexes but merely reporting the lower training error of those overlapping complexes. Indeed, adding a third of the test set to the training set makes RF-Score::Elem-v2's R_p rise from 0.803 (no overlap between training and test sets) to 0.872 (65 overlapping complexes). Clearly, the same training and test set must be used when comparing scoring functions, but unfortunately this has not been required in recent community benchmarks[48].

Furthermore, it could be argued that there is something particular about the training/test partition selected by Cheng et al. in the PDBbind benchmark. This partition was chosen to compare RF-Score against the best scoring function in that study under exactly the same experimental conditions. An experiment to investigate this question was already carried out in the original RF-Score paper[23] (Appendix A4 in the Supplementary Data of that paper[23]) and further discussed in a subsequent commentary[29] (see Figure 1 therein). The performances of RF-Score for 25 randomly generated training/test partitions with the same sizes as the benchmark partition (1105/195) were calculated. The experiment demonstrated that there is a minor difference in RF-Score performance between the benchmark partition and the median of these 25 alternative partitions.

While our study focuses on generic scoring functions, we would like to briefly comment on how RF-Score would perform on subgroups of the test set (e.g. complexes whose proteins belong to the same family). Clearly, the better the performance of RF-Score over another scoring function on the full test set, the more test subsets will be better predicted by RF-Score. To illustrate this, Appendix A1 presents the performance of the new version of RF-Score (RF-Score::Elem-v2; full test set RMSE=1.54) and RF using Credo intermolecular interaction features (RF-Score::Credo; full test set RMSE=1.72) on the four subsets resulting from partitioning the test set by binding affinity ranges. Appendix A2 presents another experiment where two small subsets of 23 complexes are generated, one containing those with the most similar ligands and the other subset with the most dissimilar ligands in terms of chemical structure. RF-Score::Elem-v2 outperforms RF-Score::Credo in all subsets but the one with the most dissimilar ligands, where RF-Score::Credo performs slightly better. These results illustrate the fact that, despite RF-Score::Elem-v2 generally performing better, there could be a few complexes (e.g. ligand-bound structures of a particular target) where other scoring functions perform better. We intend to study this issue further in the future.

On the other hand, it is noteworthy that RF-Score performed best when describing a complex using a 12Å distance cutoff between atom pairs, a distance well beyond direct interatomic contacts. The improvement over RF-Score with a more common 6Å cutoff is however modest (+0.017 in median Rp, +0.003 in median Rs and -0.05 log K units in median SD; see figure 3). This result suggests that there is a minor favourable contribution from atom pairs separated by a distance between 6 and 12Å over the 1300 considered complexes. Such long-range contributions to protein-ligand binding affinity have been attributed to the electronic properties of the binding site and ligand being affected by all protein atoms[49] and also to long-range electrostatics interactions[50]. Increasing the non-covalent cutoff to 12Å has also been found beneficial in protein folding[51] and DNA molecular dynamics[52] simulations. It seems therefore that RF-Score is able to capture long-range effects implicitly to some extent.

The main reason why RF-Score works much better than classical scoring functions at predicting binding affinity of most complexes is due to the circumvention of the strong assumption of a predetermined functional form. All classical scoring functions consists of a sum of typically nonlinear terms with respect to selected inter-atomic distances, such as van der Waals terms in empirical scoring functions or particular atom-atom potentials in knowledge-based scoring functions. For instance, the Scoring Function Consortium (SFC), in a concerted effort between 10 pharmaceutical companies and academic institutions, generated an empirical scoring function (SFCscore)[53], which by the time of its development was clearly superior to most of the then available scoring functions. Very recently, one of the leading SFC authors has demonstrated[54] that, by using Random Forest regression instead of SFCscore's additive functional form and keeping all other modeling choices unaltered (training data, test data and descriptors), performance rises from RMSE=1.84 to RMSE=1.56 (0.683 to 0.788 in the case of Rs). This is a very large improvement for a single modification in a generic model, especially taking into account that scoring functions are highly optimised due to intense work over the years in this area. Another study demonstrating that assuming an additive functional form is detrimental for the performance of empirical scoring functions is by Kinnings et al.[30] As force-field and knowledge-based scoring functions make the same assumption, these studies strongly suggest that a machine-learning version of other classical scoring functions will also result in significant improvement.

Another important conclusion of our study is that a more precise chemical description of the protein-ligand complex does not generally lead to more accurate prediction of binding affinity (see figure 4). In the first study, Li et al.[55] present a scoring function tested on exactly the same test set as us, with a much larger training set that includes ours and the use of a very precise description consisting of 50 calculated descriptors falling into nine interaction categories: van der Waals, hydrogen-bonding, electrostatic, pi-system, metal-ligand bonding, desolvation effect, entropic loss effect, shape matching, and surface property matching (Table 1 at page 593 of Li et al.'s paper). Li et al. obtained SD=1.63 and Rs=0.779 (Table 4 at page 597 of Li et al.'s paper), whereas RF-Score originally obtained SD=1.58 and Rs=0.762 on the same test set. Interestingly, these authors referenced RF-Score, but did not include it in the comparison or comment on why its performance was better in some performance measures despite using much simpler descriptors and less data for training.

The second independent study provides an even more direct comparison. Zilian and Sotriffer[51] used the same training set, test set and regression model as RF-Score. The only difference between their scoring function and ours is in the 63 used descriptors, which were one of the outcomes of the industry-academia Scoring Function Consortium. These descriptors include the number of rotatable bonds in the ligand, hydrogen bonds, aromatic interactions, polar and hydrophobic contact surfaces, among others (a full list can be found in Table 1 of page 398 of the original SFC paper[50]). Their best scoring function achieved RMSE=1.56 and Rs=0.788 (Table 1 of Zilian and Sotriffer's paper), which is slightly better than the original version of RF-Score (RMSE=1.58 and Rs=0.762). If the modelling assumptions implied in the calculation of chemical properties were generally accurate, we should have seen many scoring functions performing much better than RF-Score thanks to using a more precise chemical description. But we have actually seen the opposite in these two independent studies, once we compare the performances achieved by Li et al. (SD=1.63 and Rs=0.779) and Zilian and Sotriffer (RMSE=1.56 and Rs=0.788) to that of the new version of RF-Score (SD=1.54, RMSE=1.54 and Rs=0.797) on the same diverse test set. The new version differs from the original version of RF-Score in that features are distance-dependent, but still do not explicitly incorporate calculated protonation states.

We discuss next four convoluted factors that may contribute to this result: modeling assumptions; co-dependence of representation and regression; data restricted to the bound state; and conformational heterogeneity in data. The first factor is that more precise descriptors often mean making modelling assumptions that introduce additional error. For example, the protonation state of an atom needs to be estimated in order to assign its Sybyl type, but the local change in pH induced by hydrogen bond donors/acceptors in nearby residues and water molecules is usually not incorporated into scoring functions for binding affinity prediction. Similar arguments can be constructed regarding the calculation of donor-hydrogen-acceptor angles to perceive hydrogen bonds. The question remains as to how large the impact of this error is compared to that of not considering protonation states at all (the element descriptor scheme).

The second factor, often neglected, is the optimality of problem representation (description scheme) for the applied solution construction method (regression technique). From a purely chemical perspective, deconvoluting elemental atom types into their various hybridisation states constitutes a more precise description of the complex. However, this scheme also results in a higher number of features and thus more sparse features. The latter are detrimental for random forest regression because as many data as possible are needed to characterise the interaction between each pair of atom types, best achieved by minimising the number of different types defined. In practice, the definition of atom types must reflect a compromise between these two conflicting objectives, so as to ensure that the features are backed up by sufficient data to be statistically as well as chemically meaningful. This situation gives rise to a trade-off between the predictability and interpretability of the model, which is not uncommon in regression problems[56] and has also been observed here (see Figure 4).

For the sake of efficiency, scoring functions only exploit the information contained in the bound state of the complex, as represented by a crystal structure. However, binding affinity also depends on the energetic contributions from ligand and protein desolvation as well as induced fit upon binding. The third contributing factor is therefore the uncertainty about how well a particular description of the bound complex is also describing the complex just before desolvation and induced fit takes place. We speculate that descriptors whose values change less during the binding process might be more suitable for predicting binding affinity using only data about the bound state. For instance, element descriptors do not change much in general during the binding process, as a fixed cutoff will include roughly the same protein and ligand atoms just before and after binding. In contrast, protonation states will generally change significantly upon binding because of desolvation.

The last contributing factor comes from the uncertainty arising from the fact that the crystallographers deposit a single structural model in the PDB while several different models may fit the electron density equally well[57]. The conformational heterogeneity of a complex (i.e. several bound states are possible for this complex, at least within experimental uncertainty) means that different sets of descriptors would be generated for exactly the same binding affinity. Access to the multiple structural models of a complex that are significantly different at binding site level is likely to be helpful in deciding how to best address this issue. In particular, it would be interesting to investigate whether combining the predictions from each structure is a better strategy than simply predicting from the deposited structure.

Our finding that binding affinity can be better predicted when calculated protonation states are not explicitly incorporated into the scoring function will be certainly seen as a controversial result by most molecular modellers. We are providing next an intuitive explanation for this result. In machine learning nomenclature, the chemical description of complexes constitutes a data representation. Representations present an opportunity to incorporate domain knowledge into the problem, which in principle can help to disentangle the different explanatory factors for variation of the predicted variable (binding affinity here) and thus lead to better performance by simplifying the regression problem. However, as domain knowledge is affected by confounding factors and implemented with various degrees of efficacy, it is entirely possible to obtain better performance by incorporating less domain knowledge (i.e. introducing less noise) and hence relying more on pure inference from the data. In our problem, binding affinity is experimentally determined in solution along a trajectory in the co-dependent conformational spaces of the interacting molecules, whereas the structure represents a possible final state of that process in a crystallised environment. Consequently, very precise descriptors calculated from the structure are not necessarily more representative of the dynamics of binding than less precise descriptors. This means that a more precise description will not necessarily lead to a better prediction of binding affinity, as it has been proven here using Random Forest. Because the information content of a set of variables is independent of the adopted regression model, the use of an alternative regression technique should lead to the same conclusion, although this point is still to be confirmed experimentally. We cannot stress enough that we are not making any claim about the importance of protonation for pose generation in docking. Pose generation and re-scoring are different problems and so are the objectives that the corresponding scoring functions must fulfil.

In summary, we have seen that one can be easily fooled by uncertainty when investigating more accurate scoring functions. Given the unavoidable uncertainty, we believe that rigorous and systematic numerical studies are the most reliable way to make progress in predicting intermolecular binding affinity. We hope that the availability of the RF-Score software (links are provided in the Methods section) will encourage experts in the area to try to perform better on the PDBbind benchmark using alternative chemical descriptions as a way to investigate this issue further. The code permits reproducing the results obtained by RF-Score::Elem-v2 and can also be used as a template to test alternative regression techniques implemented in R. Without any modification, the RF-Score software can be employed to re-score ligands in crystal structures or docking poses. There is a range of applications in which more accurate prediction of binding affinity of a complex would be very useful, some of them new such as replacing force-fields in molecular dynamics simulations. Other applications include structure-based virtual screening and lead optimisation. In fact, applying a simpler variant of RF-Score::Elem-v1[23] to prospective virtual screening has already been found[37] to excel at discovering innovative inhibitors of antibacterial targets. Very recently[58], RF-Score::Elem-v1 has been incorporated into an easy-to-set large-scale docking webserver (http://istar.cse.cuhk.edu.hk/idock) to carry out virtual screening of up to 17 million purchasable molecules from the ZINC database[59], which should be upgraded soon to RF-Score::Elem-v2.

ACKNOWLEDGEMENTS

We would like to thank Dr Nidhi Tyagi (EMBL-EBI) for feedback on the manuscript.

The following sources of funding are gratefully acknowledged: the Medical Research Council for a Methodology Research Fellowship (grant no. G0902106 awarded to P.J.B.). We thank the Wellcome Trust for an SDDI grant supporting of A.S.

SUPPORTING INFORMATION

Appendix A1 (performance on subgroups of test set by binding affinity of complexes), Appendix A2 (performance on subgroups of test set by chemical similarity of ligands), Appendix A3 (classification scheme for all interaction types used in CREDO). This information is available free of charge via the Internet at http://pubs.acs.org

AUTHOR CONTRIBUTIONS

P.J.B. designed the study, implemented the new version of RF-Score and run the numerical experiments. A.S. implemented the software to calculate descriptors. P.J.B. wrote the manuscript using the information provided by A.S. on how the descriptors were implemented. All authors discussed results and commented on the manuscript.

[1] Song L, Kalyanaraman C, Fedorov AA, Fedorov EV, Glasner ME, et al. (2007) Prediction and assignment of function for a divergent N-succinyl amino acid racemase. *Nat Chem Biol* 3: 486–491.

[2] Hermann JC, Marti-Arbona R, Fedorov AA, Fedorov E, Almo SC, et al. (2007) Structure-based activity prediction for an enzyme of unknown function. *Nature* 448: 775–779.

[3] Jorgensen WL (2009) Efficient Drug Lead Discovery and Optimization. *Acc Chem Res* 42: 724–733.

[4] Schneider G, Böhm, H-J (2002) Virtual screening and fast automated docking methods. *Drug Discov. Today* 7: 64–70.

[5] Leach AR, Shoichet BK, Peishoff CE. (2006) Prediction of Protein–Ligand Interactions. Docking and Scoring: Successes and Gaps. *J Med Chem* 49: 5851–5855.

[6] Moitessier N, Englebienne P, Lee D, Lawandi J, Corbeil CR. (2008) Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br J Pharmacol* 153: S7–S26.

[7] Novikov FN, Zeifman AA, Stroganov OV, Stroylov VS, Kulkov V, et al. (2011) CSAR Scoring Challenge Reveals the Need for New Concepts in Estimating Protein–Ligand Binding Affinity. *J Chem Inf Model* 51: 2090–2096.

[8] Huang N, Kalyanaraman C, Bernacki K, Jacobson MP. (2006) Molecular mechanics methods for predicting protein-ligand binding. *Phys Chem Chem Phys* 8: 5166–5177.

[9] Ewing TJA, Makino S, Skillman AG, Kuntz ID. (2001) DOCK 4.0: Search Strategies for Automated Molecular Docking of Flexible Molecule Databases. *J Comput-Aided Mol Des* 15: 411-428.

[10] Mitchell JBO, Laskowski RA, Alex A, Thornton JM (1999) BLEEP - potential of mean force describing protein-ligand interactions: I. Generating potential. *J Comput Chem* 20: 1165–1176.

[11] Gehlhaar DK, Verkhivker GM, Rejto PA, Sherman CJ, Fogel DR et al. (1995). Molecular Recognition of the Inhibitor AG-1343 by HIV-1 Protease: Conformationally Flexible Docking by Evolutionary Programming. *Chem Biol* 2: 317-324.

[12] Muegge I, Martin YC. (1999) A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach*.* *J Med Chem* 42: 791-804.

[13] Mooij WTM, Verdonk ML. (2005) General and Targeted Statistical Potentials for Protein-Ligand Interactions. *Proteins Struct Funct Bioinf* 61: 272-287.

[14] Gohlke H, Hendlich M, Klebe G. Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions. *J Mol Biol* 295: 337-356.

[15] Böhm H-J. (1994) The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant for a Protein-Ligand Complex of Known Three-Dimensional Structure. *J Comput-Aided Mol Des* 8: 243-256.

[16] Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee R P. (1997) Empirical Scoring Functions: I. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. *J Comput-Aided Mol Des* 11: 425-445.

[17] Wang R, Lai L, Wang S. (2002) Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. *J Comput-Aided Mol Des* 16: 11-26.

[18] Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, 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.

[19] Krammer A, Kirchhoff PD, Jiang X, Venkatachalam CM, Waldman M. (2005) LigScore: A Novel Scoring Function for Predicting Binding Affinities. *J Mol Graphics Modell* 23: 395-407.

[20] Michel J, Essex JW (2010) Prediction of protein–ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *J Comput Aided Mol Des* 24: 639–658.

[21] Mobley DL (2012) Let's get honest about sampling. *J Comput Aided Mol Des* 26: 93–95

[22] Guvench O, MacKerell Jr AD (2009) Computational evaluation of protein-small molecule binding. *Curr Opin Struct Biol* 19: 56–61.

[23] Ballester PJ, Mitchell JBO (2010) A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics* 26: 1169–1175.

[24] Baum B, Muley L, Smolinski M, Heine A, Hangauer D, Klebe G (2010) Non-additivity of functional group contributions in protein-ligand binding: a comprehensive study by crystallography and isothermal titration calorimetry. *J Mol Biol* 397: 1042–1054.

[25] Arunan E, Desiraju GR, Klein RA, Sadlej J, Scheiner S, et al. (2011) Definition of the hydrogen bond (IUPAC Recommendations 2011). *Pure and Applied Chemistry* 83: 1637–1641.

[26] Snyder PW, Mecinovic J, Moustakas DT, Thomas SW 3rd, Harder M, et al. (2011) Mechanism of the hydrophobic effect in the biomolecular recognition of arylsulfonamides by carbonic anhydrase. *Proceedings of the National Academy of Sciences* 108: 17889–17894.

[27] Li L, Li J, Khanna M, Jo I, Baird JP et al. (2010) Docking to Erlotinib Off-Targets Leads to Inhibitors of Lung Cancer Cell Proliferation with Suitable in Vitro Pharmacokinetics. *ACS Med Chem Lett* 1: 229–233.

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

[29] Ballester PJ, Mitchell JBO (2011) Comments on 'Leave-Cluster-Out Cross-Validation is appropriate for scoring functions derived from diverse protein data sets': Significance for the validation of scoring functions. *J Chem Inf Model* 51: 1739–1741.

[30] Kinnings SL, Liu N, Tonge PJ, Jackson RM, Xie L, Bourne PE (2011) A Machine Learning-Based Method to Improve Docking Scoring Functions and its Application to Drug Repurposing. *J Chem Inf Model* 51: 408–419.

[31] Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH (2012) Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review. *The AAPS Journal* 14: 133–141.

[32] Lahti JL, Tang GW, Capriotti E, Liu T, Altman RB. (2012) Bioinformatics and Variability in Drug Response: A Protein Structural Perspective. *J R Soc Interface* 9: 1409–1437

[33] Sotriffer C (2012) Scoring Functions for Protein–Ligand Interactions. In: Gohlke H, editor. Protein-Ligand Interactions. *Wiley-VCH Verlag GmbH & Co. KGaA*. pp. 237–263.

[34] Das S, Krein MP, Breneman CM (2010) Binding Affinity Prediction with Property-Encoded Shape Distribution Signatures. *J Chem Inf Model* 50: 298–308.

[35] Li L, Wang B, Meroueh SO (2011) Support Vector Regression Scoring of Receptor-Ligand Complexes for Rank-Ordering and Virtual Screening of Chemical Libraries. *J Chem Inf Model* 51: 2132–2138.

[36] Durrant JD, McCammon JA (2011) NNScore 2.0: A Neural-Network Receptor–Ligand Scoring Function. *J Chem Inf Model* 51: 2897–2903.

[37] Ballester PJ, Mangold M, Howard NI, Marchese-Robinson RL, Abell C, et al. (2012) Hierarchical virtual screening for the discovery of new molecular scaffolds in antibacterial hit identification. *J R Soc Interface* 9: 3196-3207.

[38] Deng Z, Chuaqui C, Singh J (2004) Structural Interaction Fingerprint (SIFt):□ A Novel Method for Analyzing Three-Dimensional Protein–Ligand Binding Interactions. *J Med Chem* 47: 337–344.

[39] Schreyer A, Blundell T (2009) CREDO: A Protein–Ligand Interaction Database for Drug Discovery. *Chemical Biology & Drug Design* 73: 157–167.

[40] Allen FH, Baalham CA, Lommerse JPM, Raithby PR. (1998) Carbonyl-Carbonyl Interactions can be Competitive with Hydrogen Bonds. *Acta Crystallographica Section B* 54:320–329.

[41] O'Boyle NM, Morley C, Hutchison GR. (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem Central J;* 2:5.

[42] Eric Jones E., Oliphant T., Peterson P. et al. (2001) SciPy: Open Source Scientific Tools for Python. http://www.scipy.org

[43] Wang R, Fang X, Lu Y, Yang C-Y, Wang S (2005) The PDBbind Database: Methodologies and updates. *J Med Chem* 48: 4111–4119.

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

[45] Breiman L (2001) Random Forests. *Mach Learn* 45: 5–32.

[46] Rognan D, Lauemoller SL, Holm A, Buus S, Tschinke V (1999) Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J Med Chem* 42: 4650–4658.

[47] Fan H, Schneidman-Duhovny D, Irwin JJ, Dong G, Shoichet BK, et al. (2011) Statistical Potential for Modeling and Ranking of Protein–Ligand Interactions. *J Chem Inf Model* 51: 3078–3092.

[48] Smith RD, Dunbar JB, Ung PM-U, Esposito EX, Yang C-Y, Wang S, et al. (2011) CSAR Benchmark Exercise of 2010: Combined Evaluation Across All Submitted Scoring Functions. *J Chem Inf Model* 51: 2115–2131.

[49] Hayik SA, Dunbrack R, Merz KM. (2010) A Mixed QM/MM Scoring Function to Predict Protein-Ligand Binding Affinity. *J Chem Theory Comput.* 6: 3079–91.

[50] Caravella JA, Carbeck JD, Duffy DC, Whitesides GM, Tidor B. (1999) Long-Range Electrostatic Contributions to Protein−Ligand Binding Estimated Using Protein Charge Ladders, Affinity Capillary Electrophoresis, and Continuum Electrostatic Theory. *J Am Chem Soc* 121: 4340–7.

[51] Piana S, Lindorff-Larsen K, Dirks RM, Salmon JK, Dror RO, Shaw DE. (2012) Evaluating the Effects of Cutoffs and Treatment of Long-range Electrostatics in Protein Folding Simulations. *PLoS ONE* 7: e39918.

[52] Norberg J, Nilsson L. (2000) On the Truncation of Long-Range Electrostatic Interactions in DNA. *Biophysical Journal* 79: 1537–53.

[53] Sotriffer CA, Sanschagrin P, Matter H, Klebe G. (2008) SFCscore: scoring functions for affinity prediction of protein-ligand complexes. *Proteins* 73: 395–419.

[54] Zilian D, Sotriffer CA. (2013) SFCscoreRF: A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein–Ligand Complexes. *J Chem Inf Model.* 53: 1923–1933.

[55] Li G-B, Yang L-L, Wang W-J, Li L-L, Yang S-Y. (2013) ID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein–Ligand Interactions. *J Chem Inf Model* 53: 592–600.

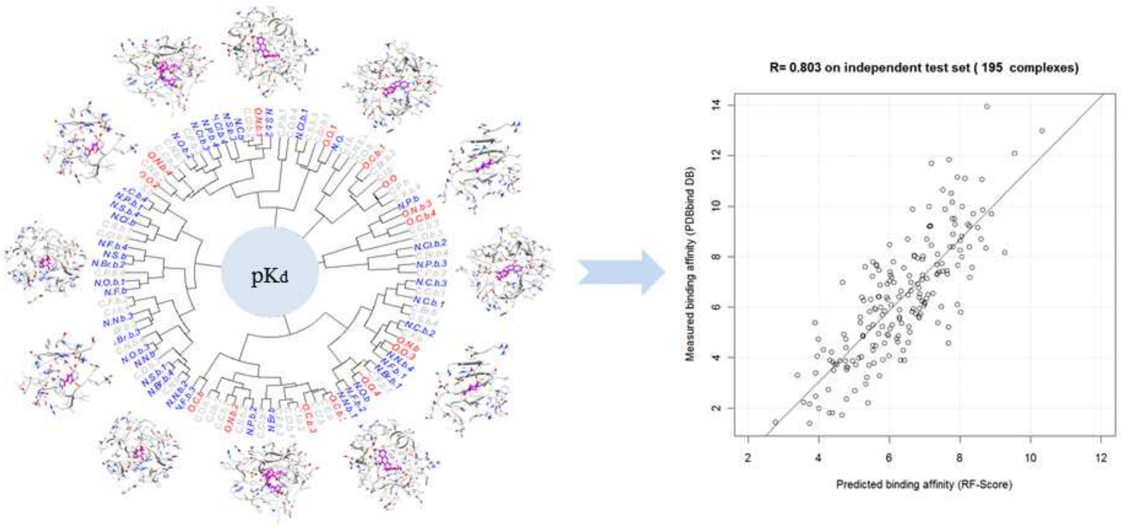
[56] Sukumar N, Das S. (2011) Current Trends in Virtual High Throughput Screening Using Ligand-Based and Structure-Based Methods. *Combin Chem High Throughput Screen* 14: 872–88.

[57] Furnham N, Blundell TL, DePristo MA, Terwilliger T (2006) Is one solution good enough? *Nat Struct Mol Biol* 13: 184-185

[58] Li H, Leung K-S, Ballester PJ, Wong M-H (2014) istar: A Web Platform for Large-Scale Online Protein-Ligand Docking. *PLoS ONE 9(1): e85678.*

[59] Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG (2012) ZINC: A Free Tool to Discover Chemistry for Biology. *J Chem Inf Model* 52:1757–68.

TOC graphic


pKd
R= 0.803 on independent test set (195 complexes)
Measured binding affinity (PDBbind DB)
Predicted binding affinity (RF-Score)