JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

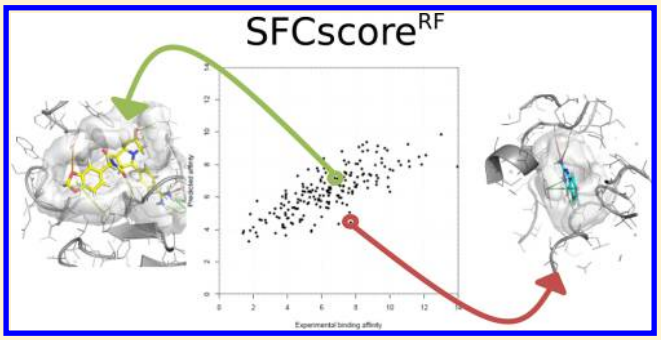
SFCscore^{RF}: A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein−Ligand Complexes

David Zilian and Christoph A. Sotriffer*

Institute of Pharmacy and Food Chemistry, University of Wuerzburg, Am Hubland, D-97074 Wuerzburg, Germany

Supporting Information

ABSTRACT: A major shortcoming of empirical scoring functions for protein−ligand complexes is the low degree of correlation between predicted and experimental binding affinities, as frequently observed not only for large and diverse data sets but also for SAR series of individual targets. Improvements can be envisaged by developing new descriptors, employing larger training sets of higher quality, and resorting to more sophisticated regression methods. Herein, we describe the use of SFCscore descriptors to develop an improved scoring function by means of a PDBbind training set of 1005 complexes in combination with random forest for regression. This provided SFCscore^{RF} as a new scoring function with significantly improved performance on the PDBbind and CSAR−NRC HiQ benchmarks in comparison to previously developed SFCscore functions. A leave-cluster-out cross-validation and performance in the CSAR 2012 scoring exercise point out remaining limitations but also directions for further improvements of SFCscore^{RF} and empirical scoring functions in general.



INTRODUCTION

Scoring functions are essential components of structure-based drug design and virtual screening. The purpose of their application is either the evaluation of docking poses, ranking of compounds with respect to the quality of their interaction with a biomolecular target, or explicit estimation of protein−ligand binding affinities. Comparative and retrospective studies show relatively high success rates for the identification of correct binding modes.[1] More error prone, however, is the accurate prediction of binding affinities and the correct ranking of compounds by their putative activity.[2−5] Despite improvements over the last years, most scoring functions, regardless whether they are empirical, force field based, or knowledge based, still suffer from a rather poor correlation with experimental binding affinity.[6,7]

Empirical scoring functions rely on the premise that the affinity between a protein and a ligand can be estimated as a sum of individual contributions or interaction terms. These are evaluated in terms of geometric descriptors of localized chemically intuitive interactions (such as hydrogen bonds) or entropic factors (such as the freezing of rotatable bonds). The weighting coefficient of each descriptor is obtained by regression analysis or related statistical methods using a training set of experimental affinity values for protein−ligand complexes of known structure.

SFCscore is an example of such an empirical scoring function.[8] It was introduced as a collection of eight alternative functions derived from large differently compiled training sets of public and industrial origin. The functions were obtained by means of multiple linear regression (MLR) or partial least squares (PLS) analysis based on classical and newly added geometric descriptors characterizing protein−ligand interactions. Although by the time of their development the performance of the SFCscore functions was clearly superior to most of the then available scoring functions,[8] the correlation with experimental affinity values across large data sets was still unsatisfying. Accordingly, efforts to further improve affinity prediction with scoring functions in general and SFCscore in particular are highly warranted.

Improvements of empirical scoring functions can be envisaged along three lines: first, by developing new descriptors; second by training on larger high-quality data sets of protein−ligand structures with experimentally determined affinities; and third, by using alternative methods for regression analysis. In the work presented herein, we retain the previously developed descriptors and address the second and third issue in order to improve SFCscore.

With respect to the regression technique, we resort to random forest as a machine learning approach. As illustrated by an increasing number of applications in the literature, machine learning algorithms (such as random forest, Bayesian classifiers, and support vector machines (SVM)), can successfully be applied in virtual screening scenarios in the context of scoring tasks. For example, Jorissen and Gilson trained an SVM model

Special Issue: 2012 CSAR Benchmark Exercise

Received: February 19, 2013
Published: May 25, 2013

based on two-dimensional molecular descriptors to enrich a molecular database with potential actives.[9] Amini et al. used SVMs in combination with inductive logic programming (ILP) to obtain rule-based scoring functions for specific targets.[10] Li et al. used a support vector regression (SVR)-based model to rank the community structure−activity resource (CSAR) data set and achieved very good results.[11] Ballester and Mitchell applied the random forest algorithm to derive the RF-Score function, which showed significantly higher correlations between predicted and experimental affinities for the PDBbind benchmark set than all other functions previously tested on this data set.[12] This seminal study prompted us to investigate the performance of random forest in combination with the SFCscore descriptors, leading to SFCscore^{RF}, as further described below.

As far as the data sets for training and testing are concerned, we used a PDBbind set[13,14] of more than 1000 complexes for training and applied the common PDBbind benchmark set[4] and the CSAR−NRC HiQ set[15] for testing and evaluation. This allows for an unbiased comparison of the SFCscore functions among each other and in comparison to other scoring functions already tested against these benchmarks. Furthermore, the most recent edition of the CSAR exercise (CSAR 2011−2012 Benchmark Exercise) provided the welcome opportunity to investigate the performance of all SFCscore functions in real-life scenarios and to draw conclusions for further improvement.

MATERIALS AND METHODS

Data Sets. *Generic Data Sets.* A large and diverse data set of high-quality crystal structures in combination with experimentally determined binding affinities (K_i or K_d values) is of utmost importance for the development of empirical scoring functions. In the present study, the PDBbind database[14,13] and the CSAR−NRC 2010 set of complexes[15] served as data sources for training and testing.

For the derivation of SFCscore^{RF}, the PDBbind refined set v2007 was used after removal of all complexes occurring in the predefined benchmarks (test sets) described below. Accordingly, the 195 complexes of the PDBbind core set and an additional 99 complexes occurring in the CSAR−NRC set were removed from the training set. In addition, one complex could not be processed for the SFCscore descriptor calculation and had to be eliminated, thus leaving 1005 complexes for the training of the model. This data set will be referred to as *training set*. The complete list of PDB codes is provided in Table S1 of the Supporting Information.

For comparative evaluation of the scoring functions on generic data sets of protein−ligand complexes, the PDBbind benchmark set[4] consisting of 195 complexes was used (in the following referred to as *PDBbind test set*). In addition, the CSAR−NRC HiQ set (referred to as *CSAR test set*) was applied for testing.[15] To make the results comparable with the extensive analysis of Smith et al., the 11 Factor Xa complexes were removed for the reasons explained in that study.[6] This left 332 complexes in the CSAR test set.

CSAR 2012 Exercise Data Sets. For further evaluation, the CSAR 2012 exercise data sets for the three targets CHK1, ERK2, and LpxC were used.

CHK1 (checkpoint kinase-1) is a kinase involved in the cell cycle control.[16] The data set comprises 47 compounds targeted against the ATP binding site. The compounds belong to three structural classes: aromatic tricycles with an imidazole substructure, tricyclic scaffolds with a central seven-membered ring lactame, and differently substituted urea derivatives. ERK2 (extracellular-signal-regulated kinase 2) is also involved in many stages of the cell cycle.[17,18] The 39 compounds of the data set address the ATP binding pocket. Three ligand series can be distinguished: pyrazoles, guanidines, and a series with no common substructure. LpxC, the UDP-3-O-(*R*-3-hydroxymyristoyl)-N-acetylglucosamine deacetylase, is a metalloamidase involved in the bacterial lipopolysaccharide synthesis.[19,20] The binding pocket contains a zinc ion as an essential cofactor. The data set consists of 16 structurally very similar compounds containing either a hydroxamic acid or a hydantoin group for metal complexation.

Validation Data Sets for CSAR 2012 Exercise. For all three targets, additional data sets were created for validating and selecting the docking and scoring procedure to be applied on the CSAR exercise data sets. The corresponding data were collected from the PDB[21] and ChEMBL[22] databases. For validating the docking procedure, the PDB was searched for crystal structures of target−inhibitor complexes with a resolution better than 2.5 Å. The resulting hits were visually inspected to ensure that binding pockets and ligands bound therein were fully resolved. This resulted in 11 structures for CHK1, 10 structures for ERK2, and only 3 for LpxC.

To estimate the performance of the scoring functions on the three targets, a scoring validation set was created for each target by searching the ChEMBL database for compounds with experimental inhibition constants (K_i values). This yielded 13 entries for CHK1, spanning a pK_i range from 4.7 to 9.6. For ERK2, 41 compounds were found, covering a range between 4.6 and 8.6 in terms of pK_i. For LpxC, 42 compounds with IC_{50} values measured in the same assay were obtained (the pIC_{50} ranging between 4.2 and 6.9).

Data Preparation for SFCscore^{RF} Development. For descriptor calculation with SFCscore,[8] proteins, ligands, and cofactors need to be provided as separate files. In the training set, PDBbind test set, and CSAR test set, the ligands were directly available as separate files. Cofactors were extracted with SYBYL-X (SYBYL-X 1.0, Tripos International, St. Louis, Missouri, U.S.A.) as mol2 files. All water molecules, ions, or buffer components were deleted from the crystal structures. Descriptors were calculated directly from these files without any further optimization.

SFCscore provides a collection of 66 descriptors for representing protein−ligand complexes, including purely ligand-dependent descriptors (such as the number of rotatable bonds), descriptors for specific interactions (such as hydrogen bonds or aromatic interactions), and surface characteristics (such as polar or hydrophobic contact surfaces). A full list with detailed definitions of the descriptors is provided in the SFCscore publication.[8] For the study presented herein, the descriptors MW (molecular weight), NAtoms (number of heavy atoms), and TotLigSurf (total ligand surface) were excluded because of their frequently observed trivial correlation with the affinity irrespective of the actual quality of interaction. This left 63 descriptors for calibration of new functions.

Random Forest for Regression. In its original form, SFCscore is a collection of eight empirical scoring functions based either on multiple linear regression (MLR) or partial least squares (PLS) analysis.[23] To overcome some of the general limitations of these methods, the random forest (RF) approach was followed to train new regression models. RF is an ensemble learning method based on a large number of decision trees.[24] The number is tunable, but typically 500 or more trees

are used. The performance does not sensitively depend on the tree number. Depending on the nature of the underlying trees, the method can be used for classification or for regression. The prediction of a test object is based on the majority votes (classification) or obtained by averaging the prediction of all trees (regression).

A single decision tree is grown by splitting the initial data set into subsets that have a higher enrichment of the target variable at each node of the tree. In the case of a regression tree, the splitting criterion is the minimization of the sum of squares of the target variable in each subset. In RF, each of the trees is trained on a different subset of the training set (approximately 60%) and at every splitting node with a different subset of variables (descriptors). This adds variability to the model and is the main reason for the improved robustness of RF compared to a single decision tree. The parameter m_{try}, the number of variables used at each splitting node, is the only tunable parameter that significantly influences the performance of the model. Default values are $p/3$ for regression and $p^{1/2}$ for classification trees (where p is the total number of variables available). Empirical experiments suggest that the model performance is stable over a wide range of m_{try} values.[25]

As an intrinsic method for internal validation, the RF algorithm offers the out-of-bag (OOB) prediction. For each tree, about one-third of the training set objects are "out-of-bag" because they are not used for training. For the OOB prediction, every complex in the training set is predicted only by those trees for which it has not been used for training. Svetnik et al. showed that the OOB estimations are in good agreement with the results of a cross-validation.[25] Here, we used the error of the OOB predictions to assess the optimal value of m_{try}. A systematic search was performed, evaluating the OOB error for every possible m_{try} value. The algorithm of this systematic search is based on the code provided by Ballester and Mitchell in the Supporting Information.[12] The search resulted in an optimal m_{try} of 21, which is in good agreement with the rule of thumb suggesting one-third of the number of descriptors as a suitable value.

The derivation and analysis of the RF models was carried out with R,[26] using the randomForest package.[27] The descriptor tables for all data sets used in this study are in the Supporting Information (with an explanation in Table S2). The applied R code is available from the authors upon request.

Data Preparation for CSAR 2012 Exercise. *Generation of Docking Poses.* Binding poses were generated with the docking program Glide (version 5.7, Schrödinger, LCC, New York, 2011).[28−30] The ligand setup was carried out with LigPrep (version 2.5, Schrödinger, LLC, New York, 2011). Ligand preparation included the conversion to 3D structures (where necessary), generation of low-energy ring conformations, and protonation and tautomerization with Epik (version 2.2, Schrödinger, LLC, New York, 2011).[31,32] The pH was set according to the assay conditions (±0.5) provided with the CSAR data sets. For LpxC inhibitors, "metal binding states" were also calculated with Epik. As protein structures, the PDB codes suggested by the CSAR organizers were used (CHK1, PDB 2e9n;[16] ERK2, PDB 3i5z;[18] LpxC, PDB 3p3e[20]), unless otherwise noted. Protein preparation for docking was based on the Protein Preparation Wizard as integrated in Maestro.[33] All water molecules and compounds present from the crystallization buffer were removed. Hydrogens were added using standard protonation states and subsequent optimization of the hydrogen bond network. No manual intervention was made at this stage.

Upon grid generation for docking with Glide, hydroxyl groups in the binding pocket were kept rotatable; for LpxC a metal constraint was defined with respect to the catalytic zinc ion. Docking runs were performed in XP (extra precision) mode[30] using standard settings except for a few modifications. Epik state penalties were added to the docking score in all cases. For the CHK1 ligands, multiple low-energy ring conformers were docked. Ten docking poses were generated for each compound.

Pose Selection and Scoring. The docking poses obtained from Glide were rescored with DrugScoreX (DSX),[34] and the lists ranked by GlideScore and DSX were visually inspected for final pose selection. The most important criterion upon visual inspection was the consistency of binding modes for ligands belonging to the same series. The selected pose for each ligand was submitted to rescoring with SFCscore v1.1.[8] The scores of all available SFCscore functions were calculated but only the results of two for each target were submitted to the exercise based on the performance in the scoring validation. Nevertheless, the results of all SFCscore functions will be presented and discussed in the following.

RESULTS AND DISCUSSION

Derivation of SFCscore^{RF}. Using SFCscore descriptors and experimental pK_i values of the training set complexes as input data, SFCscore^{RF} was trained as a regression-based random forest model. The obtained regression and correlation statistics are listed in Table 1. The regression resulted in 0.972 for Pearson's correlation coefficient R and a median error of 0.33 pK_i units for the training set. Although this indicates that the model is well capable of correlating binding features in terms of the applied descriptors with experimental binding affinities, it is clear that training set regression statistics do not provide suitable estimates of the actual predictivity.

Table 1. Performance Measures of SFCscore^{RF} for the Training Set (Regression) and Two Test Sets (Correlation)[a]

	R	ρ	τ	RMSE	MedError
training set	0.972	0.974	0.865	0.60	0.33
training set (OOB)	0.707	0.716	0.530	1.49	0.85
CSAR test set	0.730	0.716	0.533	1.53	0.96
PDBbind test set	0.779	0.788	0.592	1.56	1.06

[a] R is Pearson's correlation coefficient. ρ is Spearman's correlation coefficient. τ is Kendall's tau. RMSE is root mean square error. MedError is Median of residuals. OOB is out-of-bag.

Accordingly, for internal validation the OOB predictions were calculated, leading to Pearson's R of 0.707 and a median error of 0.85 (Table 1). These values provide a more realistic estimate of the actual model performance and compare well with the validation results obtained for the external test sets described below. Figure 1 illustrates that the deviations of the OOB predictions are largest in the high and low affinity areas. The tendency to best predict affinities in the middle range is frequently observed for regression-based (empirical) scoring functions and an inherent problem of the approach as long as training objects are most abundant for the micro- to nanomolar affinity range (pK_i between 4 and 9) and rather sparse for higher or lower values.

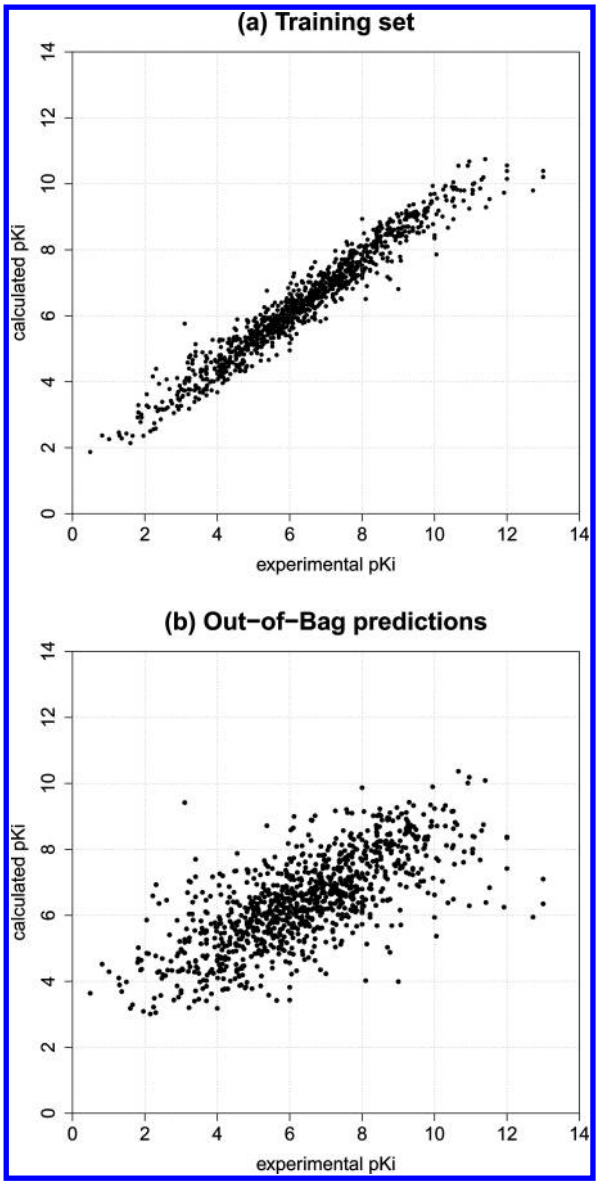


Figure 1. Correlation between experimentally measured and calculated affinities for the training set in terms of the pK_i values: (a) complete regression model and (b) out-of-bag predictions.

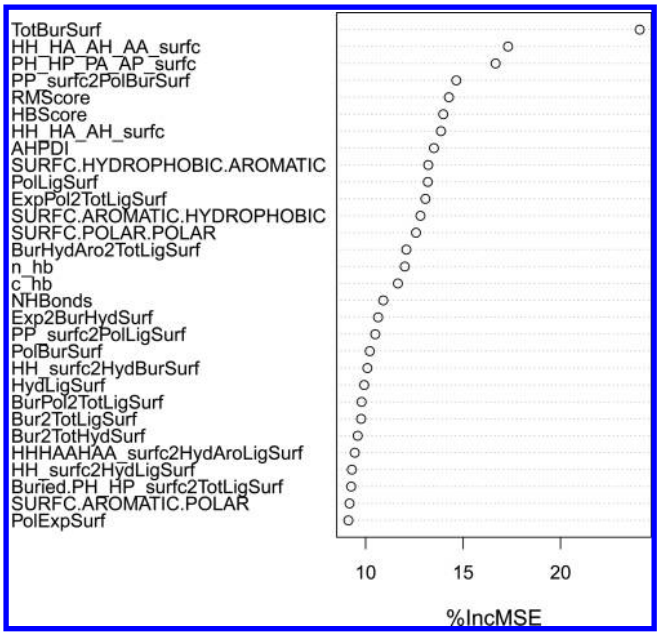


Figure 2. Importance of single descriptors measured as the (normalized) increase of the mean standard error (IncMSE given in %) when the variable listed in the left column is randomly permuted (descriptor abbreviations are listed and explained in the work introducing SFCscore).[8]

Unlike all previous SFCscore functions, SFCscore^{RF} is trained with almost the full set of available descriptors. As no a priori dimensionality reduction was carried out, this implies that also correlated descriptors are contained in the data set. For the training set, a total of 12 groups (9 pairs, 3 triplets) of descriptors show a Pearson correlation coefficient of >0.85. The RF algorithm, however, is able to handle this high dimensionalty as well as correlated and unimportant descriptors.[25]

The interpretation of an RF model with respect to descriptor contributions is not as straightforward as in the case of MLR or PLS functions. Rather, the relevance of a particular descriptor is assessed indirectly by randomly permuting its values and measuring the decrease in accuracy in the prediction results. Plotting the increase in the mean standard error (MSE), expressed as percentage, for each descriptor in decreasing order provides the importance plot shown in Figure 2. TotBurSurf, the total buried surface, is by far the most important descriptor. This is not too surprising, given that it is part of all classical (MLR and PLS) SFCscore functions.[8] In contrast to the pure ligand surface or the molecular weight (which have been excluded as descriptors for the RF model), the total buried surface is not merely depending on ligand size but primarily on the shape complementarity between ligand and binding pocket and hence a fully valid descriptor of protein−ligand interactions. The three descriptors appearing at ranks 2−4 are more detailed measures of surface complementarity, including the sum of all hydrophobic (H) and aromatic (A) contact surfaces between the protein and ligand (HH_HA_AH_AA_surfc), sum of all polar (P)−hydrophobic (H) and polar (P)−aromatic (A) contact surfaces (PH_HP_PA_AP_surfc), and purely polar contact surface related to the polar buried surface of the ligand (PP_surfc2PolBurSurf). These are followed by the first nonsurface-related descriptors, i.e., RMScore (ring−metal interaction score) and HBScore (the total hydrogen bond score). Overall, this speaks for an interaction model that incorporates significant contributions from very different interaction types. Most of these descriptors appeared also with prominent contributions in at least some of the original SFCscore functions, even though n_hb (neutral hydrogen bond score) was generally preferred over HBScore and PP_surfc2PolLigSurf (polar contact surface related to the polar ligand surface) over PP_surfc2PolBurSurf. Surprisingly, no rotatable bond score appears under the top 30 descriptors in the importance plot of SFCscore^{RF}, whereas it is contained in most original SFCscore functions. Apparently, over the large training set, the rotatable bond scores contribute much less to the overall predictivity than generally assumed. Notably, this finding is not an artifact related to the presence of three correlated descriptors providing different measures of rotatable bonds (n_rot, NRotBonds, and RBScore). Rather, also in a model generated without NRotBonds and RBScore, the remaining descriptor n_rot leads to an increase in the MSE of only 9.4% (compared to 8.7% in the original model).

Evaluation of SFCscore^{RF} and Classical SFCscore Functions. SFCscore^{RF} and the original SFCscore functions were first evaluated with the PDBbind test set as a common benchmark for scoring functions.[4] The results are shown in Table 2. The original functions achieve values between 0.581 and 0.644 in terms of Pearson's R. This is generally in line with the evaluation results obtained previously on other large test sets,[8] indicating a rather moderate degree of correlation and predictivity. Nevertheless, these functions would all range among the top three when compared to the numerous scoring functions tested by Cheng et al., who obtained correlation coefficients between 0.216 and 0.644.[4] SFCscore^{RF} clearly outperforms these functions, reaching values of 0.779 and 0.788 for Pearson's R and Spearman's ρ, respectively (Figure 3). To the best of our knowledge, the best-performing function hitherto reported for this test set is the RF-Score function, which achieved values of 0.776 for R and 0.762 for ρ.[12] SFCscore^{RF} slightly exceeds these values, reaching essentially the same level of performance. Interestingly, RF-Score and SFCscore^{RF} differ only in the type of descriptors, whereas the training set and the machine learning method used for regression are basically the same. Apart from the slight performance advantage, the SFCscore descriptors are also more readily interpretable in terms of intermolecular interaction features than the occurrence counts between atom type pairs used by RF-Score.

Table 2. Performance of SFCscore^{RF} on the PDBbind Test Set in Comparison to the Original SFCscore Functions[a]

function	R	ρ	τ	RMSE	MedError
sfc_rf	0.779	0.788	0.592	1.56	1.06
sfc_229m	0.587	0.644	0.460	1.93	1.17
sfc_229p	0.589	0.631	0.454	1.96	1.16
sfc_290m	0.587	0.645	0.462	1.95	1.16
sfc_290p	0.620	0.664	0.475	1.88	1.03
sfc_855	0.602	0.635	0.451	1.92	1.15
sfc_frag	0.581	0.645	0.464	2.69	1.56
sfc_met	0.644	0.683	0.495	1.84	1.09
sfc_ser	0.588	0.629	0.448	1.97	1.14

[a] R is Pearson's correlation coefficient. ρ is Spearman's correlation coefficient. τ is Kendall's tau. RMSE is Root mean square error. MedError is median of residuals.

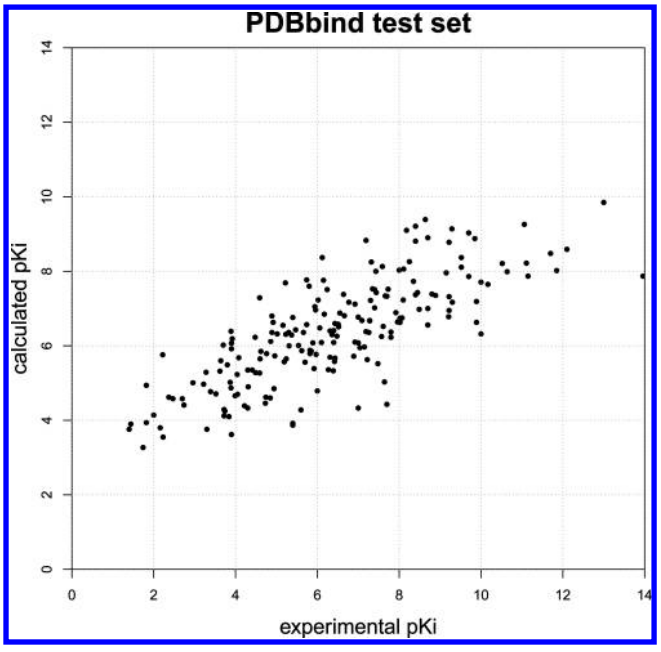


Figure 3. Correlation between SFCscore^{RF} predictions (calculated pK_i) and experimentally measured pK_i values for the PDBbind test set.

Further evaluation was carried out by means of the CSAR test set. The results for this larger and more recent benchmark (listed in Table 3) confirm the findings obtained with the PDBbind test set. The original SFCscore functions showing Pearson's correlation coefficients between 0.575 and 0.620 are clearly outperformed by SFCscore^{RF}, which reaches an R value of 0.730 and a median error of 0.96 pK units. The functions tested in the original study by Smith et al. show Pearson's correlation coefficients between 0.35 and 0.76.[6] SFCscore^{RF} would rank second among these functions. Even though the overall performance is respectable, it is also clear that at such a degree of correlation significant deviations between predicted and experimental binding affinities still exist. The overall picture is very similar as shown in Figure 3 for the PDBbind test set. The largest residuals occur in the high- and low-affinity ranges, but deviations are observed anywhere on the pK_i scale. Obviously, the performance can vary strongly depending on the nature of the complex and, hence, on the composition of the data set.

Table 3. Performance of SFCscore^{RF} on the CSAR Test Set in Comparison to the Original SFCscore Functions[a]

function	R	ρ	τ	RMSE	MedError
sfc_rf	0.730	0.716	0.533	1.53	0.96
sfc_229m	0.578	0.576	0.408	1.86	1.21
sfc_229p	0.575	0.563	0.395	1.86	1.28
sfc_290m	0.577	0.574	0.406	1.84	1.21
sfc_290p	0.578	0.583	0.410	1.84	1.20
sfc_855	0.593	0.589	0.419	1.78	1.10
sfc_frag	0.616	0.600	0.430	2.61	1.57
sfc_met	0.596	0.592	0.422	1.79	1.12
sfc_ser	0.620	0.618	0.438	1.76	1.17

[a] R is Pearson's correlation coefficient. ρ is Spearman's correlation coefficient. τ is Kendall's tau. RMSE is root mean square error. MedError is median of residuals.

Leave-Cluster-Out Cross-Validation. An impression of the target class-dependent performance of SFCscore^{RF} can be obtained from a "leave-cluster-out cross-validation". This method has been suggested by Kramer and Gedeck as a more thorough validation strategy particularly for generic scoring functions trained on the PDBbind database and tested against PDBbind benchmark sets.[35] In contrast to standard cross-validation, the complexes are not divided into random subsets but grouped into clusters according to the protein class assignment of the PDBbind. These clusters are then predicted as OOB subsets.[35] For comparability with the results obtained by Kramer and Gedeck, the PDBbind refined version 2010 was used for this purpose.[13,14] Each protein cluster was eliminated in turn from this data set, and SFCscore^{RF} was rederived from the remaining complexes and used to predict the eliminated cluster. The results of this leave-cluster-out validation of the SFCscore random forest models are listed in Table 4 and illustrated in Figure 4.

dx.doi.org/10.1021/ci400120b | J. Chem. Inf. Model. 2013, 53, 1923−1933

Table 4. Results of the Leave-Cluster-Out Cross-Validation as Proposed by Kramer and Gedeck[35a]

biological target	cluster	samples	R	R^2	RMSE
HIV protease	A	187	0.15	0.02	1.74
trypsin	B	73	0.76	0.58	0.91
thrombin	C	53	0.66	0.44	1.53
carbonic anhydrase	D	57	0.45	0.20	2.18
PTP1B	E	32	0.70	0.49	1.04
factor Xa	F	32	0.30	0.09	1.82
urokinase	G	29	0.60	0.37	1.19
different similar transporters	H	29	0.24	0.06	1.03
c-AMP dependent kinase	I	17	0.57	0.32	1.36
β-glucosidase	J	17	0.54	0.29	1.20
antibodies	K	16	0.77	0.59	1.54
casein kinase II	L	15	0.75	0.56	0.64
ribonuclease	M	15	0.61	0.38	1.03
thermolysine	N	14	0.79	0.63	0.93
CDK2 kinase	O	13	0.35	0.12	1.30
glutamate receptor 2	P	13	−0.10	0.01	1.00
P38 kinase	Q	13	0.65	0.42	0.74
β-secretase I	R	12	0.84	0.70	1.66
tRNA-guanine transglycosylase	S	12	−0.13	0.02	1.38
endothiapepsin	T	11	0.52	0.27	1.27
α-mannosidase 2	U	10	0.48	0.23	1.55
carboxypeptidase A	V	10	0.80	0.63	1.85
penicillopepsin	W	10	−0.50	0.25	2.02
clusters with 4−9 complexes	X	379	0.66	0.43	1.51
clusters with 2−3 clusters	Y	334	0.58	0.33	1.53
singletons	Z	316	0.47	0.22	1.69

[a] Every protein class (cluster) is predicted by a model trained with the remainder of the data set, i.e., all other protein classes. R is Pearson's correlation coefficient. RMSE is root mean square error.

Not unexpectedly, the observed performance is very heterogeneous across the different target classes. For some classes, very good predictions are obtained (e.g., trypsin, casein kinase II, and thermolysin, which show R values of at least 0.75 and RMSE values less than 1.0), while for other target classes, virtually no correlations (e.g., HIV protease, factor Xa), or even negative correlation coefficients (e.g., tRNA-guanine transglycosylase and penicillopepsin) are observed. Such findings are common for current scoring functions, indicating that even training sets consisting of more than 1000 complexes are not sufficient to adequately cover all features of molecular recognition in protein−ligand complexes. Kramer and Gedeck, in fact, obtained overall similar results in their leave-cluster-out cross-validation of RF-score.[35] Only 7 of the 23 target clusters A through W differ by more than 0.20 (absolute value) in R (with better values obtained by SFCscore^{RF} for thrombin, different similar transporters, casein kinase II, ribonuclease, and α-mannosidase 2, but worse values for CDK2 and tRNA-guanine transglycosylase) and only 6 of the 23 clusters differ by more than 0.20 units (absolute value) in RMSE (with smaller RMSE obtained by SFCscore^{RF} for thrombin, casein kinase II, and α-mannosidase 2, and larger RMSE observed for carbonic anhydrase, urokinase, and tRNA-guanine transglycosylase).

The reasons for the large spread of performance across the different clusters are manifold, and a detailed analysis for every target class would be beyond the scope of this paper. However, some more general points should be outlined. In their reply to Kramer and Gedeck, Ballester and Mitchell question the practical value of leave-cluster-out cross-validation and provide important considerations in this context.[36] Indeed, the leave-cluster-out cross-validation may provide a somewhat too pessimistic estimate of the actual performance. In real-world applications, the type of complexes that the approach is actually meant to work on would never be eliminated from a training set. Rather, the training set composition should reflect as closely as possible the actual objects on which the empirical approach will be applied. For example, from the 187 complexes of the HIV protease data set, many would normally enter the training set, thus ensuring that the descriptor space is adequately covered to predict the affinity of other HIV protease complexes. Eliminating all 187 complexes, however, leads to an imbalance between test and training set, as HIV protease inhibitors are on average much larger than the ligands of the other targets. The boxplot in Figure 5 shows the distribution of the TotBurSurf (total buried surface) descriptor values of the HIV protease test set (cluster A) and the training set (clusters B through Z). Obviously, there is a clear difference with respect to this important descriptor, which also explains why the HIV protease complexes generally receive too high scores in the leave-cluster-out cross-validation. Interestingly, analyzing the OOB predictions (as reported in Table 1) of the actual RF-model used for SFCscore^{RF} by protein class shows much better results for the HIV protease class (N = 97, R = 0.60, RMSE = 1.26). This illustrates that the leave-cluster-out results should not directly be interpreted as performance measures of SFCscore^{RF} on particular protein classes.

Another critical point is the limited size of many clusters and the small range of experimental affinity values therein, making a correct prediction of the ranking rather challenging if not impossible. Looking at the RMSE in relation to the correlation coefficient shows, for example, that for cluster P (glutamate receptor 2), the RMSE is only 1.00 even though the calculated values do not show any correlation with the experimental affinities. In fact, 9 of 13 complexes in cluster P have pK_i values within a range of 1.5 units. Most of the calculated values do not deviate much from the experimental ones on an absolute scale, but the relative ranking is very different due to varying residuals. Nevertheless, there are also cases where the prediction works well despite a relatively small affinity range. For example, in cluster L (casein kinase II), the experimental pK_i values span a range of slightly more than 2 units, yet the predictions show a good correlation (R = 0.75) and small residuals (RMSE = 0.64).

In summary, even though good and practically useful results can already be obtained for a variety of different target classes with scoring functions derived from large generic data sets of approximately 1000 complexes, such functions are not yet able to provide robust predictions across all different targets. Apparently, there are still not all interaction features captured that might be responsible for particular structure−activity relationships of certain target classes. As long as this situation cannot be overcome, careful validation of a given scoring function for application on a particular target class or ligand series is of utmost importance.

Validation Data Sets for CSAR 2012 Exercise. Prior to the application of SFCscore to data sets of the CSAR 2012 challenge, the performance of the individual functions was evaluated by means of validation data sets generated for the corresponding targets as described in Materials and Methods. The required docking poses were generated with Glide after carrying out redocking and crossdocking experiments with a set of high quality crystal structures (data not shown). Visual

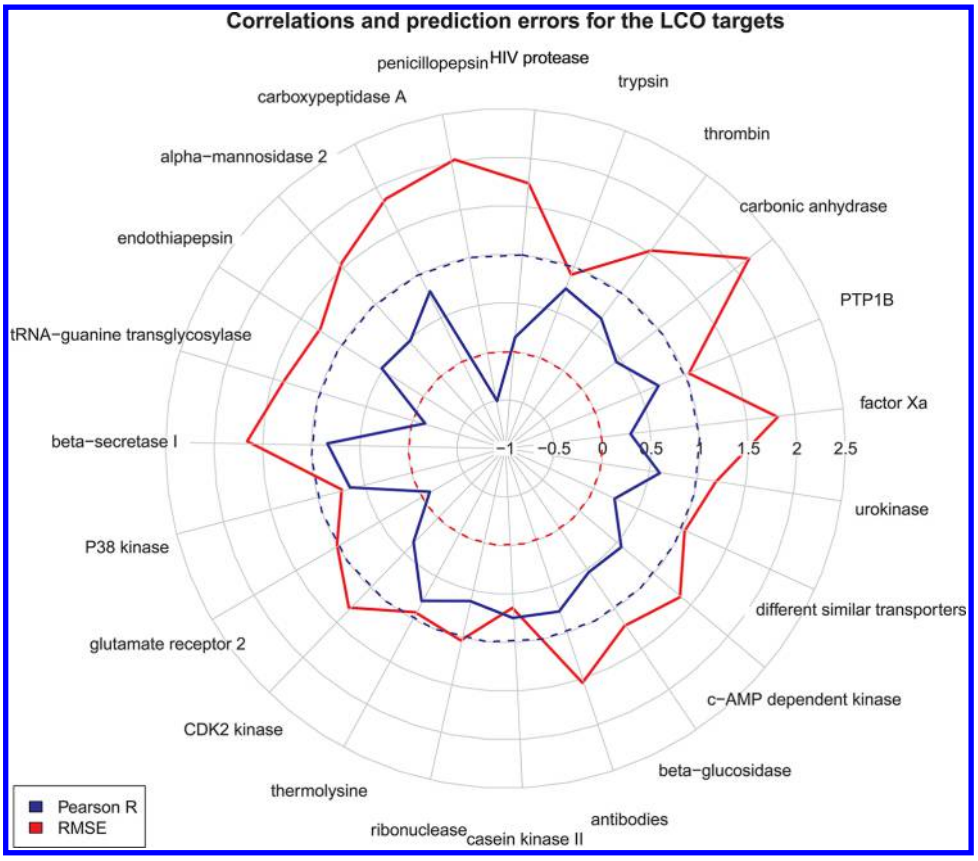


Figure 4. Graphical representation of the leave-cluster-out cross-validation results for the SFCscore random forest approach. The closer the blue line (R) to 1 and the red line (RMSE) to 0 (dotted lines), the better the corresponding prediction.

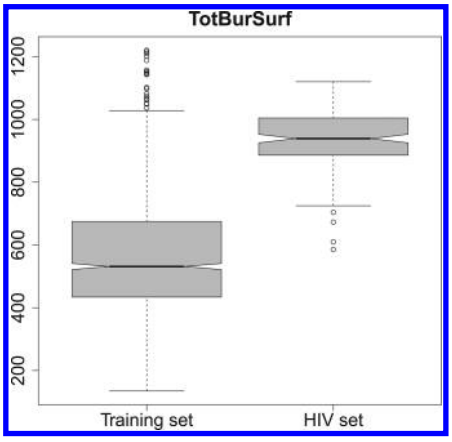


Figure 5. Boxplot illustrating the distribution of the descriptor values for TotBurSurf (total buried surface of the ligand) in the training set without HIV protease complexes (left) and in the HIV protease test set (right).

inspection of the docking poses generated for the validation data sets showed in the case of CHK1 and ERK2 that neither GlideScore nor DSX rescoring provided reasonably consistent binding modes within the different ligand series. Under the assumption of an approximately conserved binding mode for congeneric ligands of an SAR series, a visual inspection and selection of the docking poses was carried out, guided by the information from the available crystal structures.

The criteria for pose selection in case of CHK1 ligands were hydrogen bond interactions with the backbone nitrogen of Cys87 as well as the backbone oxygen of Cys87 or Glu85. Similarly, docking poses of ERK ligands had to show at least two backbone hydrogen bonds: one with Met106 nitrogen and a second with the oxygen of Met106 or Asp104. In cases where multiple binding modes satisfied the criteria, the pose with the better GlideScore was chosen. For some compounds, no docking poses consistent with the common binding mode of related ligands and satisfying the hydrogen bond criteria could be obtained. Because of their highly uncertain binding mode, these compounds were excluded from the subsequent scoring. For the LpxC data set, no manual pose selection was necessary, as the poses suggested by Glide and DSX were consistent within the ligand series and also in good agreement with the crystal structure PDB 3p3e showing the binding mode of a ligand similar to the investigated series.

The selected poses were scored with all SFCscore functions, including SFCscore^{RF} (sfc_rf). Table 5 shows the obtained correlations between the experimentally determined affinities and the SFCscore predictions for the three validation data sets. For CHK1, good correlations were obtained with most of the functions. SFCscore^{RF} and the MLR functions sfc_229m and sfc_290m are in the top ranks, with virtually identical performance (R between 0.892 and 0.904). Larger differences among the functions and worse overall performance were observed for the ERK2 data set, where SFCscore^{RF} and sfc_frag showed the best results (0.726 and 0.742, respectively). Finally, with the LpxC data set, most functions achieve values only between 0.5 and 0.6, with sfc_229m and sfc_290m as exceptions (0.642 and 0.634, respecively). On the basis of these results, SFCscore^{RF} (for CHK1 and ERK2), sfc_229m (for CHK1 and LpxC), and sfc_frag (for ERK2) were selected as functions for submitting predictions to the CSAR exercise.

dx.doi.org/10.1021/ci400120b | J. Chem. Inf. Model. 2013, 53, 1923−1933

Table 5. Correlation between Experimental and Predicted Binding Affinities in Terms of Pearson's Correlation Coefficient *R* for the SFCscore Functions Applied on the Three Validation Data Sets (Consisting of 13 CHK1 Complexes, 41 ERK2 Complexes, and 42 LpxC Complexes)[a]

	CHK1	ERK2	LpxC
sfc_rf	**0.892**	**0.726**	0.523
sfc_229m	**0.903**	0.500	**0.640**
sfc_229p	0.879	0.450	0.571
sfc_290m	0.904	0.476	0.634
sfc_290p	0.870	0.385	0.605
sfc_855	0.789	0.612	0.520
sfc_frag	0.820	**0.742**	0.549
sfc_met	0.822	0.480	0.570
sfc_ser	0.840	0.471	0.560

[a]The functions selected for submission to the scoring exercise are marked in **bold**.

Nevertheless, the results of all SFCscore functions are presented and analyzed below.

CSAR 2012 Exercise Data Sets. The results for the CSAR 2012 exercise data sets in terms of correlation coefficients and RMSE are summarized in Table 6. Unexpectedly, the overall trends in the performance of the functions differ quite significantly from those observed with the validation data sets. In the following, the analysis and discussion is carried out individually for each target.

CHK1. For the CHK1 data set, none of the SFCscore functions obtained satisfying results. With an *R* value of 0.38, sfc_229m and sfc_290m show the best correlation coefficient, whereas no correlation at all is observed for SFCscore^{RF}. However, on the basis of the summarized results of all submissions released later on by the CSAR organizers, the CHK1 data set in general was very demanding, as the majority of the functions earned poor correlations.

To understand and explain these results, the correctness of the binding modes obtained by docking might be the most important factor, at least in our case. Comparison of the docking poses with the crystal structures of 12 compounds disclosed by the CSAR organizers shows that in only 7 cases the docked binding modes were correct. Especially in ligand series 3 (urea derivatives), multiple cases are observed where the selection criterion (two backbone hydrogen bonds) is satisfied, yet one hydrogen bond is formed by the wrong nitrogen atom of the ligand, leading to a 180° flip of the structure (Figure 6). The correlation coefficients split by ligand series shown in Table 7 clearly demonstrate that none of the functions achieves meaningful predictions for this series 3. On the other hand, ligand series 1 is predicted much better by most

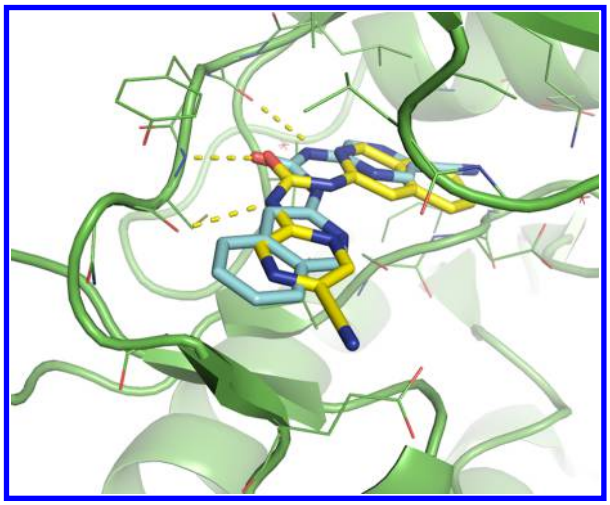


Figure 6. Comparison of docked pose (yellow) and crystallized pose (cyan) of the ligand chk_35. The RMSD is 8 Å because the hydrogen bond with the protein backbone is formed by different nitrogens of the urea substructure. The figure was prepared with PyMOL.[37]

Table 7. Pearson's Correlation Coefficients *R* for the CHK1 Data Set Split by Ligand Series[a]

function	series 1	series 2	series 3
sfc_rf	0.55	0.0055	−0.54
sfc_229m	0.67	0.50	−0.53
sfc_229p	0.59	0.46	−0.35
sfc_290m	0.67	0.50	−0.55
sfc_290p	0.49	0.32	−0.63
sfc_855	0.18	0.47	−0.40
sfc_frag	0.45	0.50	−0.38
sfc_met	0.45	0.29	−0.70
sfc_ser	0.56	0.35	−0.67

[a]Ligand series 1 corresponds to the imidazole derivatives (14 compounds), series 2 to the lactame derivatives (16 compounds), and series 3 to the urea derivatives (17 compounds).

Table 6. Overview of Results Obtained with SFCscore Functions for the CSAR 2012 Exercise Data Sets CHK1, ERK2, and LpxC[a]

	CHK 1			ERK 2			LpxC		
	R	ρ	RMSE	*R*	ρ	RMSE	*R*	ρ	RMSE
sfc_rf	−0.02	−0.02	1.28	0.49	0.45	1.13	0.90	0.76	0.55
sfc_229m	0.38	0.31	1.25	0.45	0.47	1.11	0.56	0.02	1.06
sfc_229p	0.30	0.11	1.59	0.35	0.37	1.96	0.70	0.48	0.73
sfc_290m	0.38	0.30	1.25	0.44	0.45	1.13	0.52	0.02	1.05
sfc_290p	0.23	0.15	1.31	0.44	0.46	1.16	0.58	0.10	0.91
sfc_855	0.19	0.17	1.24	0.47	0.44	1.30	0.55	0.43	0.81
sfc_frag	0.24	0.20	1.27	0.54	0.53	1.50	0.80	0.57	1.10
sfc_met	0.18	0.12	1.36	0.44	0.48	1.29	0.63	0.43	0.89
sfc_ser	0.22	0.17	1.29	0.46	0.48	1.13	0.51	0.35	0.96

[a]*R* is Pearson's correlation coefficient. ρ is Spearman's correlation coefficient. RMSE is root mean square error.

dx.doi.org/10.1021/ci400120b | *J. Chem. Inf. Model.* 2013, 53, 1923−1933

of the functions. In fact, this series also showed the most consistent and reliable docking results.

In this context, it is worth noting that SFCscore generally predicts more accurate scores for crystal structures, not only in the obvious cases where the deviation between docked and crystallized pose is high but also in some cases with relatively small structural differences. The most remarkable example is chk1_6, where the RMSD between crystal structure and docked pose is only 0.47 Å, yet the score of SFCscore^{RF} differs by 0.55 pK_i units (docked pose, 6.95; crystal pose, 7.50; experimental pK_i, 8.80). Certainly, the quality of the ligand poses is an essential prerequisite for more successful affinity prediction.

ERK 2. The results obtained for the ERK 2 data set are clearly better than the CHK1 results, but are still not satisfying, given correlation coefficients between 0.35 and 0.54. Nearly all functions, including SFCscore^{RF}, perform very similarly on this data set. Again, the quality of the docking poses appears to have large influence. Of the 12 complexes with disclosed crystal structure, only 5 had docking poses with an RMSD < 2.5 Å. In some cases, the backbone interaction used as pose-selection criterion is formed by different functional groups. In contrast to the CHK1 data set, the problem is not limited to one ligand series.

Expectedly, the performance of SFCscore improves when the crystal structures are rescored instead of the docked poses. Figure 7 shows the Pearson correlation coefficients for the prediction of the 12 crystal structures in comparison to the predictions of the docked poses. With the exception of two functions (sfc_met and sfc_290p), all functions perform better for the crystal structures. For some of the functions, the improvement is significant, e.g., SFCscore^{RF} shows a correlation of around 0.5 for the docked poses and improves to nearly 0.8 for the crystal structures.

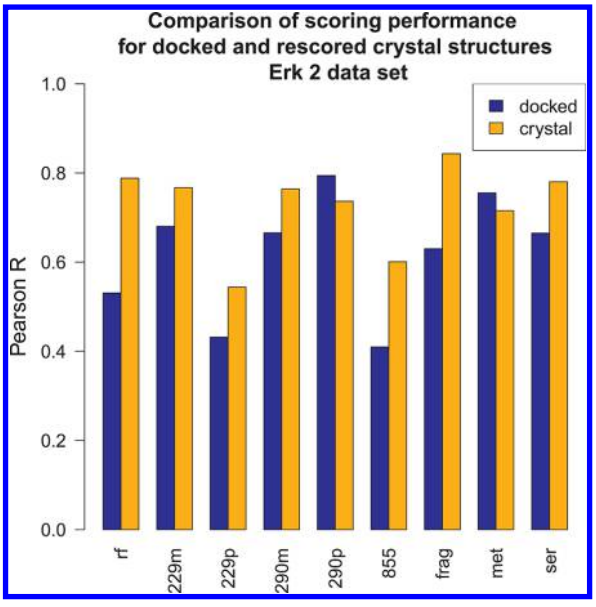


Figure 7. Pearson's correlation coefficients *R* for the predition of the 12 crystal poses in comparison to the docked poses of the respective compounds.

LpxC. The results obtained for the LpxC data set are the best of the three tested data sets. Pearson's correlation coefficients range from 0.51 to 0.90, with SFCscore^{RF} as the top-ranked function. The prediction errors are in most cases <1 p*K* unit, which is fully acceptable for many purposes. The better prediction quality is most likely due to the more reliable docking poses, as there is essentially only one reasonable possibility how the ligands might bind to the pocket. Nevertheless, the SFCscore functions show relatively large differences in their performance, with some of them displaying very low Spearman's correlation coefficients, such as sfc_229m (it should be noted, though, that the data set consists of only eight active compounds, five of which displaying very similar affinities; this makes the Spearman rank correlation very susceptible to outliers).

To identify possible reasons for this varying performance, the training data of the three functions SFCscore^{RF}, sfc_229p (both performing comparatively well), and sfc_229m (showing surprisingly poor results for the LpxC data set) were subjected to three individual PCA analyses in terms of the descriptors used to derive the corresponding function. In each case, the first two principal components explain a significant amount of variance in the data sets, as indicated by the percentages on the axis labels in Figure 8. By plotting these two components and calculating the distribution density of the objects (complexes) in this PCA space, a two-dimensional heat map was generated, illustrating the descriptor space used to derive the function. Projecting the objects of the LpxC test set onto this heat map provides an impression whether the test set objects lie within the area spanned and covered by the training set. The plots in Figure 8 show that this is very well the case for SFCscore^{RF} and—to a slightly lower degree—for sfc_229p. In contrast, for sfc_229m, the LpxC complexes do not map to well-occupied areas of the training set. This provides an explanation of why the sfc_229m function was found to be less well suited for estimating the affinities of the LpxC compounds.

CONCLUSIONS

SFCscore^{RF} is a new scoring function based on SFCscore descriptors and derived with the random forest machine learning approach from a large training set of PDBbind protein−ligand complexes. Its performance on two common benchmarks (the 195 PBDbind test set complexes and the 332 CSAR−NRC HiQ complexes) ranks it among the best functions hitherto reported for these validation sets. Leave-cluster-out cross-validation indicates target-dependent performance, with some target classes being described very well, yet others without any degree of correlation. This cross-validation exercise, however, provides an estimate of the performance of the generic SFCscore^{RF} approach on a truly new target protein not represented by the clusters of the training set. For the more common real-world scenario of application on closely related protein targets, a more stable performance can be expected, as illustrated also by the application of SFCscore^{RF} on the LpxC test set of the CSAR 2012 Exercise.

As the development of SFCscore^{RF} has been conducted along the lines of the seminal RF-Score study,[12] a direct comparison of the two functions is possible. Because the random forest approach has been applied in complete analogy and the training sets are nearly identical (with the exception that about 10% of the complexes had to be removed from the SFCscore^{RF} training set to avoid overlap with the CSAR test set), any difference in performance is primarily related to the applied descriptors. In contrast to the completely nonparametric approach of RF-Score based on atom-pair contact counts, the SFCscore descriptors imply some modeling assumptions regarding

dx.doi.org/10.1021/ci400120b | J. Chem. Inf. Model. 2013, 53, 1923−1933

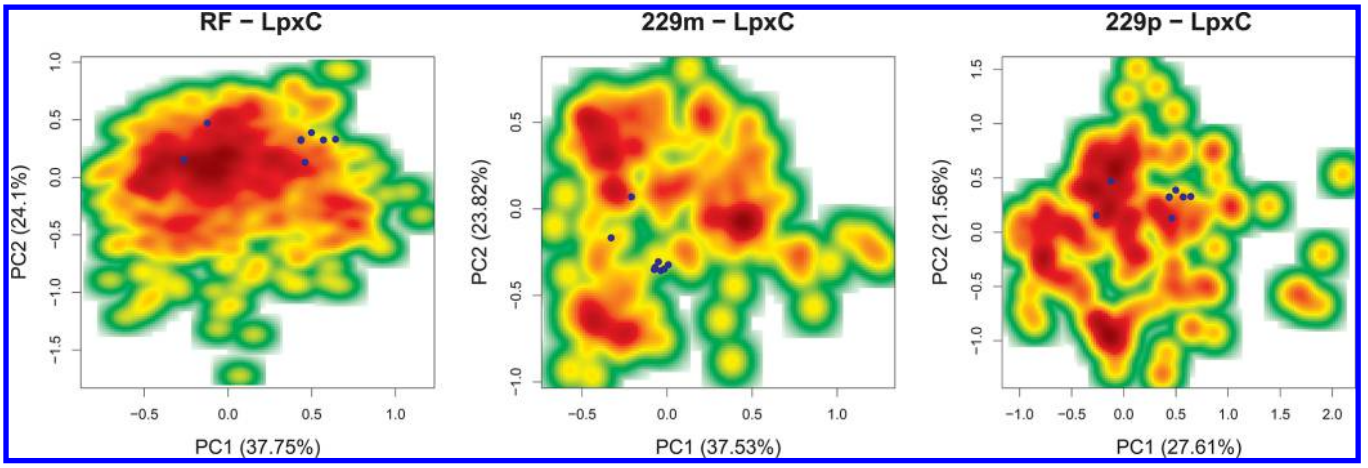


Figure 8. Density of the training set objects in the descriptor space spanned by the first two principal components of the descriptors used to derive the functions SFCscore^{RF} (left), sfc_229m (center), and sfc_229p (right). The variance explained by each principal component is noted in brackets. The density is displayed as a heat map, with red indicating high density and green low density. The position of the LpxC test set objects projected onto these PCA plots is shown as blue dots.

molecular recognition features (as for example in the hydrogen bond descriptors). This does not appear to deteriorate the resulting function, as virtually the same performance is obtained with SFCscore^{RF} and RF-Score for the same benchmark. An advantage of the SFCscore descriptors might be seen in the possibility for some degree of interpretation, as illustrated in the discussion of the importance plot. On the basis of such an analysis of the role of single descriptors, further improvements can be envisaged, either by fine tuning current descriptors or by developing complementary ones.

Application of SFCscore^{RF} and the classical SFCscore functions on the CSAR 2012 exercise data sets CHK1, ERK2, and LpxC resulted in predictions of strongly varying quality. In the self-compiled validation sets for the three targets, SFCscore^{RF} showed the best overall performance. In the exercise data sets, however, SFCscore^{RF} gave good results only for LpxC. For CHK1 and ERK2, the results were below the expectations, although it should be mentioned that none of the functions reached a satisfying degree of correlation and prediction accuracy. A major problem in this context was the quality of the docking poses, which was overall insufficient for CHK1 and ERK2. Apart from the misplacement of individual ligands, which makes the subsequent scoring futile, generally tolerated deviations of docking poses from the native crystal pose affect the scoring results often negatively. In particular, SFCscore^{RF}—the scoring function trained on the largest data set of high-quality crystal structures—was found to sensitively depend on the quality of the ligand poses.

Besides pose quality, the applicability domain is an essential issue for empirical scoring functions. In general, reliable predictions can only be expected within adequately covered areas of the chemical space in terms of the applied descriptors of the training set. As shown for the LpxC data set, mapping of this chemical space via PCA analysis and projecting the test set objects onto this map can help to estimate whether a specific function can be expected to perform well on a new data set or might be less suited for reliable predictions. Although this approach was used retrospectively in this study, it can be suggested for prospective application and selection of the most suitable scoring function variant. Caution is warranted, though, when using docking poses in this context. Erroneous ligand poses (as seen in the CHK1 and ERK2 prediction sets) can also mislead such an analysis, highlighting once more the need for precise binding modes as prerequisite for successful scoring.

■ ASSOCIATED CONTENT

Supporting Information

Table with the PDB codes of all training set complexes (Table S1) and the SFCscore descriptors for all data sets that were used in this study (Table S2). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: sotriffer@uni-wuerzburg.de.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Financial support by the DFG (KFO216) is gratefully acknowledged.

■ REFERENCES

(1) Sotriffer, C. Scoring Functions for Protein-Ligand Interactions. In *Protein−Ligand Interactions*; Methods and Principles in Medicinal Chemistry Series; Gohlke, H., Ed.; Wiley-VCH: Weinheim, Germany, 2012; Chapter 12, pp 237−263.

(2) Warren, G.; Andrews, C.; Capelli, A.; Clarke, B.; LaLonde, J.; Lambert, M.; Lindvall, M.; Nevins, N.; Semus, S.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.

(3) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. *Br. J. Pharmacol.* **2008**, *153* (Suppl), S7−S26.

(4) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.* **2009**, *49*, 1079−1093.

(5) 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.

(6) Smith, R. D.; Dunbar, J. B.; Ung, P. M.-U.; 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.
(7) Sotriffer, C.; Matter, H. The Challenge of Affinity Prediction: Scoring functions for Structure-Based Virtual Screening. In *Virtual Screening*; Methods and Principles in Medicinal Chemistry Series; Sotriffer, C., Ed.;Wiley-VCH: Weinheim, Germany, 2011; Chapter 7, pp 177−221.
(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) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *45*, 549−561.
(10) Amini, A.; Shrimpton, P.; Muggleton, S.; Sternberg, M. A general approach for developing system-specific functions to score protein−ligand docked complexes using support vector inductive logic programming. *Proteins* **2007**, *69*, 823−831.
(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) Ballester, P. J.; Mitchell, J. B. O. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26*, 1169−1175.
(13) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind database: Methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.
(14) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.
(15) Dunbar, J. B.; Smith, R. D.; Yang, C.-Y.; Ung, P. M.-U.; Lexa, K. W.; Khazanov, N. A.; Stuckey, J. A.; Wang, S.; Carlson, H. A. CSAR Benchmark Exercise of 2010: Selection of the protein−ligand complexes. *J. Chem. Inf. Model.* **2011**, *51*, 2036−2046.
(16) Tong, Y.; Claiborne, A.; Stewart, K. D.; Park, C.; Kovar, P.; Chen, Z.; Credo, R. B.; Gu, W.- Z.; Gwaltney, S. L., II; Judge, R. A.; Zhang, H.; Rosenberg, S. H.; Sham, H. L.; Sowin, T. J.; Lin, N. H. Discovery of 1,4-dihydroindeno[1,2-c]pyrazoles as a novel class of potent and selective checkpoint kinase 1 inhibitors. *Bioorg. Med. Chem.* **2007**, *15*, 2759−2767.
(17) Sebolt-Leopold, J. S.; Herrera, R. Targeting the mitogen-activated protein kinase cascade to treat cancer. *Nat. Rev. Cancer* **2004**, *4*, 937−947.
(18) Aronov, A. M.; Tang, Q.; Martinez-Botella, G.; Bemis, G. W.; Cao, J.; Chen, G.; Ewing, N. P.; Ford, P. J.; Germann, U. A.; Green, J.; Hale, M. R.; Jacobs, M.; Janetka, J. W.; Maltais, F.; Markland, W.; Namchuk, M. N.; Nanthakumar, S.; Poondru, S.; Straub, J.; ter Haar, E.; Xie, X. Structure-guided design of potent and selective pyrimidylpyrrole inhibitors of extracellular signal-regulated kinase (ERK) using conformational control. *J. Med. Chem.* **2009**, *52*, 6362−6368.
(19) Barb, A. W.; Zhou, P. Mechanism and inhibition of LpxC: An essential zinc-dependent deacetylase of bacterial lipid A synthesis. *Curr. Pharm. Biotechnol.* **2008**, *9*, 9−15.
(20) Lee, C.-J.; Liang, X.; Chen, X.; Zeng, D.; Joo, S. H.; Chung, H. S.; Barb, A. W.; Swanson, S. M.; Nicholas, R. A.; Li, Y.; Toone, E. J.; Raetz, C. R.; Zhou, P. Species-specific and inhibitor-dependent conformations of LpxC: Implications for antibiotic design. *Chem. Biol.* **2011**, *18*, 38−47.
(21) 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.
(22) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2011**, *40*, D1100−D1107.
(23) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.
(24) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.
(25) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.
(26) 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.
(27) Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18−22.
(28) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.
(29) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.
(30) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein−ligand complexes. *J. Med. Chem.* **2006**, *49*, 6177−6196.
(31) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: A software program for pK(a) prediction and protonation state generation for drug-like molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 681−691.
(32) Greenwood, J. R.; Calkins, D.; Sullivan, A. P.; Shelley, J. C. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 591−604.
(33) *Schrödinger Suite 2011 Protein Preparation Wizard*; *Epik*, version 2.2, *Impact*, version 5.7, *Prime*, version 2.3; Schrödinger, LLC: New York, 2011.
(34) Neudert, G.; Klebe, G. DSX: A knowledge-based scoring function for the assessment of protein−ligand complexes. *J. Chem. Inf. Model.* **2011**, *51*, 2731−2745.
(35) Kramer, C.; Gedeck, P. Leave-cluster-out cross-validation is appropriate for scoring functions derived from diverse protein data sets. *J. Chem. Inf. Model.* **2010**, *50*, 1961−1969.
(36) Ballester, P. J.; Mitchell, J. B. O. 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.* **2011**, *51*, 1739−1741.
(37) *The PyMOL Molecular Graphics System*, version 1.5.0.5; Schrödinger, LLC: New York, 2012.