Leave-Cluster-Out Cross-Validation Is Appropriate for Scoring Functions Derived from Diverse Protein Data Sets

Christian Kramer* and Peter Gedeck

Novartis Institutes for BioMedical Research, Novartis Pharma AG, Forum 1, Novartis Campus, CH-4056 Basel, Switzerland

Received July 9, 2010

With the emergence of large collections of protein−ligand complexes complemented by binding data, as found in PDBbind or BindingMOAD, new opportunities for parametrizing and evaluating scoring functions have arisen. With huge data collections available, it becomes feasible to fit scoring functions in a QSAR style, i.e., by defining protein−ligand interaction descriptors and analyzing them with modern machine-learning methods. As in each data modeling ansatz, care has to be taken to validate the model carefully. Here, we show that there are large differences measured in R (0.77 vs 0.46) or R^2 (0.59 vs 0.21) for a relatively simple scoring function depending on whether it is validated against the PDBbind core set or validated in a leave-cluster-out cross-validation. If proteins from the same family are present in both the training and validation set, the estimated prediction quality from standard validation techniques looks too optimistic.

INTRODUCTION

Structure-based drug design and the application of docking in virtual screening have become standard tools in the drug discovery process over the past decade.[1−4] The application of scoring functions is crucial for evaluating different protein−ligand interaction hypotheses by differentiating between good and bad hypotheses. Additionally, it would be desirable to have scoring functions available that are able to correctly predict the binding free energy of a given complex. No scoring function published so far has achieved this goal to a degree which is competitive with experimentally determined binding constants.[5] Although the docking technology is rather old (the first docking program DOCK by Kuntz and co-workers[6] was published in the 1980s), the correlations of predicted and measured activities validated on a large diverse data set are rather small.[7]

Traditionally, scoring functions can be divided into three different approaches:

(I) Force-field-based scoring functions that purely rely on molecular dynamics force fields
(II) Empirical scoring functions containing additional terms for entropic contributions
(III) Knowledge-based scoring functions which use atom- or fragment-pair potentials based on statistical analysis of crystal contacts

In the field of knowledge-based scoring functions, a very interesting publication of a scoring function called RF-score was recently published by Ballester and Mitchell.[8] This scoring function is based on ligand_atom-protein_atom pair counts, which are combined into a scoring function using a Random Forest in classical QSAR style. This scoring function and the way it is derived are interesting for the following three reasons:

(I) It outperforms all other standard scoring functions tested on the PDBbind2007 core set.
(II) It is created in classical QSAR style, using protein−ligand complex descriptors and a nonlinear learning algorithm.
(III) It has been trained on the PDBbind2007 refined set, which contains more than 1000 complexes and thus is probably the largest data set that has ever been used for generating scoring functions.

In particular, the first reason makes this function attractive. In comparison to all of the other scoring functions, RF-score is based on a much simpler functional. The second item illustrates that the full breadth of QSAR methodology, developed over the past 40 years, is applicable to the scoring problem. The third point, finally, is important, because a QSAR model can be retrained frequently, and with an, over time, increasing data set, it gets better or at least more reliable. If the number of protein−ligand complexes published increase as strongly as the number of publicly available protein crystals structures, there will be an ever increasing wealth of data to develop and improve models. The QSAR methodology allows for a profit from such an increase in a straightforward manner. The application of QSAR algorithms to scoring allows the development of a whole new class of scoring functions, if we can develop meaningful and general protein−ligand interaction descriptors.

Each QSAR model requires thorough validation by determining the performance of the model on an external test set. This external test set must contain data points that have not been used during model development. There are various ways of choosing such a test set. The most obvious one is by random selection. In some cases, however, systematic approaches may be more suitable, but care must be taken that this does not introduce an optimistic bias, which means that the performance of the model is overestimated. In the case of the RF-score publication, the PDBbind core set[7] was used as an external validation. With this procedure,

* Contact email: Christian.Kramer@novartis.com.

10.1021/ci100264e
Published on Web 10/12/2010

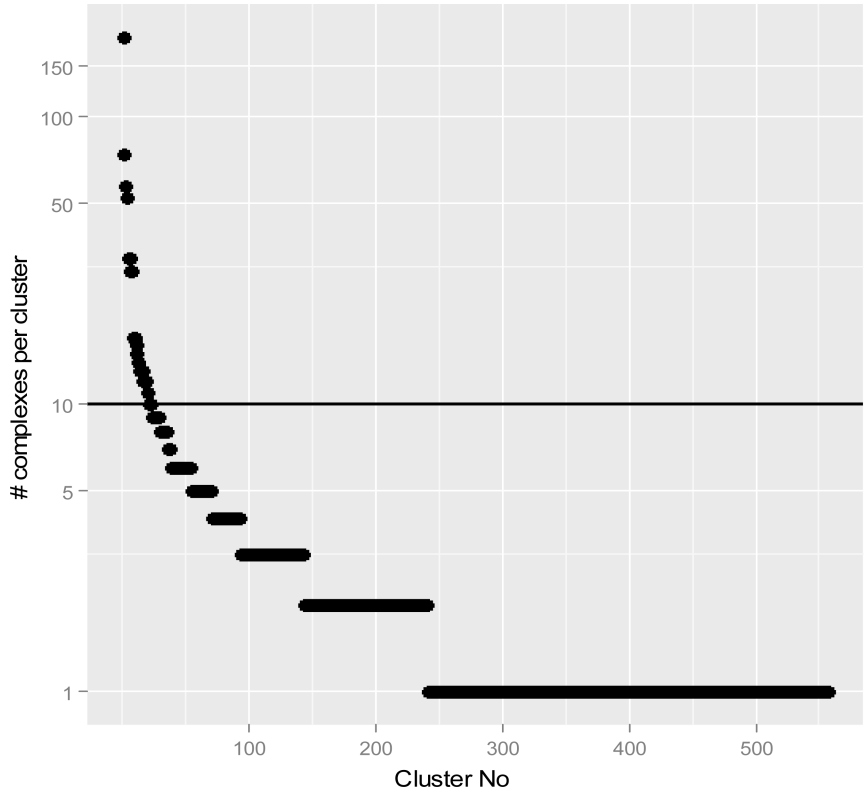

Figure 1. Distribution of cluster size in the PDBbind09 database. The horizontal line shows the 10 complexes per cluster limit.

Ballester and Mitchell followed the setup previously used by Wang et al.[7] in a study where they compared the performance of different scoring functions in the PDB core set. Briefly, the PDBbind core set is assembled from all protein families (subsequently called clusters) that contain at least four members within the PDBbind07 refined set. From every cluster, the complex with the maximum binding energy, the minimum binding energy, and the complex with the binding energy closest to the cluster average are selected. This procedure yields a diverse set of protein ligand complexes spanning a large magnitude of possible binding data. This data subset is interesting for the evaluation of scoring functions that have been built on the basis of data outside of the PDBbind database. However, as will be demonstrated in this communication, care must be taken when interpreting the performance quality measures, since a significant part of the variance can be regressed using knowledge about the cluster membership only. For the assessment of scoring functions, it is often more interesting how the performance within certain clusters varies compared to how much it varies across clusters. In the following, we analyze the PDBbind core and refined sets, propose a clustering scheme for the PDBbind database,[9,10] evaluate the performance of RF-score and a variant with a different cutoff for the atom−atom pair descriptors using this clustering scheme, and highlight the advantages and necessity of leave-cluster-out cross-validation for PDBbind-based scoring functions. The results of this study also apply to any work with similar collections of protein−ligand complexes and their activities, such as BindingMOAD,[11] another database used for the development of scoring functions.

METHODS

PDBbind, PDBbind Refined Set, and PDBbind Core Set. In order to understand the necessity of leave-cluster-out cross-validation, the composition of the training data set has to be understood. The PDBbind database was first published in 2004 as a collection of all high-quality protein−ligand crystal structure complexes published in the PDB.[9] These complexes have been complemented with binding free energy data determined from K_i's and K_d's. From all complexes, a refined set is generated which consists of all complexes that have a resolution of less than 2.5 Å and contain only noncovalent complexes with small molecules or small peptides. Being updated every year, the refined set reached a size of 1741 high-quality complexes in 2009 which are curated and prepared for automated processing. The protein sequences were further clustered according to a 90% BLAST sequence similarity into 237 protein clusters, as was done by Wang et al. for assembling the PDBbind core set. The largest cluster of the PDBbind 2009 version comprises 188 complexes of HIV-protease. The second largest consists of 74 complexes of trypsin, and the third largest cluster contains 54 complexes of carbonic anhydrase. The PDBbind data set also has 321 singletons that are not assigned to any cluster. The distribution of the cluster size is highly skewed, as shown in Figure 1.

A weekly updated clustering of the PDB according to BLAST sequence similarities is available from www.pdb.com. Using the cluster assignment downloaded on April 21, 2010, the clusters in PDBbind2009 were assigned. The assignment for each cluster can be found in the Supporting Information.

Since 2007, a core set is assembled each year from the clustering which consists of three complexes per cluster

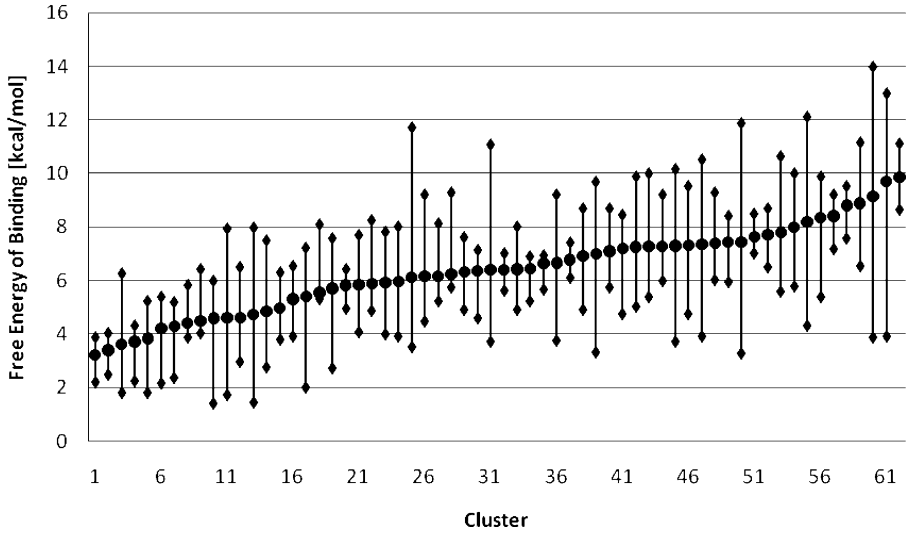

Figure 2. Distribution of PDBbind07 core set activity data, sorted according to the middle value.

created for all clusters that contain at least four complexes. It contains the complex with the highest binding energy, the lowest binding energy, and the binding energy closest to the mean. The core set of the PDBbind2007 database has been used for validating a large set of scoring functions and the RF-score scoring function. In the version that we downloaded, there are 210 complexes. A total of 195 of these complexes have been used for a large validation study of scoring functions. From the 195 complexes mentioned in the original publication,[7] four (1fzj, 1g7f, 2aov, and 2azr) are not in the PDBbind2007 database that we downloaded. For the within cluster analysis of the PDBbind2007 database, we consequently discarded the whole cluster of these complexes, leaving 186 complexes in 62 clusters. The list of PDB IDs can be found in the Supporting Information. The distribution of activity values in the core set is illustrated in Figure 2.

The distribution of free energies within the clusters is narrower than the distribution of free energies across the whole data set. In 10 clusters, the difference between the most active and the least active compound is less than 2 kcal/mol, while the largest difference is around 10 kcal/mol. Some of the least active compounds of clusters with a very high average activity have a higher free energy of binding than the most active compounds from clusters with a very low average activity. There is a significant variance of activities between different clusters.

The PDBbind database is a very valuable source of protein−ligand complex data, and it can be used to both parametrize scoring functions and assess their predictive power on a broad structural data basis. As the data set contains several clusters of homogeneous compound series, standard cross-validation approaches cannot be used, as will be shown below. In contrast, we suggest a leave-cluster-out cross-validation scheme: compared to standard cross-validation, the subsets left out for validation are not selected randomly but according to the cluster to which they belong. Each complete cluster is iteratively left out of the training set and used to assess the predictive quality of the generated model. The overall predictivity can be obtained from either a weighted average over all clusters or from inspecting each cluster individually. The leave-cluster-out cross-validation scheme is depicted in Figure 3.

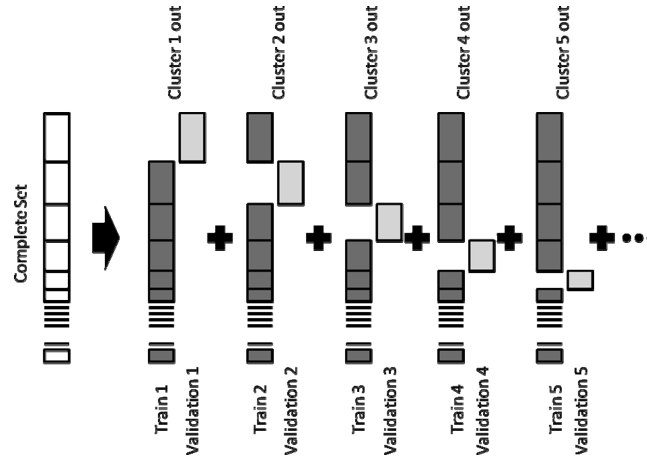

Figure 3. Out-of-cluster cross-validation scheme.

Descriptors and Learning Algorithm. For this study, the atom−atom pair count descriptors as published by Ballester and Mitchell were used. Briefly, the pairs of Ligand_atom−Protein_atom in a radius of 12 Å for the elements C, N, O, F, P, S, Cl, Br, and I of the ligand and C, N, O, and S of the protein are counted. This gives 36 descriptors which are used to train a random forest model for the prediction of free energies. The code provided in the Supporting Information of the RF-score publication was used to calculate the descriptors. The random forest models[12] were generated in the *R* version 2.8.1[13] with the randomForest package.[14] Standard settings (1000 trees, minimum node size = 5) were used for training. Preliminary tests showed that the prediction performance varies only very slightly upon changing the number of parameters evaluated for each split, so the default value of 5 was used for all models.

Measures of Performance. We use the root mean squared error (RMSE) and the Pearson correlation coefficient *R* to assess the quality of the models. Additionally, we give the square of the Pearson correlation coefficient R^2, because it estimates the proportion of the variance of the data set explained by the model. The RMSE measures the accuracy of the prediction. This means how well the experimental value is reproduced by the model. *R* and R^2 assess how well models predict the relative order within a data set. A model that has a high RMSE (bad) and a high R^2 (good) can still

Table 1. Results for the RF-Score Scoring Functions Evaluated with Different Cutoffs for the Descriptors[a]

binding site cutoff	R	R^2	RMSE
4 Å	0.7	0.49	1.81
6 Å	0.76	0.58	1.63
8 Å	0.76	0.58	1.61
12 Å	**0.77**	**0.59**	**1.58**
16 Å	0.77	0.59	1.60

[a] The 12 Å cutoff used in the original RF-score publication is highlighted.

be useful to prioritize candidates for screening in typical scoring applications.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{i,\mathrm{pred}} - y_{i,\mathrm{meas}})^2}$$

$$R = \frac{\sum_{i=1}^{N}(y_{i,\mathrm{pred}} - \bar{y}_{\mathrm{pred}})(y_{i,\mathrm{meas}} - \bar{y}_{\mathrm{meas}})}{\sqrt{\sum_{i=1}^{N}(y_{i,\mathrm{pred}} - \bar{y}_{\mathrm{pred}})^2(y_{i,\mathrm{meas}} - \bar{y}_{\mathrm{meas}})^2}}; \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$R^2 = R \times R$$

Here, N is the number of instances, $y_{i,\mathrm{pred}}$ is the predicted binding energy, and $y_{i,\mathrm{meas}}$ is the measured binding energy. The average values are calculated from the complete subset, i.e., either the training or validation set.

RESULTS

PDBbind Core Set Validation. To validate our setup, we initially reproduced the original RF-score publication with nearly the same results as published: $\mathrm{RMSE}_{\mathrm{validation\_set}} = 1.58$, $\mathrm{RMSE}_{\mathrm{out\text{-}of\text{-}bag}} = 1.61$, $R_{\mathrm{validation\_set}} = 0.774$, and $R_{\mathrm{out\text{-}of\text{-}bag}} = 0.773$. Surprisingly, the RF-score performs better than any other scoring function tested on the PDBbind07 core set, although the scoring functions might even have been parametrized on some members of the validation set. Additionally, the physical description of the binding motifs, the 36 RF-score descriptors, is much simpler than descriptions used in any of the other scoring functions tested.

In order to test whether the RF-score model can be improved by varying the cutoff radius for the atom counts we tested cutoff values of 4, 6, 8, and 16 Å. The results for the PDBbind07 core set are summarized in Table 1.

It is surprising that the performance does not improve when reducing the cutoff to 6 Å, although one would expect to find the most relevant interactions within a smaller area around the ligand and the space further away to introduce noise. There is a clear drop in performance in the model at the 4 Å cutoff. The variation of the performance using the other cutoffs seems not to be significant.

The quality of performance might be explained by the ability of the random forest model to correctly assign the different clusters. This is suggested by an analysis of variances (ANOVA). The ANOVA of the PDBbind07 core set shows the following: The overall variance of the PDBbind07 core set is 5.81. The variance of the average activities of each cluster from the validation set is 2.40. The remaining variance within the clusters is 3.41. A model that assigns the average activity of the cluster as prediction would have a RMSE of sqrt(3.41) = 1.85 and an R^2 of 0.41, which is not too different from the above results. Thus, any Scoring Function providing predictions on the PDBbind07 core set with an average standard deviation higher than 1.85 might achieve its complete predictive power from assigning the mean or median of the corresponding cluster only.

Table 2. Analysis of the Ranking of Predictions for the PDBbind07core Set Clusters[a]

	order	number of occurrences
correct order	1−2−3	30
median wrong	2−1−3, 1−3−2	20
extremes wrong	2−3−1, 3−1−2	6
reverse order	3−2−1	6
		total: 62

[a] Lowest activity complex = 1, median activity complex = 2, highest activity complex = 3.

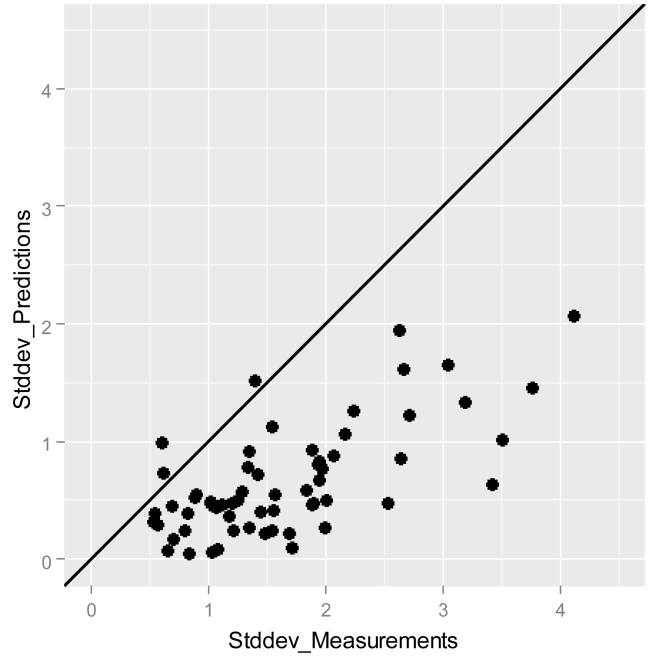

Figure 4. Standard deviations of measured and predicted activities within the clusters.

Scoring functions might unintentionally be trained to simply do a cluster assignment by counting a characteristic number of specific atoms in the binding site vicinity and look like having significant predictive power. If this was the case, the predictions would have a very small variation compared to the measurements, and the ranking within the cluster would be arbitrary. In order to test whether the RF-score scoring function benefits purely from the assignment to the correct cluster, the ranking within the clusters and the spread of measured and predicted activities has been analyzed. Table 2 shows the distribution of the rankings within the clusters. Figure 4 shows a plot of the standard deviations within the clusters of the measured versus the predicted activities.

This analysis shows that the RF-score scoring function is able to distinguish between low affinity and high affinity binders of a certain protein to some extent. However, the standard deviations of the measured activities within the clusters are higher than the standard deviations of the predicted activities in nearly all cases.

Out-of-Bag Validation. Two of the standard methods for validating QSAR models are cross-validation and out-of-bag validation. In both cases the created model is trained with a randomly chosen subset, and the remaining data points are used to assess the predictive quality. For cross-validation, the data set is usually randomly divided into m subsets, and each subset is consecutively used to assess the predictive quality of the models generated. The overall predictive quality is then assessed from the united test set predictions.

In out-of-bag validation, the training is repeated n times with a randomly chosen subset, and the samples not used for training are predicted with the generated model. Usually, this is repeated many times ($n \gg m$), such that each sample is predicted several times with a different model. The overall predictive accuracy is then assessed from the average prediction for each sample. This is the standard way of assessing the predictive power of random forests. Note that in both cases the training set is assembled randomly. For the PDBbind data set, this means that for every test/training set configuration it is very likely that there are members from each cluster in the training set.

We examined the predictive quality for a standard random forest model with random selection of the out-of-bag sets of the PDBbind09 refined set. The model obtained has correlation coefficients $R = 0.72$, $R^2 = 0.52$, and RMSE = 1.38.

Leave-Cluster-Out Validation. In order to obtain a more detailed picture of the real performance of the RF-score scoring function, we did a leave-cluster-out cross-validation. For this, we used the 2009 version of the PDBbind database. The results obtained with the new validation strategy introduced here cannot be compared to the results for other scoring functions from the previous validation studies, since we use another training and validation strategy.

For the leave-cluster-out cross-validation, we treated clusters with more than nine members individually (clusters A−W) and combined clusters with less than 10 members. These clusters were combined into three sets: (X) clusters with four to nine members, (Y) clusters with two and three members, and (Z) all singletons. These new sets represent diverse sets from different protein families which can give an impression of the overall performance of scoring functions across different unknown protein families. However, the clustering scheme ensures that the training set does not include proteins from the same family as the validation set.

In the original PDB clustering, some complexes are assigned to two or more clusters. Fortunately, this does not create conflicts for the clustering of the PDBbind subset. Only in one case does a complex belong to two different clusters with several members. It turns out that the protein units of both clusters are antibodies, so the two clusters were united.

With the clusters defined as above, the RF-score modeling strategy was evaluated with a leave-cluster-out cross-validation. The prediction results for the one-family clusters and the cluster assignments are summarized in Table 3.

The prediction quality for different protein families varies strongly. The best predictions in terms of R^2 are obtained for the β-secretase 1 complexes. The best predictions in terms of RMSE are obtained for the P38 kinase complexes. For five out of the 23 clusters, the correlation coefficient exceeds $R^2 = 0.5$. For seven complexes, the correlation coefficient is below 0.1. The out-of-bag prediction quality for the training set is highly constant across all samples. $R = 0.71 \pm 0.01$ and RMSE = 1.38 ± 0.01 vary only slightly across all sets (details not shown). They improve slightly when clusters that are predicted badly are not part of the training set. There is no correlation with training set size that is observable.

Table 3. Leave-Cluster-Out Cross-Validation Performance of RF-score, Descriptors Calculated with 12 Å Cutoff[a]

biological target			cluster left out		
validation set	cluster	#samples	R	R^2	RMSE
HIV protease	A	188	0.11	0.01	1.91
trypsin	B	74	**0.73**	**0.53**	1.04
carbonic anhydrase	C	57	0.56	0.31	1.68
thrombin	D	52	0.37	0.14	2.03
PTP1B (protein tyrosine phosphatase)	E	32	0.63	0.40	1.02
factor Xa	F	32	0.19	0.04	1.76
urokinase	G	29	**0.78**	**0.61**	0.95
different similar transporters	H	29	−0.12	0.01	1.17
c-AMP dependent kinase (PKA)	I	17	0.54	0.29	1.26
β-glucosidase	J	17	0.59	0.35	1.13
antibodies	K	16	0.58	0.34	1.57
casein kinase II	L	16	0.44	0.19	1.10
ribonuclease	M	15	0.18	0.03	1.20
thermolysin	N	14	0.68	0.46	1.09
CDK2 kinase	O	13	0.64	0.41	1.11
glutamate receptor 2	P	13	−0.20	0.04	1.16
P38 kinase	Q	13	**0.79**	**0.62**	0.59
β-secretase 1	R	12	**0.93**	**0.86**	1.51
tRNA-guanine transglycosylase	S	12	0.12	0.01	1.08
endothiapepsin	T	11	0.60	0.36	1.34
α-mannosidase 2	U	10	−0.17	0.03	1.88
carboxypeptidase A	V	10	**0.78**	**0.61**	1.71
penicillopepsin	W	10	−0.42	0.18	2.22

[a] R and RMSE for the out-of-bag predictions are $R = 0.71 \pm 0.01$ and RMSE = 1.38 ± 0.01 respectively. Clusters with R^2 results greater than 0.5 are highlighted.

Table 4. Leave-Cluster-Out Cross-Validation Performance of RF-score on the Mutiple-Family Clusters, Descriptors Calculated with 12 Å Cutoff[a]

biological target			cluster left out		
validation set	cluster	#samples	R	R^2	RMSE
all clusters with 4−9 complexes	X	387	0.56	0.31	1.63
all clusters with 2−3 complexes	Y	340	0.53	0.28	1.61
singletons	Z	321	0.44	0.19	1.75

[a] R and RMSE for the out-of-bag predictions are $R = 0.73 \pm 0.02$ and RMSE = 1.34 ± 0.03, respectively.

The prediction results and the cluster assignments for the multiple-family clusters are summarized in Table 4.

The out-of-bag prediction performance is much better in terms of both R and RMSE than the performance on the validation sets. This is to be expected, since the models are trained on the basis of members of the same protein family.

Plots of the measured versus predicted activity for each single cluster left out are shown in Figure 5.

As can be seen from the correlation coefficients and the plots, the predictions for different clusters vary strongly. There is a striking difference in the quality of performance in the different clusters. The best performance in terms of the correlation coefficient with $R = 0.93$ is achieved on the β-secretase 1 cluster (R). The best result in terms of RMSE = 0.59 is achieved on the p38 kinase cluster. The worst performance in terms of both $R = -0.417$ and RMSE = 2.22 is achieved on the penicillopepsin cluster. For the large

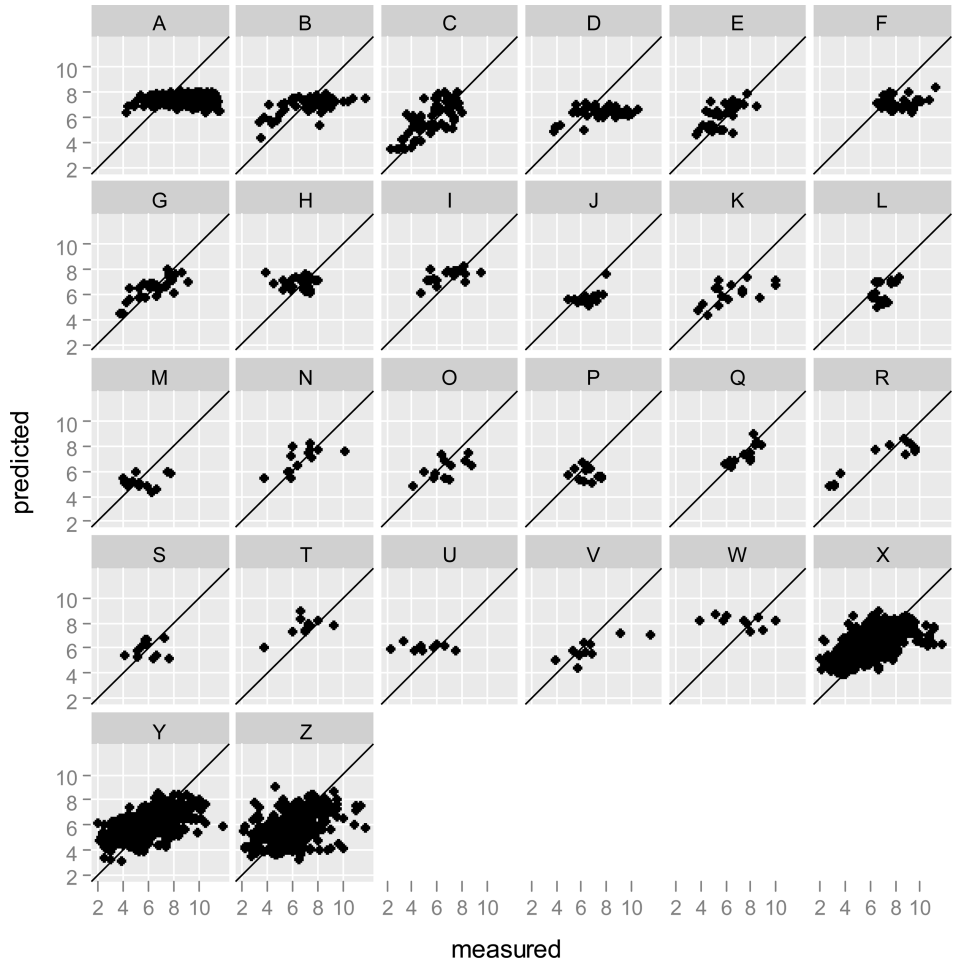

Figure 5. Measured versus predicted activity for each single cluster with models trained using leave-cluster-out training sets.

mixed clusters X, Y, and Z, it seems that the performance tends to go to $R = 0.5$ and RMSE = 1.65. This contrasts with the average performance of $R = 0.42$, RMSE = 1.40, and $R^2 = 0.29$ across the clusters and the average performance weighted by the number of complexes in each cluster $R = 0.46$, RMSE = 1.60, and $R^2 = 0.25$. However, the variation between the performance on different clusters is very high, with a standard deviation of R of $\sigma(R) = 0.35$ and a standard deviation of the RMSEs of σ(RMSE) = 0.39.

In most cases, the range of predicted values is much smaller than the range of measured values. The variances of the measurements and the predictions of each single validation set are shown in Figure 6.

We examined a smaller cutoff at the descriptors of 6 Å for the leave-cluster-out cross-validation. The performance on the different clusters is shown in Table 5. The prediction results based on the 6 Å cutoff for the multiple-family clusters are summarized in Table 6.

The average performance is $R = 0.49$, RMSE = 1.38, and $R^2 = 0.31$. The weighted average performance is $R = 0.49$, RMSE = 1.59, and $R^2 = 0.27$. This is a bit better than the performance using the 12 Å cutoffs, although not significant. The standard deviations of R $\sigma(R) = 0.26$ and RMSE σ(RMSE) = 0.38 are also smaller, indicating that the model based on the descriptors calculated with the 6 Å cutoff are not worse than the models based on the 12 Å cutoff.

The measures of quality for all different models generated are summarized in Table 7.

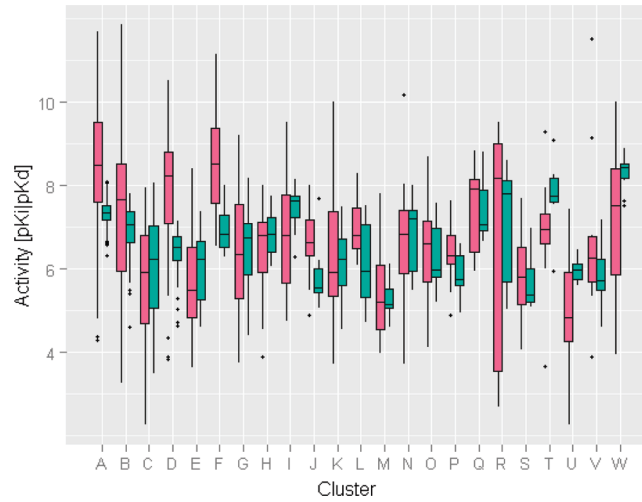

Figure 6. Distributions of measured and predicted activities for each separate cluster. Descriptors calculated with 12 Å cutoff. Measured values = red, calculated values = green.

The average performance of the leave-cluster-out cross-validation shows a strong decrease in the correlation coefficients R and R^2. The RMSE of prediction only gets a little worse. However, R is probably more important when comparing the performance of different scoring functions, since many scoring functions only provide a pseudo-energy from which an RMSE cannot directly be calculated. The relative performance of the 6 Å cutoff descriptors compared

Table 5. Leave-Cluster-Out Cross-Validation Performance of RF-score, Descriptors Calculated with 6 Å Cutoff

biological target	cluster	#samples	R	R^2	RMSE
HIV protease	A	188	0.18	0.03	1.79
trypsin	B	74	**0.73**	**0.53**	1.05
carbonic anhydrase	C	57	0.63	0.40	1.75
thrombin	D	52	0.59	0.35	2.14
PTP1B (protein tyrosine phosphatase)	E	32	0.59	0.35	1.08
factor Xa	F	32	0.12	0.01	1.67
urokinase	G	29	**0.74**	**0.55**	0.99
different similar transporters	H	29	0.33	0.11	1.12
c-AMP dependent kinase (PKA)	I	17	0.65	0.42	1.07
β-glucosidase	J	17	0.53	0.28	1.28
antibodies	K	16	0.58	0.34	1.52
casein kinase II	L	16	0.44	0.19	1.15
ribonuclease	M	15	0.23	0.05	1.12
thermolysin	N	14	0.57	0.32	1.19
CDK2 kinase	O	13	0.60	0.36	1.2
glutamate receptor 2	P	13	−0.07	0.00	1.08
P38 kinase	Q	13	**0.83**	**0.69**	0.57
β-secretase 1	R	12	**0.93**	**0.86**	1.34
tRNA-guanine transglycosylase	S	12	0.20	0.04	1.05
endothiapepsin	T	11	**0.77**	**0.59**	1.13
α-mannosidase 2	U	10	0.30	0.09	1.68
carboxypeptidase A	V	10	**0.77**	**0.59**	1.96
penicillopepsin	W	10	−0.04	0.00	1.88

Table 6. Leave-Cluster-Out Cross-Validation Performance of RF-score on the Multiple-Family Clusters, Descriptors Calculated with 6 Å Cutoff

biological target	cluster	#samples	R	R^2	RMSE
mixed clusters, 4−9 complexes each	X	387	0.56	0.31	1.63
mixed clusters, 2−3 complexes each	Y	340	0.56	0.31	1.58
singletons	Z	321	0.41	0.17	1.77

Table 7. Summary of the Performance According to Different Validation Schemes Examined

model	assessment	R	R^2	RMSE
RF-score 12 Å cutoff	PDBbind core set	0.77	0.59	1.58
	leave-cluster-out cross-validation	**0.46**	**0.21**	1.60
RF-score 6 Å cutoff	PDBbind core set	0.76	0.58	1.63
	leave-cluster-out cross-validation	**0.49**	**0.24**	1.59

to the 12 Å cutoff descriptors improves a little bit, but not significantly.

Since there is an obvious difference in the performance of the scoring function depending on whether samples from the same cluster are in the training set, it is interesting whether there are systematic differences between the clusters. We performed multidimensional[15] scaling to two dimensions on the basis of clusters A−W using the descriptors for a 12 Å cutoff. Figure 7 shows the proximities between the clusters and different members of the clusters. Most clusters occupy a distinct region, and most of their members lie close. Thus, it should be probable that a learning algorithm could assign cluster membership on the basis of the RF-score descriptors.

Within Cluster Performance. Target-specific scoring functions are different from general scoring functions. When developing target-specific scoring functions, data from complexes for the same or similar targets are available. This is the situation where it makes sense to fit the scoring function to data from the same cluster and use out-of-bag cross-validation in order to assess the predictivity. The out-of-bag performance for the four largest clusters and the mixed-family clusters trained exclusively on the cluster compounds is summarized in Table 8.

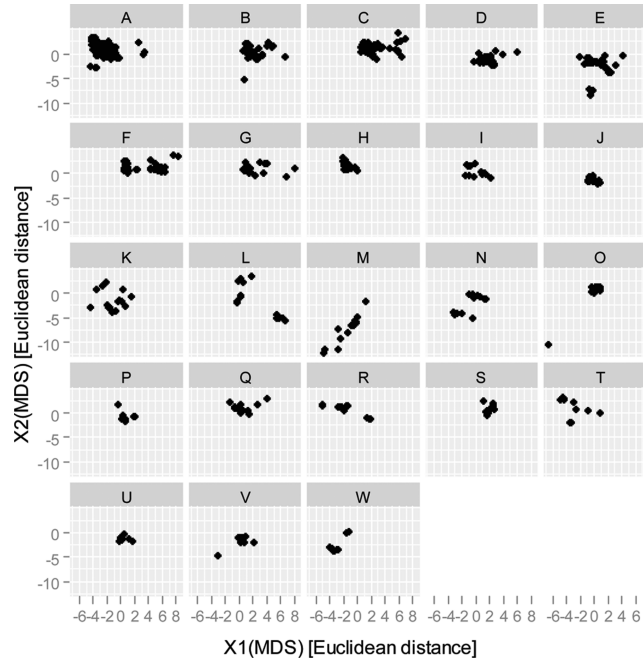

Figure 7. Distribution of clusters and their members in proximity space.

In internal out-of-bag validation, the predictions for HIV protease, thrombin, and trypsin improve strongly. The predictions for carbonic anhydrase improve slightly. This clearly shows that internal fitting can improve scoring functions systematically. The same effect can be shown with the mixed-family clusters: the more members from the same family within the cluster, the better the prediction becomes compared to the leave-cluster-out validation. The average performance for the cluster with four to nine complexes per family improves from RMSE = 1.63 to RMSE = 1.38. The average performance for the cluster with 2−3 complexes per family decreases in RMSE by 0.1 but improves in terms of R from 0.53 to 0.63. The average performance for the cluster with the singletons slightly improves in terms of RMSE but decreases in terms of R. This trend can be explained by the size of the training set, which is much smaller in the internal out-of-bag validation as compared to training based on all other external complexes. However, if there are already four to nine complexes from the same family present in the training set, the results on the validation set clearly improve.

DISCUSSION

Recently, the RF-score scoring function was introduced by Ballester and Mitchell.[8] This represents a new way of parametrizing scoring functions, departing from the traditional functional fitting procedures toward a QSAR-type fitting based on protein−ligand interaction descriptors. Despite the simplistic physical description of the interaction, RF-score performs surprisingly well, outperforming all other more complex scoring functions on the diverse PDBbind07 core validation set.

Because of the nature of the PDBbind core set, it cannot be used to give a realistic estimate of the performance of a scoring function parametrized on part of or the full PDBbind

Table 8. Comparison of Performance upon Out-of-Bag Validation Trained within the Cluster and Leave-Cluster-Out Validation

biological target			out-of-bag within cluster			cluster left out		
validation set	cluster	#samples	R	R^2	RMSE	R	R^2	RMSE
target family clusters								
HIV protease	A	188	0.67	0.45	1.17	0.11	0.01	1.91
trypsin	B	74	0.86	0.74	0.71	0.73	0.53	1.04
carbonic anhydrase	C	57	0.62	0.38	1.64	0.56	0.31	1.68
thrombin	D	52	0.69	0.48	1.09	0.37	0.14	2.03
mixed family clusters								
mixed clusters, 4−9 complexes each	X	387	0.72	0.52	1.38	0.56	0.31	1.63
mixed clusters, 2−3 complexes each	Y	340	0.63	0.40	1.71	0.53	0.28	1.61
singletons	Z	321	0.41	0.17	1.69	0.44	0.19	1.75

database. Here, we have shown that scoring functions can gain significant predictive power by just assigning the protein family to which a certain complex belongs. This prediction cannot be the target of scoring functions, and therefore a thorough validation strategy should exclude effects based on the assignment of the target family. Thus, we propose using leave-cluster-out cross-validation for scoring functions. This ensures that there cannot be an artificial performance improvement by cluster-assignment and that the variance between different clusters can be estimated on a statistically sound basis.

Here, we have executed the leave-cluster-out cross-validation of RF-score on the PDBbind09 refined set. We find that overall performance in terms of R decreases significantly from $R = 0.77$ to $R = 0.49$ and from $R^2 = 0.59$ to $R^2 = 0.21$. The overall performance in terms of RMSE does not change much. However, in many cases, RMSE cannot be used for directly comparing different scoring functions since many scoring functions only output pseudo-energies where the absolute values cannot be used. Additionally, one can see that the performance is very different on individual clusters, ranging between zero correlation to almost perfect correlation ($R = 0.93$ and RMSE $= 0.59$). To avoid an overoptimistic bias due to the validation strategy, we recommend using leave-cluster-out cross-validation for all future studies that compare the performance of scoring functions on diverse protein−ligand complexes.

We have shown on HIV-protease, trypsin, thrombin, and carbonic anhydrase that scoring functions for specific targets can, but must not, be significantly improved by using target-specific training data. Further, we find that the performance in the leave-cluster-out cross-validation does not decrease upon using RF-score descriptors with a 6 Å cutoff instead of a 12 Å cutoff.

While there is large variability between the overall performance within different clusters, the average performance for large data set goes toward RMSE $= 1.6$ and $R^2 = 0.2-0.3$. This should be seen in context. The RMSE of a model that always predicts the average of the complete PDBbind2009 refined set would be 1.98, equaling the standard deviation of the measured activities. A model that always predicts the average of the cluster using the PDBbind09 refined set has an RMSE of 1.74 and an R^2 of 0.20. Any future scoring function will have to outperform the lower limits defined by this baseline model. The descriptors used in RF-score may be used to generate baseline models that have to be improved upon by models including more physical complexity. However, care must be taken when comparing other scoring functions that have been built in a traditional manner using weighted distance-dependent potentials, since they have probably been built on the basis of a subset of the PDBbind refined set. With the QSAR-type parametrization procedure that Ballester and Mitchell have published with RF-score, it will hopefully be possible to generate better scoring functions based on better descriptors for the significant protein−ligand interactions.

Future evaluations of RF-score will have to show whether it is not only able to predict the free energy of binding of complexes with known geometry but if it is also able to perform well in real docking scenarios with decoys and unknown complex structures.

SUMMARY AND CONCLUSIONS

Scoring functions are still far from perfect. However, in RF-score, a new way of QSAR-type parametrization for scoring functions has been shown to not only be feasible but also provide results that are very competitive with those of other scoring functions. For future scoring functions based and/or validated on the PDBbind database or a subset thereof, it is necessary to employ a different validation strategy, namely, leave-cluster-out cross-validation. The presence of a number of protein families within the PDBbind database allows scoring functions to look much better than they really are, if the same protein families are present in both the training and validation set. Additionally, leave-cluster-out cross-validation provides information of the variance of the performance between different clusters. Here, we have shown that the performance of RF-score between different clusters indeed varies strongly, tending to give an average performance of $R = 0.5$ ($R^2 = 0.2-0.3$) and RMSE $= 1.6$ for large diverse validation sets.

Given the bad performance on a very general and diverse set such as the PDBbind of other scoring functions trained on smaller data sets, improving standard scoring schemes should be possible by simply refitting the functions to a larger data set. With this publication, we provide a list of cluster assignments for all the complexes in the PDBbind 2009 refined set that can be used to carry out reasonable leave-cluster-out cross-validation for future scoring functions and comparisons among different scoring functions.

ACKNOWLEDGMENT

C.K. thanks the Education Office of NIBR for a Postdoctoral Fellowship. We would also like to thank Reinhard Bergmann for advice on statistical questions.

Supporting Information Available: The cluster annotation is available. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Coupez, B.; Lewis, R. A. Docking and Scoring - Theoretically Easy, Practically Impossible. *Curr. Med. Chem.* **2006**, *13*, 2995−3003.

(2) Kroemer, R. T. Structure-based drug design: docking and scoring. *Curr. Protein Pept. Sci* **2007**, *8*, 312–328.

(3) Jain, A. N. Scoring functions for protein-ligand docking. *Curr. Protein Pept. Sci* **2006**, *7*, 407–420.

(4) Kontoyianni, M.; Madhav, P.; Suchanek, E.; Seibel, W. Theoretical and practical considerations in virtual screening: a beaten field. *Curr. Med. Chem* **2008**, *15*, 107–116.

(5) Kolb, P.; Irwin, J. J. Docking screens: right for the right reasons. *Curr. Top. Med. Chem.* **2009**, *9*, 755–770.

(6) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.

(7) 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.

(8) 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.

(9) 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.

(10) Wang, R.; Fang, X.; Lu, Y.; Yang, C.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.

(11) Benson, M. L.; Smith, R. D.; Khazanov, N. A.; Dimcheff, B.; Beaver, J.; Dresslar, P.; Nerothin, J.; Carlson, H. A. Binding MOAD, a high-quality protein ligand database. *Nucleic Acids Res.* **2008**, *36*, D674−678.

(12) 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.

(13) R Development Core Team. *R: A Language and Environment for Statistical Computing*; The R project for statistical computing; R Foundation for Statistical Computing: Vienna, Austria, 2009.

(14) Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.

(15) Gower, J. C. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **1966**, *53*, 325–338.

CI100264E