JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

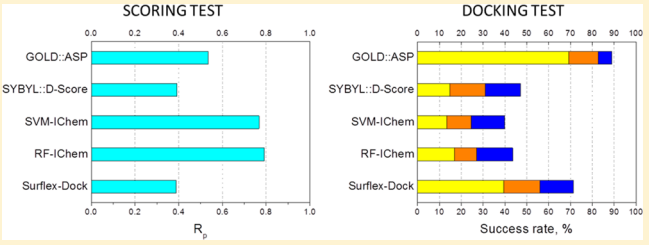
Beware of Machine Learning-Based Scoring Functions—On the Danger of Developing Black Boxes

Joffrey Gabel, Jérémy Desaphy, and Didier Rognan*

Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, 74 route du Rhin, F-67400 Illkirch, France

Supporting Information

ABSTRACT: Training machine learning algorithms with protein−ligand descriptors has recently gained considerable attention to predict binding constants from atomic coordinates. Starting from a series of recent reports stating the advantages of this approach over empirical scoring functions, we could indeed reproduce the claimed superiority of Random Forest and Support Vector Machine-based scoring functions to predict experimental binding constants from protein−ligand X-ray structures of the PDBBind dataset. Strikingly, these scoring functions, trained on simple protein−ligand element−element distance counts, were almost unable to enrich virtual screening hit lists in true actives upon docking experiments of 10 reference DUD-E datasets; this is a a feature that, however, has been verified for an a priori less-accurate empirical scoring function (Surflex-Dock). By systematically varying ligand poses from true X-ray coordinates, we show that the Surflex-Dock scoring function is logically sensitive to the quality of docking poses. Conversely, our machine-learning based scoring functions are totally insensitive to docking poses (up to 10 Å root-mean square deviations) and just describe atomic element counts. This report does not disqualify using machine learning algorithms to design scoring functions. Protein−ligand element−element distance counts should however be used with extreme caution and only applied in a meaningful way. To avoid developing novel but meaningless scoring functions, we propose that two additional benchmarking tests must be systematically done when developing novel scoring functions: (i) sensitivity to docking pose accuracy, and (ii) ability to enrich hit lists in true actives upon structure-based (docking, receptor−ligand pharmacophore) virtual screening of reference datasets.



INTRODUCTION

Predicting the absolute binding free energy from atomic three-dimensional (3D) coordinates of protein−ligand complexes is one of the remaining grand challenges for computational chemists. When applied to drug discovery, it should enable to significantly enhance, for the right reasons, hit rates in structure-based virtual screening, and guide medicinal chemists in the hit to lead optimization of experimentally confirmed hits.

From the pioneering work of Böhm[1] in the early 1990s on designing a fast scoring function from first-principles, many approaches (regression-based empirical scoring function, potential of mean force, molecular mechanics), descriptors and protein−ligand datasets have been utilized to predict inhibition constants from protein−ligand atomic coordinates.[2−4] If current fast scoring functions have proven their capacity to discriminate true actives from nonbinding ligands in hundreds of structure-based virtual screening reports;[5] they are still unable, with few exceptions, to predict binding constants at a precision (ca. 1 pK_i unit) required for structure-based hit to lead optimization. Numerous public[6,7] and private initiatives[8] have been addressed to the community of computational chemists, to enhance the accuracy of fast scoring functions. Among the main directions that have been followed are (i) the design of larger,[8] higher-quality,[9] and more-diverse[10] training/test sets of protein−ligand complexes; (ii) the use of novel protein−ligand interaction descriptors;[8] and (iii) the application of nonlinear regression methods[11,12] to link descriptors to experimental data.

Despite time and expertise spent on this issue, very modest improvements had been achieved until a recent series of independent studies reporting the usage of machine learning algorithms to predict absolute binding free energies.[13−15] A striking example was illustrated by two reports of Ballester et al.[14,15] training a Random Forest (RF) model on very simple descriptors (protein−ligand element−element distance counts), and predicting binding constants with a standard deviation of 1.5 pK_i units, whereas all previous classical empirical scoring functions had been leveling off at a plateau close to 2 pK_i units.[4,14] Very intriguingly, the accuracy of the corresponding RF score was shown to be inversely correlated to the physical relevance of the descriptors on which the model was trained on (element-dependent distance counts > atom type-dependent distance counts > true interaction descriptors).[14]

From our viewpoint, it is difficult to understand why ultra-simple descriptors (element−element distance counts) would outperform interaction-driven attributes in predicting binding free energies. Therefore, in this manuscript, we give a close inspection to the RF score and develop two machine learning (ML)-based scoring functions following prior Ballester's work.[14,15]

Received: July 8, 2014

   dx.doi.org/10.1021/ci500406k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

The ML scoring functions were challenged in three independent tests: scoring power, docking power, virtual screening power. If they successfully pass the screening power test, they largely failed in the remaining tests. Looking in depth for reasons explaining this discrepancy, we demonstrate that ML-based functions are overtrained on descriptors (protein−ligand element−element distance counts) that do not depict protein−ligand interactions but interaction-independent counts.

COMPUTATIONAL METHODS

Protein−Ligand Datasets. *PDBBind.* The PDBBind (v.2007) dataset[16] is a standard benchmark tool for comparing scoring functions aimed at predicting binding free energies (pK_i values) from protein−ligand X-ray structures of known experimental binding constants. It consists in two sets of carefully selected X-ray structures from the Protein DataBank (PDB).[17] A training set of 1105 complexes is used to calibrate the scoring function. A second test set of 195 complexes (core set) is then utilized to monitor the predictive ability of the derived function. The scoring function is classically evaluated in two tests, inspecting different properties. The "scoring power test" evaluates the ability of the scoring function to predict experimental binding constants for complexes of the core set. Success is inferred from the correlation coefficient between predicted and experimental binding constants: the higher the correlation coefficient, the better. The most accurate scoring function (RF score: Elem-v2) to date presents, for this core set, a Pearson's correlation coefficient of $R_p = 0.803$.[14]

The "docking power test", then looks at the sensitivity of the scoring function to docking pose accuracy. For that purpose, a collection of docking poses (decoys) was generated from four different docking tools for each complex of the core set.[16] Success of the scoring function is evaluated by measuring the percentage of entries for which the top-ranked docking pose is close (root mean square deviation < 2 Å) to the X-ray solution. The most accurate scoring function (GOLD::ASP) to date presents, for this core set, a success rate of 90%.[16]

The PDBBind (v.2007) dataset (ligand input file in sd format, protein input file in pdb format, docking decoys in mol2 file format) was downloaded from the PDBBind-CN Web site.[18] For the 1105 entries of the training set, PDB input files for the protein were converted to the mol2[19] file format, using an in-house SYBYL[19] programming language (SPL) script. Water, co-factor, and ion molecules not present in the binding site were explicitly deleted. In addition, the coordinates of polar hydrogens, as well as tautomeric and protonation states (binding site amino acids, ligand), were manually modified in order to optimize the intramolecular and intermolecular hydrogen bonding networks. For the 195 entries of the core set, the same procedure as those described above was utilized. Docking decoys (mol2 file format) of the docking power test were used without any further modification. All curated structures (protein, ligands) are available as Supporting Information (mol2 file format).

DUD-E Target and Ligand Sets. The DUD-E dataset[20] is a standard benchmark set for evaluating the virtual screening accuracy of docking programs. It consists of 22 886 active compounds, and their affinities, against 102 targets, with an average of 224 ligands per target. For each target, the DUD-E database provides a set of decoys (50 decoys for each active) having similar physicochemical properties but dissimilar two-dimensional (2-D) topology. The DUD-E dataset is here used to infer the accuracy of the scoring function on a "virtual screening power test". For diverse DUD-E targets, the corresponding actives and decoys are docked and ranked according to the scoring function under evaluation. The accuracy of the function is measured by computing by the area under the receiver operating characteristic (ROC) curve,[21] for binary classification models (active/inactive) based on the docking score values. Areas under the ROC curves above 0.7−0.75 are generally considered as acceptable for predictive models.

For the present project, DUD-E active and decoy ligand sets for 10 targets covering 5 different protein families: G protein-coupled receptors (adenosine A2A receptor (AA2AR), adrenergic beta2 receptor (ADRB2)), nuclear hormone receptors (androgen receptor (AND), glucocorticoid receptor (GCR)); other enzymes (adenosine deaminase (ADA), prostaglandin G/H synthase 2 (PGH2)); proteases (angiotensin-converting enzyme (ACE), renin (RENI)); protein kinases (fibroblast growth factor receptor 1 (FGFR1), RAC-alpha serine/threonine-protein kinase (AKT1)) were downloaded in 3D mol2 file format from the DUD-E[20] Web site (http://http://dude.docking.org/). For each target, a unique representative X-ray structure was selected for docking and processed as described above (PDBBind dataset).

Protein−Ligand Descriptors. Protein−ligand element−element distance counts were computed following Ballester's report,[14] using the Ichem toolkit.[22] For the ligand, nine elements were considered (C, N, O, F, P, S, Cl, Br, I). For the protein, we chose six elements (C, N, O, P, S, M) in which M stands for any of the five following metal ions (Zn^{2+}, Fe^{2+}, Mg^{2+}, Mn^{2+}, Ca^{2+}). All pairwise element−element counts were computed and stored in seven bins, depending on the observed distance (0−12 Å, 0−2 Å, 2−4 Å, 4−6 Å, 6−8 Å, 8−10 Å, 10−12 Å). Several modifications to the original descriptors of Ballester were introduced:

(i) halogens were not considered as possible protein atoms,
(ii) bound water molecules were kept as part of the protein as long as the water oxygen atom was less than 6 Å away from any ligand heavy atom and that the water could geometrically form two hydrogen bonds with protein atoms,
(iii) ions and co-factors present in the binding site (defined as any amino acid within a 6.5-Å-radius sphere centered on the bound-ligand center of mass) were retained as part of the protein structure, and
(iv) all protein coordinates (and not only the ligand-binding pocket) were used as input to build element−element distance counts.

Hence, the total number of counts in the 10−12 Å interval, as reported by Ballester,[14] is underestimated, since pocket atoms are derived in this study from a 12 Å-radius sphere centered on the ligand center of mass and therefore miss a few counts between peripheral ligand atoms and residues at the sphere border. A vector of 378 descriptors (counts) was outputted for every complex and directly read as input file for machine learning.

Machine Learning Regression Models. *Random-Forest Regression (RF-IChem).* A total of 500 decision trees (ntree parameter) were trained on all 378 protein−ligand descriptors of the PDBBind training set (n = 1105), but varying the number of variables (m_{try} parameter) at each splitting node. For each m_{try} value (10, 20, 30, 40, 50, 60, 70, 80), 10 models were built from the training set, using different random seeds,

dx.doi.org/10.1021/ci500406k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Statistical Evaluation of Scoring Functions in Predicting the Experimental Binding Constant of 195 Complexes from the PDBBind Core Set

function	R_p[a]	R_s[b]	SD[c]	RMSE[d]
RF-score::Elem-v2[e]	0.803	0.797	1.54	1.53
RF-IChem[f]	0.791	0.787	1.55	1.55
SVM-IChem[g]	0.767	0.754	1.54	1.54
Surflex-Dock[h]	0.388	0.398	3.02	3.01
X-score::HMScore[i]	0.644	0.705	1.83	n.a.[j]
SYBYL::F-score[k]	0.216	0.243	2.35	n.a.

[a]Pearson's correlation coefficient. [b]Spearman's correlation coefficient. [c]Standard devation, given in pK_i units. [d]The root-mean-square error, given in pK_i units. [e]Data taken from Ballester et al.[14] [f]Best RF model out of 80 runs. [g]Best SVM model out of 6000 5-fold cross-validated runs. [h]pK_d score, as predicted by the Surflex-Dock native scoring function. [i]Best out of 16 scoring functions in the PDBBind (V.2007) scoring power test.[16] [j]Not available in the work reported by Cheng et al.[16] [k]Worst out of 16 scoring functions in the PDBBind (v.2007) scoring power test.[16]

and further applied to the core set (n = 195). All RF models were built using the RandomForest 4.6−7 library[23] in the R statistical package.[24] The m_{try} value leading to the best predictive model (highest Pearson correlation coefficient in predicting pK_i values of the core set) was further retained.

Support Vector Machine Regression (SVM-IChem). Out of the initial 378 variables, only 126 with non-null values for all training samples were selected. Optimal values for gamma (γ-) and c-parameters were found after systematic variation of both parameters in a 5-fold cross-validation procedure, using the rbf kernel, applied to the entire training set. We first iterate γ from 0.01 to 0.1, using a increment of 0.01, and we iterate c from 1 to 60 with an increment equal to 1. The best average F-measure of the 5-fold runs gave optimal γ- and c-values. The model leading to the best F-measure (γ = 0.05 and c = 26) was finally selected for external predictions. All SVM models were built using the e1071 libsvm and rpart libraries in the R statistical package.[24]

Surflex-Dock Docking. All ligands from the DUD-E dataset were docked into their original X-ray structure with Surflex-Dock.[25] Protomols were first generated from the list of cavity-lining residues (see above definition). Compounds were then docked with default settings (excepted for the "-pgeom" option) of the docking engine, keeping the best 10 poses, according to the native Surflex-Dock scoring function. For each target, the quality of the docking-based *in silico* screen was assessed by computing the area under the ROC curve,[21] for binary classification models (active/inactive) based on score values (Surflex-Dock, RF-IChem score, SVM-IChem score). ROC and Boltzmann-enhanced discrimination of ROC (BEDROC)[26] curves were computed using standard settings of the CROC program.[27]

Surflex-Dock Scoring. Simple scoring the 195 protein−ligand complexes of the PDBBind core set was done from protein and ligand MOL2 files (see the PDBBind section), using the "opt" option of the Surflex-Dock (v.2.601) software.[25] The outputed "initial score" was taken as the Surflex-Dock score.

RESULTS AND DISCUSSION

RF and SVM Models Trained on Simple Descriptors Outperform a Prototypical Scoring Function in Predicting Binding Constants from Atomic Coordinates (Scoring Power Test). Despite small modifications in our implementation of protein−ligand element−element distance counts (see Computational Methods), we could reproduce results reported by Ballester et al.[14] when training a RF model on element−element distance counts from the PDBBind training set and predicting the experimental binding constant of 195 core-set ligands to their cognate PDBBind target (see Table 1, Figure 1).

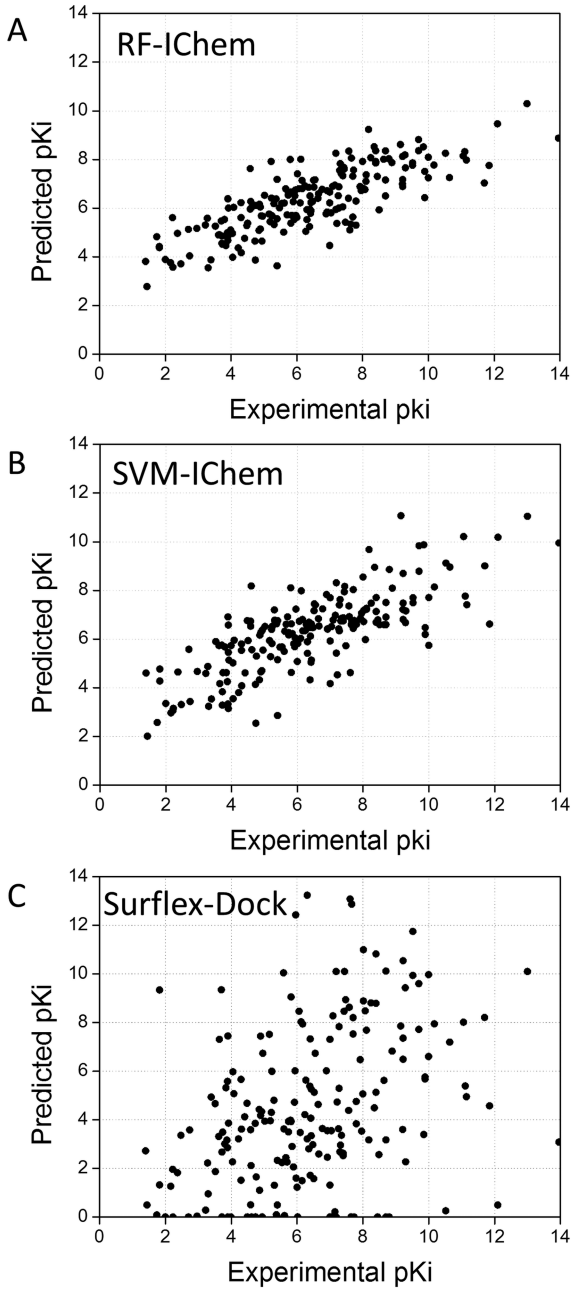


Figure 1. Prediction of the binding constant of 195 test protein−ligand complexes with two machine learning methods ((A) RF-IChem and (B) SVM-IChem) and a prototypical scoring function ((C) Surflex-Dock).

Observed correlation coefficients (Pearson and Spearman) were stable (R_p = 0.791 ± 0.02, R_s = 0.787 ± 0.02, m_{try} = 70, N = 10), quite similar to those previously described[14] and superior to that observed for 16 scoring functions in a prior benchmark study (Table 1).[16] Interestingly, the best SVM model was slightly less accurate (see Table 1, Figure 1) than the optimal RF model,

dx.doi.org/10.1021/ci500406k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 2. Area under the ROC Curve of a Binary Classification (Active, Inactive) of Docked Poses to the X-ray Structure of 10 Representative Targets

		DUD-E[a]		RF-IChem		SVM-IChem		Surflex-Dock	
target	PDB code[b]	actives	decoys	ROC[c]	BEDROC[d]	ROC	BEDROC	ROC	BEDROC
G protein-coupled receptors									
adenosine A2A receptor (AA2AR)	3pwh	482	31500	0.497	0.042	0.494	0.026	0.736	0.214
Beta2 adrenergic receptor (ADRB2)	3ny8	231	15000	0.631	0.121	0.639	0.095	0.854	0.457
Nuclear hormone receptors									
androgen receptor (ANDR)	2am9	269	14350	0.610	0.090	0.646	0.156	0.470	0.060
glucocorticoid receptor (GCR)	1p93	258	15000	0.676	0.136	0.581	0.141	0.557	0.183
Proteases									
angiotensin-converting enzyme (ACE)	3zqz	282	16900	0.658	0.183	0.753	0.358	0.840	0.383
renin (RENI)	3sfc	104	6958	0.754	0.138	0.826	0.300	0.878	0.478
Protein kinases									
RAC-alpha protein kinase (AKT1)	4ekl	293	16450	0.650	0.138	0.623	0.085	0.759	0.310
fibroblast growth factor receptor 1 (FGFR1)	3tt0	139	8700	0.480	0.017	0.428	0.059	0.721	0.213
Other enzymes									
adenosine deaminase (ADA)	1a4l	93	5450	0.630	0.129	0.537	0.131	0.759	0.274
prostaglandin G/H synthase 2 (PGH2)	3nt1	435	23150	0.555	0.070	0.469	0.021	0.721	0.113

[a]Database of useful decoys: Enhanced.[20] [b]PDB identifier chosen for the target protein structure. [c]Area under the ROC curve for a binary classification of ligands (actives, decoys) from their top-scored docked poses. [d]Area under the Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) curve.

although still in the same accuracy range for predicting pK_i values for the 195 core-set complexes. As to be expected from previous numerous reports,[2−4,14,16] a prototypical empirical scoring function such as that embedded in the Surflex-Dock docking program, although focusing on real protein−ligand interactions, is far less accurate (R_p = 0.388) with a root-mean-square error in prediction close to 3 pK_i units (see Table 1, Figure 1). Our results using the Surflex-Dock scoring function are consistent with a previous report predicting binding constants for the entire PDBBbind set (1300 complexes) using a slightly older release of the docking software (Surflex-Dock, v. 2.2).[4]

It is puzzling to consider that descriptors that do not explicitly account for intermolecular interactions behave so accurately to train a machine learning algorithm. Four possible reasons were tentatively provided by Ballester et al.:[14]

(i) lack of modeling assumptions usually required by more-sophisticated descriptions of protein−ligand interactions (e.g., protonation state of ligand and protein atoms),
(ii) co-dependence of representation and regression (a high number of sparse features is undesirable here),
(iii) lower dependency to specific bound states that do not mirror desolvation and entropy changes upon binding, and
(iv) conformational heterogeneity of the complexes not depicted in static X-ray structures.

The possibility that the ML models might have been overtrained, notably considering the ratio of descriptors (123 in Ballester et al.,[14] 378 in the present study) to objects (1105), cannot be ruled out. In our opinion, two very important tests were not applied to the original Ballester's scoring function: (i) the ability to discriminate true actives from decoys in docking experiments (virtual screening power test), and (ii) sensitivity of the scoring function to the ligand pose quality (docking power test). If ML-based scoring functions demonstrate such a superiority over empirical scoring functions in predicting binding affinities for a standard test set, this trend should also be observed in real docking experiments aimed at discriminating true actives from chemically similar decoys. Moreover, the true actives should be picked for the right reason. In other words, increasing root-mean square deviation (rmsd) values of the bound ligand structure from the true X-ray pose should be penalized by the scoring function and results in predicted lower pK_i values. Therefore, we investigated the properties of both ML-based functions (RF-IChem, SVM-IChem) and the prototypical Surflex-Dock function along these lines in the next sections.

RF- and SVM-Based Models Do Not Discriminate DUD-E Actives from Decoys in Docking Experiments (Virtual Screening Power Test). We previously reported the usage of Surflex-Dock in docking large DUD-E active/decoy sets to 10 pharmaceutical targets of importance, covering 5 main target families (G protein-coupled receptors, nuclear hormone receptors, proteases, protein kinases, diverse enzymes (see Table 2)).[22] This dataset was thus an excellent choice to evaluate the accuracy of RF-IChem and SVM-IChem scoring functions in real virtual screening conditions using a single set of already existing docking poses. Conversely to data from the screening power test, the two ML-based scoring functions were far less accurate than the native Surflex-Dock function in discriminating actives from decoys (Table 2, Figure 2). Considering the quality of the ROC-based classification of each score-ranked hit list by the area under the curve (Figure 2), the *a priori* less-accurate Surflex-Dock scoring function was determined to be good in 3 out of 10 cases (area of >0.80) and fair (0.7 < area < 0.8) for the 5 remaining cases. Considering early enrichments in true actives with Boltzmann-Enhanced ROC plots, Surflex-Dock succeeded in 8 out of 10 cases, which is a performance consistent with previous independent studies on diverse target sets.[28,29] Disappointingly, the RF-IChem and SVM-IChem functions were only successful in 2 out of 10 cases (AUROC > 0.7), and almost always were inferior to the Surflex-Dock function (Table2, Figure 2). For only the two protease sets (angiotensin-converting enzyme, and renin) would we recommend the SVM-IChem scoring function for a real virtual screening. Even worse, the RF-IChem score, which is the most suitable to predict binding constants from PDB

dx.doi.org/10.1021/ci500406k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

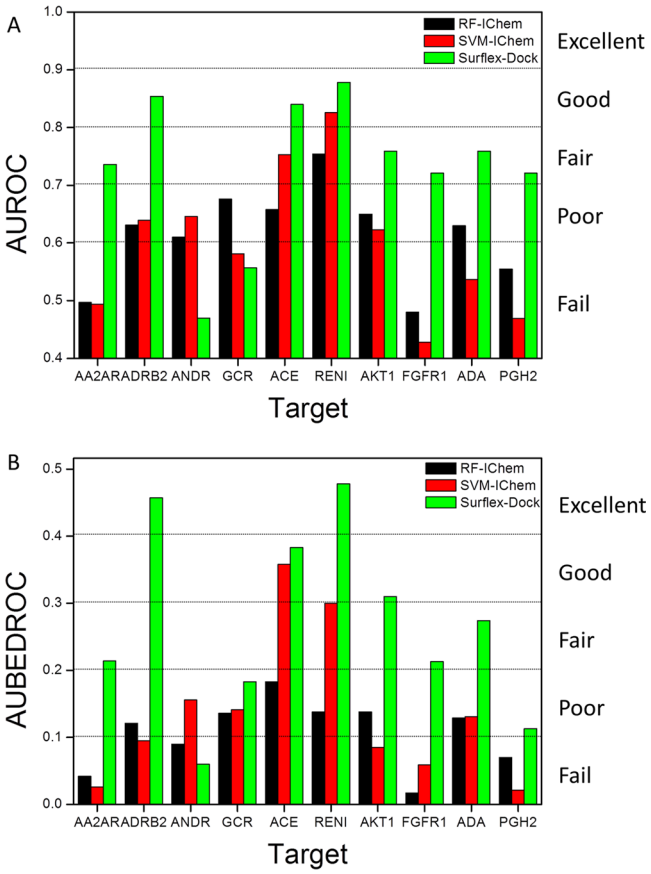


Figure 2. (A) Area under the ROC curve (AUROC) and (B) area under the Boltzmann-Enhanced ROC curve (AUBEDROC) for binary classification models aimed at discriminating DUD-E actives and decoys from docking poses scored by two machine learning models (RF-IChem, SVM-IChem) and a prototypical scoring function (Surflex-Dock). Abbreviations for the 10 representative targets are given as follows: adenosine A2A receptor, AA2AR; Beta2 adrenergic receptor, ADRB2; androgen receptor, ANDR; glucocortocoid receptor, GCR; adenosine deaminase, ADA; prostaglandin G/H synthase 2, PGH2; angiotensin-converting enzyme, ACE; renin, RENI; fibroblast growth factor receptor 1, FGFR1; and RAC-alpha protein kinase, AKT1. The quality of the model (fail, poor, fair, good, excellent) is determined by thresholds of both ROC[21] and BEDROC[26] areas under the curve.

stuctures (see preceding section), would only be applicable to one target (renin) for a docking-based virtual screening.

RF-IChem and SVM-IChem Are Insensitive to Docking Pose Accuracy (Docking Power Test). How could a scoring function nicely predict the pK_i values for 195 diverse protein−ligand X-ray structures and consistently fail in distinguishing true actives from presumably inactives (decoys) in several independent virtual screens? Apart from the possibility that some decoys may have been incorrectly chosen and be true actives if they had to be tested experimentally (this scenario however applies to any scoring function), it might be possible that the machine learning-based scoring functions may only be applicable to high-resolution protein−ligand structures on which they have been trained, and not to lower resolution docking poses. To ascertain this hypothesis, we decided to challenge the ML-based scoring functions on the PDBBind docking power test.[16] For each entry of the PDBBind core set, the ligand X-ray pose as well as up to 100 carefully selected docking decoys[16] have been scored and ranked by decreasing docking score. Success is evaluated by the ability of the scoring function to rank near-native decoys (low rmsd to the X-ray pose) before meaningless decoys (higher rmsd). As to be expected, deviation of docking poses from native solutions is penalized by Surflex-Dock, as a function of the progressive loss of attractive interactions and the number of steric and electrostatic clashes that occur. Most docking decoys are assigned a predicted pK_i value much lower than the experimental one (Figure 3A).

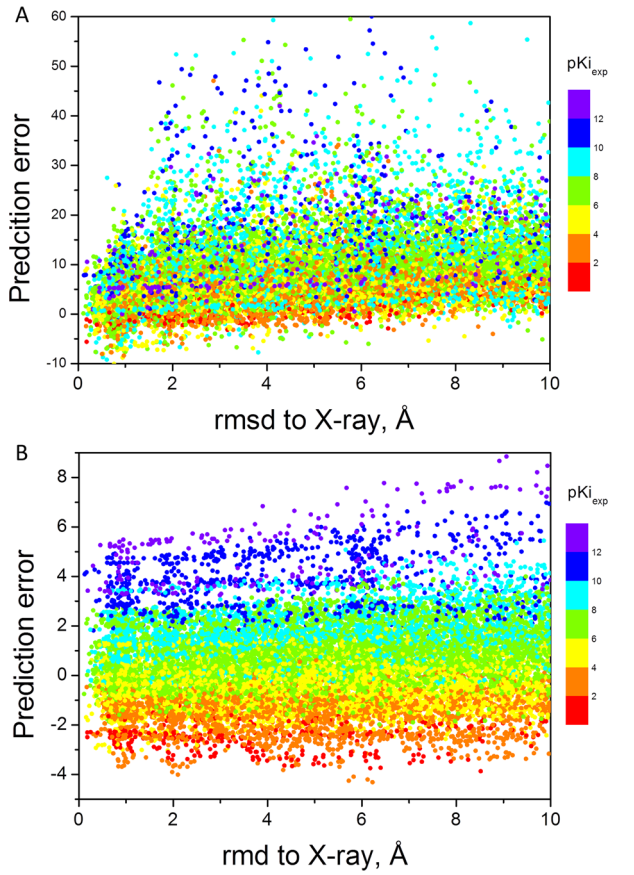


Figure 3. Error in predicting the inhibition constant (experimental pK_i − predicted pK_i), with respect to the ligand pose accuracy, as determined by the rmsd to the native X-ray pose. For each ligand of the PDBBind core set (n = 195), decoy poses of increasing rmsd were scored with (A) Surflex-Dock or (B) RF-IChem. Data points are colored according to the experimental pK_i value of the corresponding protein−ligand complex.

Intriguingly, RF scores appear to be totally independent of the rmsd to the X-ray pose (Figure 3B) and follow a layered distribution, depending on the difference between the experimental pK_i and a mean value at 5−6 pK_i unit (Figure 3B). For a low affinity complex (e.g., 1hi4, experimental pK_i = 4.50), all poses are overestimated to a pK_i value between 4.5 and 5.5, regardless of the rmsd to the X-ray pose (Figure 4A). For medium affinity complexes (e.g., 1xgj, experimental pK_i = 6.00), predictions are all in the 4.5−5 pK_i range. Last, a high affinity complex (e.g., 1sqa, experimental pK_i = 9.21) is systematically underestimated, to a mean pK_i value of ca. 6.50 (Figure 4B).

The top-ranked pose, as scored by both ML-based scoring functions, is rarely close to the X-ray pose (Figure 5A). Considering a rmsd of 1 Å between the top-ranked pose and the true X-ray pose, RF-IChem and SVM-IChem pass the

dx.doi.org/10.1021/ci500406k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

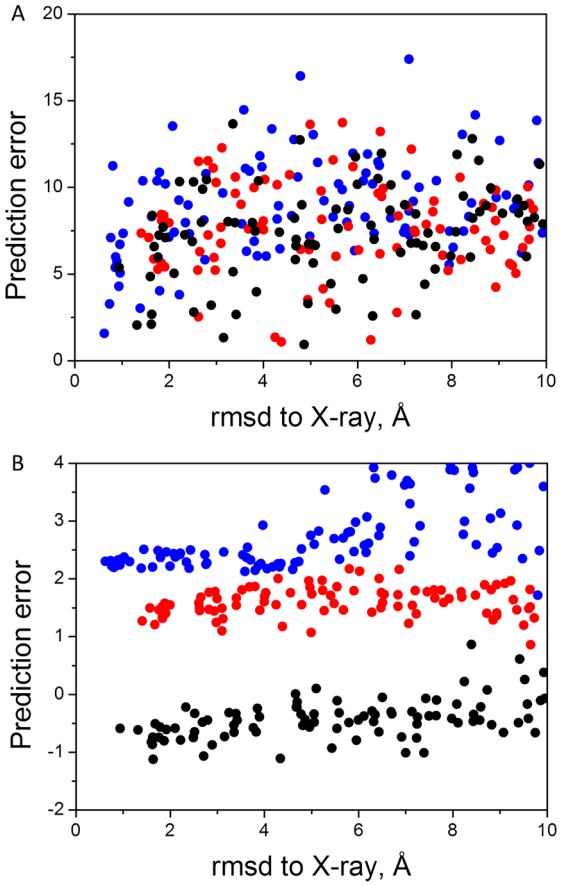


Figure 4. Error in predicting the inhibition constant (experimental pK_i − predicted pK_i), with respect to the ligand pose accuracy, for three complexes (1hi4, pK_{iexp} = 4.5, black dots; 1xgj, pK_{iexp} = 6.00, red dots; 1sqa, pK_{iexp} = 9.21, blue dots). For each ligand, decoy poses of increasing rmsd to the X-ray pose were scored with (A) Surflex-Dock or (B) RF-IChem.

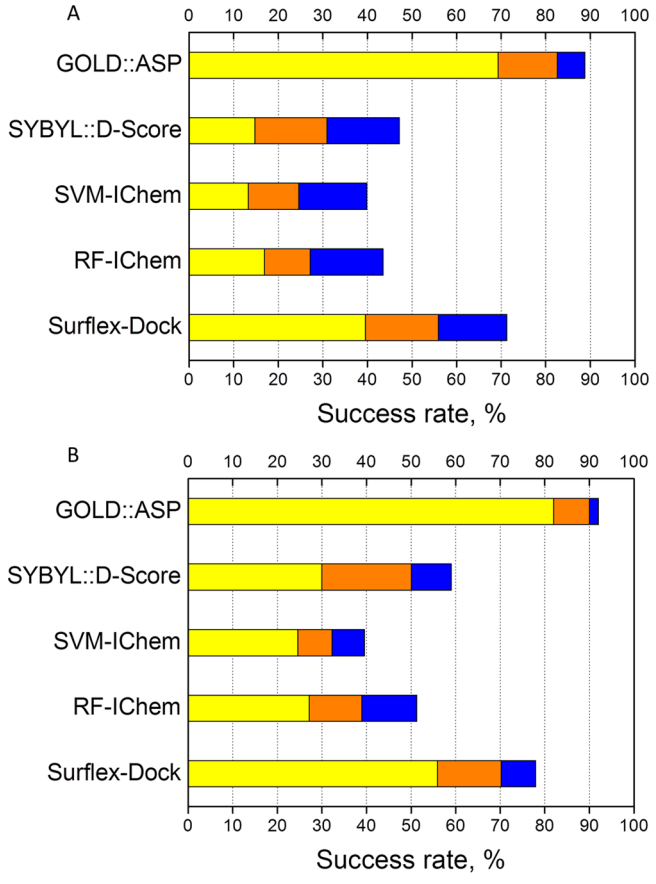


Figure 5. Comparison of the success rate of 5 scoring functions on the PDBBind (v.2007) docking power test, with the X-ray pose included in the decoy sets. GOLD::ASP and SYBYL::D-score are the two extreme scoring functions (best and worse, respectively) of the 16 scoring schemes originally investigated:[16] (A) Success rate in selecting a top-ranked pose within 1 Å (yellow), 2 Å (orange), and 3 Å (blue) rmsd to the X-ray pose. (B) Success rate in selecting the top one (yellow), the top two (orange), and the top three (blue) best-scored poses within a rmsd of 2 Å to the X-ray pose.

docking power test in only 17% and 13% of the cases. Increasing the rmsd threshold to 2 and 3 Å, the success rate of both ML-based scoring function remains very weak (Figure 5A) and even inferior to that of the less-performing scoring function (SYBYL D-score) among the 16 evaluated in the original PDBBind benchmark.[16] By comparison, observed success rates for Surflex-Dock are much higher (40% at rmsd = 1 Å, 56% at rmsd = 2 Å, 72% at rmsd = 3 Å; Figure 5A), which locate the Surflex-Dock scoring function between the worse (SYBYL D-score) and the best (GOLD ASP) of the 16 benchmarked scoring functions in PDBBind.[16] Selecting not only the top one but the top two or top three ranked poses, success rates are logically increased but with no major changes in the relative hierarchy of investigated scoring schemes (Figure 5B). Both ML-based scoring functions are the two less-performing, with respect to either Surflex-Dock or the 16 functions benchmarked by Cheng et al.[16]

If both ML-based functions are very powerful in a scoring scenario (predicting binding constants) using X-ray structures, they are unambiguously not suited to score docking poses. Inspecting the predicted pK_i values of DUD-E docking poses (both actives and decoys) confirm this observation, whatever the 10 investigated targets. RF-IChem and SVM-IChem scores share a very tiny distribution (6.5−7.5 pK_i units), whereas Surflex-Dock scores vary over a much broader range of values logically describing true actives (high scores) and inactives (low scores, Figure 6). Both ML-based scoring functions are clearly biased to predict pK_i values close to the mean pK_i value of the 1105 complexes (mean = 6.36), upon which the machine learning methods have been trained.

The observed insensitivity of the RF-IChem scoring function to the ligand pose suggests that the descriptors (protein−protein element−element distance counts) do not describe true protein−ligand interactions. To ascertain this hypothesis, we repeated the RF learning on the 1105 PDBBind complexes, but using a single distance bin (either 0−2 Å, 2−4 Å, 4−6 Å, 6−8 Å, 8−10 Å, or 10−12 Å) for counting protein−ligand element−element distances. Excepted for the RF model trained on protein−ligand clashes (0−2 Å distance bin), all other RF models perform similarly well, when predicting the 195 binding constants of the core set, with a Pearson correlation between 0.70 and 0.77 (see Figure 7). Therefore, here, we demonstrate that the machine learning model is not learning any type of protein−ligand interaction, since we do not observe a drop in performance when descriptors are focused only on very weak and long-range interactions (e.g., 10−12 Å distance count). The independency of the RF model of the protein−ligand distance counts used for training suggests that the model just learns protein and ligand atom counts separately, and not in conjunction with possible protein−ligand interactions.

dx.doi.org/10.1021/ci500406k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

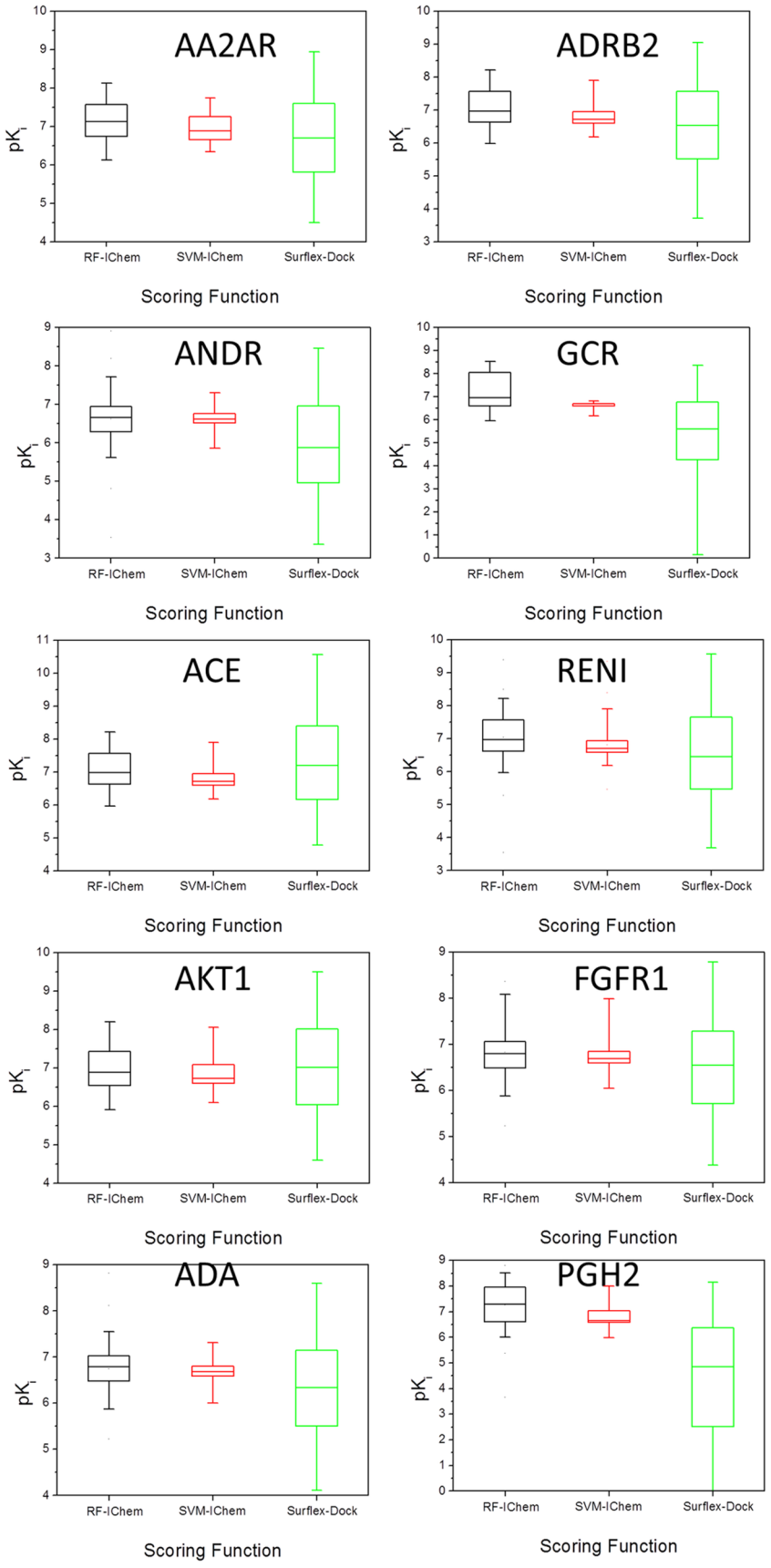


Figure 6. Distribution of predicted binding constants (pK_i units) for docking poses of 10 target-specific active and decoy sets. The box delimit the 25th and 75th percentiles, the whiskers delimit the 5th and 95th percentiles. The median values are indicated by a horizontal line in the box. Target names are abbreviated as in Figure 2.

dx.doi.org/10.1021/ci500406k | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

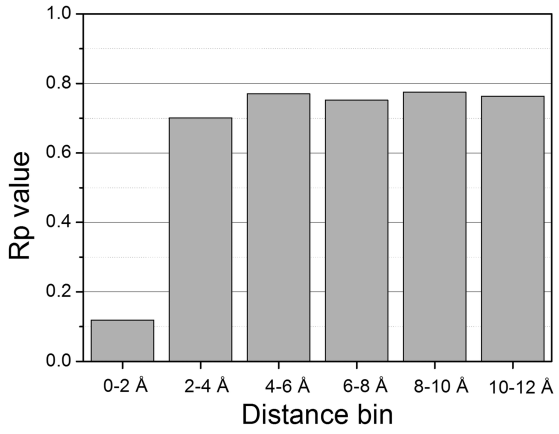


Figure 7. Accuracy of RF models estimated by the Pearson correlation coefficient (R_p), trained on a single bin registering protein−ligand element−element distance counts, in predicting the binding constant of 195 PDBBind (v.2007) complexes of the core set.

On the Danger of Predictive Modeling with Irrelevant Metrics and Descriptors. Several examples highlighting the misuse of descriptors and/or metrics in predictive modeling have already been reported. One of the first demonstrations that computer modeling may predict the good biological property for an incorrect reason was provided upon docking-based virtual screening of the Cambridge structural database for searching HIV-1 protease inhibitors. Among the select hits, the antipsychotic haloperidol was indeed confirmed experimentally as a weak nonpeptide HIV-1 protease inhibitor and further optimized to a promising lead.[30] Later, co-crystallization of the lead with the intended target however revealed that the true binding mode was markedly different from that predicted by docking (rmsd of ca. 5 Å).[31] In other words, experimental conformation of a structure-based virtual screening hit does not guarantee that the compound has been selected for the good reasons. 3D-QSAR is also prone to severe misinterpretations and errors. Muegge et al. reported, for a series of MMP-3 inhibitors, that 3D-QSAR models based on protein-free ligand alignments were markedly different from alignments derived from the corresponding protein−ligand X-ray structures, but nevertheless provided better statistics (cross-validated correlation coefficient Q^2 in predicting binding affinities) than alignments derived from co-crystals.[32] Similarly, Golbraikh et al. demonstrated, for several independent datasets, the lack of correlation between the leave-one-out cross-validated correlation coefficients (Q^2) of a QSAR model for a training set and its predictive ability for an external test set.[33]

We herewith provide another example in the field of training machine learning algorithms to predict protein−ligand binding constants. One well-known drawback of machine learning models is their lack of interpretability, although workarounds have been reported.[13] Therefore, it is of the utmost importance to ascertain that the property predicted by any predictive modeling is quantitatively well predicted for meaningful reasons. Although classical linear regression-based scoring functions suffer from a lack of accuracy, they present the advantage to be easily interpretable just by looking at the physical meaning of either positive or negative weights assigned to favorable and unfavorable energy terms, respectively. This interpretation is much more difficult with more-sophisticated nonlinear models. Despite excellent statistics in predicting binding constants (good correlation coefficient after cross-validation on an external test set, success of a *y*-scrambling test),[14,15] very simple descriptors (protein−ligand element−element distance counts) unfortunately do not allow a RF or SVM model to meaningfully predict binding constants from protein−ligand atomic coordinates. The main reason for this discrepancy lies in the disconnection of two issues that should be addressed simultaneously: (i) on one hand, the pure scoring accuracy of the machine learning model, and (ii) on the other hand, its accuracy in a docking-based virtual screening experiment.

One should not forget that the basic objective of a scoring function is not to predict (usually known) binding constants for high-resolution X-ray structures, but to predict the binding constant of novel ligands docked to an existing protein X-ray structure; or even more difficult, the binding constant of novel ligands to a protein of unknown X-ray structure (e.g., homology model). In absence of any rigorous validation in a virtual screening scenario, the quality of any given scoring function notably those arising from sophisticated machine learning methods remains questionable. Therefore, it is of the utmost importance to multiply quality checks and rely on publicly available benchmark datasets like PDBBind,[16,34] CSAR,[6] or the herein-presented virtual screening power set, to faithfully evaluate novel scoring functions.

■ CONCLUSIONS

The current study highlights the danger of training sophisticated machine learning algorithms on irrelevant descriptors. We certainly do not want to blame the two seminal reports[14,15] proposing the usage of protein−ligand element−element distance counts to train a machine learning-based regression models. The studies were carefully conducted using a state-of-the-art dataset and a rigorous computational setup. However, the final conclusion—*that an RF model trained on protein−ligand element−element distance counts is able to accurately predict binding constants*—is wrong, because of insufficient validations, notably with respect to the biophysical meaning of the model. We notably suggest, in addition to the classical prediction of binding constants from protein−ligand complexes, two series of additional benchmarking tests assessing (i) the dependency of the scoring function to the ligand pose quality and (ii) their capacity to enrich hit lists in true actives upon structure-based virtual screening of reference targets. Given the increasing popularity of machine learning methods, we hope that the guidelines proposed herein will positively assist our community in the development of more-robust scoring functions to predict binding constants from protein−ligand atomic coordinates.

■ ASSOCIATED CONTENT

Supporting Information

Curated input files (mol2 file format) of protein and ligand structures of the PDBBind core set; ligand, protein input files and Surflex-Dock poses (10/ligand) for 10 DUD-E target sets. This material is available free of charge via the Internet at http://pubs.acs.org

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: +33 3 68 85 42 35. Fax: +33 3 68 85 43 10. E-mail: rognan@unistra.fr.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work has been published within the LABEX ANR-10-LABX-0034_Medalis and received a financial support from French government managed by "Agence Nationale de la Recherche" under "Programme d'investissement d'avenir".

■ REFERENCES

(1) Bohm, H. J. 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.* **1994**, *8*, 243−256.
(2) Damm-Ganamet, K. L.; Smith, R. D.; Dunbar, J. B., Jr.; Stuckey, J. A.; Carlson, H. A. CSAR benchmark exercise 2011−2012: evaluation of results from docking and relative ranking of blinded congeneric series. *J. Chem. Inf. Model.* **2013**, *53*, 1853−1870.
(3) Smith, R. D.; Dunbar, J. B., Jr.; Ung, P. M.; Esposito, E. X.; Yang, C. Y.; Wang, S.; Carlson, H. A. CSAR benchmark exercise of 2010: Combined evaluation across all submitted scoring functions. *J. Chem. Inf. Model.* **2011**, *51*, 2115−2131.
(4) Plewczynski, D.; Lazniewski, M.; Augustyniak, R.; Ginalski, K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *J. Comput. Chem.* **2011**, *32*, 742−755.
(5) Ripphausen, P.; Nisius, B.; Peltason, L.; Bajorath, J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J. Med. Chem.* **2010**, *53*, 8461−8467.
(6) Dunbar, J. B., Jr.; Smith, R. D.; Damm-Ganamet, K. L.; Ahmed, A.; Esposito, E. X.; Delproposto, J.; Chinnaswamy, K.; Kang, Y. N.; Kubish, G.; Gestwicki, J. E.; Stuckey, J. A.; Carlson, H. A. CSAR data set release 2012: Ligands, affinities, complexes, and docking decoys. *J. Chem. Inf. Model.* **2013**, *53*, 1842−1852.
(7) Dunbar, J. B., Jr.; Smith, R. D.; Yang, C. Y.; Ung, P. M.; Lexa, K. W.; Khazanov, N. A.; Stuckey, J. A.; Wang, S.; Carlson, H. A. CSAR benchmark exercise of 2010: Selection of the protein−ligand complexes. *J. Chem. Inf. Model.* **2011**, *51*, 2036−2046.
(8) Sotriffer, C. A.; Sanschagrin, P.; Matter, H.; Klebe, G. SFCscore: Scoring functions for affinity prediction of protein-ligand complexes. *Proteins* **2008**, *73*, 395−419.
(9) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein−ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726−741.
(10) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind database: Methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.
(11) Li, L.; Wang, B.; Meroueh, S. O. Support vector regression scoring of receptor−ligand complexes for rank-ordering and virtual screening of chemical libraries. *J. Chem. Inf. Model.* **2011**, *51*, 2132−2138.
(12) Muegge, I.; Oloff, S. Advances in virtual screening. *Drug Discovery Today: Technol.* **2006**, *3*, 405−411.
(13) Zilian, D.; Sotriffer, C. A. SFCscore(RF): A random forest-based scoring function for improved affinity prediction of protein-ligand complexes. *J. Chem. Inf. Model.* **2013**, *53*, 1923−1933.
(14) Ballester, P. J.; Schreyer, A.; Blundell, T. L. Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity? *J. Chem. Inf. Model.* **2014**, *54*, 944−955.
(15) Ballester, P. J.; Mitchell, J. B. A machine learning approach to predicting protein−ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26*, 1169−1175.
(16) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.* **2009**, *49*, 1079−1093.
(17) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(18) http://www.pdbbind-cnr.org/download/CASF-2007.tar.gz (accessed June 2014).
(19) *SYBYL, version X2.1*; Certara: St. Louis, MO, 2012.
(20) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* **2012**, *55*, 6582−6594.
(21) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*, 2534−2547.
(22) Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. Encoding protein-ligand interaction patterns in fingerprints and graphs. *J. Chem. Inf. Model.* **2013**, *53*, 623−637.
(23) Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18−22.
(24) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2011 (ISBN 3-900051-07-0).
(25) Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499−511.
(26) Truchon, J. F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the ″early recognition″ problem. *J. Chem. Inf. Model.* **2007**, *47*, 488−508.
(27) Swamidass, S. J.; Azencott, C. A.; Daily, K.; Baldi, P. A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. *Bioinformatics* **2010**, *26*, 1348−1356.
(28) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455−1474.
(29) Ben Nasr, N.; Guillemain, H.; Lagarde, N.; Zagury, J. F.; Montes, M. Multiple structures for virtual ligand screening: Defining binding site properties-based criteria to optimize the selection of the query. *J. Chem. Inf. Model.* **2013**, *53*, 293−311.
(30) DesJarlais, R. L.; Seibel, G. L.; Kuntz, I. D.; Furth, P. S.; Alvarez, J. C.; Ortiz de Montellano, P. R.; DeCamp, D. L.; Babe, L. M.; Craik, C. S. Structure-based design of nonpeptide inhibitors specific for the human immunodeficiency virus 1 protease. *Proc. Natl. Acad. Sci. U.S.A* **1990**, *87*, 6644−6648.
(31) Rutenber, E.; Fauman, E. B.; Keenan, R. J.; Fong, S.; Furth, P. S.; Ortiz de Montellano, P. R.; Meng, E.; Kuntz, I. D.; DeCamp, D. L.; Salto, R.; et al. Structure of a non-peptide inhibitor complexed with HIV-1 protease. Developing a cycle of structure-based drug design. *J. Biol. Chem.* **1993**, *268*, 15343−15346.
(32) Muegge, I.; Podlogar, B. L. 3D-Quantitative Structure Activity Relationships of Biphenyl Carboxylic Acid MMP-3 Inhibitors: Exploring Automated Docking as Alignment Method. *Quant. Struct.-Act. Relat.* **2001**, *20*, 215−222.
(33) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graph. Model.* **2002**, *20*, 269−276.
(34) Li, Y.; Liu, Z.; Li, J.; Han, L.; Liu, J.; Zhao, Z.; Wang, R. Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set. *J. Chem. Inf. Model.* **2014**, *54*, 1700−1716.