JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

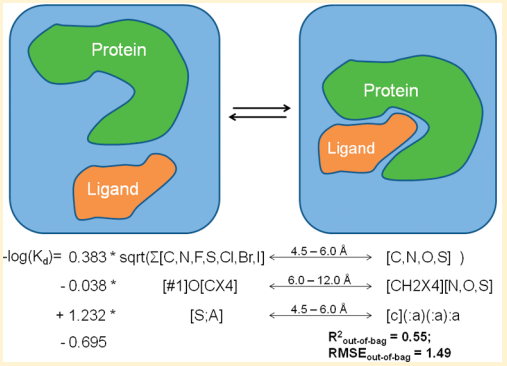
pubs.acs.org/jcim

Three Descriptor Model Sets a High Standard for the CSAR-NRC HiQ Benchmark

Christian Kramer*,† and Peter Gedeck†

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

ABSTRACT: Here we report the results we obtained with a proteochemometric approach for predicting ligand binding free energies of the CSAR-NRC HiQ benchmark data set. Using distance-dependent atom-type pair descriptors in a bagged stepwise multiple-linear regression (MLR) model with subsequent complexity reduction we were able to identify three descriptors that can be used to build a very robust regression model for the CSAR-NRC HiQ data set. The model has an R^2_{cv} of 0.55, a MUE_{cv} of 1.19, and an $RMSE_{cv}$ of 1.49 on the out-of-bag test set. The descriptors selected are the count of protein atoms in a shell between 4.5 Å and 6 Å around each heavy ligand atom excluding oxygen and phosphorus, the count of sulfur atoms in the vicinity of tryptophan, and the count of aliphatic ligand hydroxy hydrogens. The first two descriptors have a positive sign indicating that they contribute favorably to the binding energy, whereas the count of hydroxy hydrogens contributes unfavorably to the binding free energy observed. The fact that such a simple model can be so effective raises a couple of questions that are addressed in the article.



INTRODUCTION

Binding to a specific binding site of the target receptor is the *conditio sine qua non* of the overwhelming majority of drugs and agrochemicals.[1] Docking/scoring techniques have been used for a long time to predict and rationalize this most central binding event.[2−7] While the *a priori* prediction of interaction geometries has been proven feasible in many cases, prediction of interaction energies remains a challenging task.

In the last years, a couple of databases of protein−ligand interaction structures plus binding energies have been assembled from primary literature, with the most prominent among them being PDBbind[8,9] and BindingMOAD.[10] These data collections offer the unique possibility of separating the pose finding problem from the energy prediction problem. However, even if the crystal structures of the protein−ligand complexes are available their interpretation is not always straightforward, since the resolution is limited, hydrogens are not visible, and histidine or glutamine rotamers are not distinguishable from the X-ray resolution. We thus very much appreciate the efforts of the CSARdock team of Carlson et al. to assemble the CSAR-NRC HiQ data set, where all these issues have been addressed and solved to a scientifically agreeable level.

Recently, we introduced a new set of descriptors describing protein−ligand interactions.[11] Protein and ligand atoms are assigned a Crippen-like atom type[12] and for each pair of protein atom type/ligand atom type the distances are binned and occurrences counted. We have shown that using standard QSAR methods and these descriptors, free energy scoring functions for the PDBbind09 database can be generated that currently are at least among the best scoring functions validated on diverse collections of binding data plus crystal structures. In this manuscript we describe how the fitting method can be extended with a complexity reduction step to further improve the scoring function. We apply the descriptors and the fitting method to the CSAR-NRC HiQ data set and show that a simple QSAR-type scoring function based on three interpretable descriptors is capable of predicting 55% of the total variance of the CSAR-NRC HiQ data set. This result comes as a surprise, since it is as good as or better than most other, more complex protein−ligand scoring functions based on distance- and angle dependent terms.

MATERIALS AND METHODS

We have used the optimized CSAR-NRC HiQ data set as available on http://csardock.org/ on the 16th of November, 2010 for fitting the models. This data set consists of 343 protein−ligand complexes that should be of very high quality.

Distance-dependent atom type pair descriptors have been calculated for all complexes. They consist of counts of pairs of atom types in specific ranges. We used an atom typing scheme very similar to the Crippen approach, which has proven very successful for a couple of QSAR/QSPR models such as for example prediction of $logP_{ow}$.[12] A more detailed manuscript about the atom typing scheme including SMARTS patterns was recently published in this journal.[11] In brief, for the ligand there are 84 different atom types with 29 atom types for carbons, 17 atom types for nitrogen, 14 atom types for oxygen, 16 atom types for hydrogen, 3 atom types for sulfur. For F, P, Cl, Br, I, and

Special Issue: CSAR 2010 Scoring Exercise

Received: January 21, 2011
Published: May 30, 2011

each metal (protein only) there is one atom type. For the protein there are overall 39 atom types, since approximately half of the Crippen atom types do not occur in the twenty natural amino acids and cofactors in the crystals. For every complex, all pairs of ligand atom type protein atom type are counted and summarized in ranges with thresholds 3.0, 3.5, 4.0, 4.5, 6.0, and 12.0 Å. Additionally the counts are summarized for each atom type on the ligand and for all protein atoms that are directly in the binding pocket, which we defined as all protein atoms that closer than 4 Å to any ligand atom. On the CSAR-NRC HiQ data set this gives a set of 6061 descriptors with at least five nonzero entries on each descriptor.

Additionally we have calculated distance-dependent element counts, where the highly specific atom types are replaced by the parent elements (excluding hydrogens). This gives 288 descriptors. After discovering that the single C−C pair count between 4.5 and 6 Å has a correlation of $R^2 = 0.41$ with the binding affinity, we generated a new descriptor by combining all C, N, F, S, Cl, Br, and I atoms for the ligand and all heavy atoms for the protein. This increased the correlation with the free energy of binding to $R^2 = 0.44$. Further inclusion of ligand oxygen and ligand phosphorus leads to a decrease in the correlation. Finally we found that taking the square root of this new descriptor increases the correlation with the free energy of binding observed to 0.45. As a last descriptor set we have added simple ligand-only descriptors from MOE like the SlogP, the count of rotatable bonds, atoms counts, counts of donors and acceptors, and aromatic rings. These are 23 descriptors.

For fitting the models we used a bagged stepwise multiple linear regression[13] with the descriptor-pool-size adjusted F-value as stopping criterion[14] and subsequent pruning. For each of the 50 independent bagging samples 25% of the whole data set are randomly selected and left out as independent test set. The remaining 75% of the data set are used as training set. Descriptors are selected iteratively until the gain in prediction performance on the training set is no better than the gain obtained from 95% of random descriptors. For each single bagging model the left-out test set is predicted. In the end, all predictions are united by averaging the predictions for the out-of-bag test cases. On average every compound has been part of the test set 12.5 times. The overall model and predictions are obtained from averaging the coefficients and the predictions from all single bagging models.

The significance of the descriptors can be inferred from the number of times they have been selected in all models. Highly significant descriptors should be selected in most models, whereas descriptors that happen to just pass the significance threshold for the training set of one specific submodel are probably insignificant overall. The frequency criterion can only be used in models which are based on several similar submodels. It can particularly not be used with a single test/training set split. In our experience, usually only a few descriptors are selected frequently, while most descriptors are selected once or twice. Thus in a final complexity reduction step descriptors that have been selected only a few times can be removed − here we have chosen to remove all descriptors that have been selected in less than half of the models. This should make the final model even more robust, since it now consists of only a few descriptors.

Measures of Quality. We use the root mean squared error (RMSE), the mean unsigned error (MUE), the median absolute error ($M_{ed}AE$), the predictive R^2, Spearman's rank correlation coefficient ($R_{Spearman}$), and Kendall's Tau ($\tau_{kendall}$) as correlation coefficient to assess the quality of the relative predictions. We use R^2 instead of R because it estimates the proportion of the variance of the data set explained by the model and since the seminal works of Gauss[15] and Laplace[16] at the beginning of the 19th century it is known that there are a lot of good theoretical reasons to reduce the square of the error (i.e., the remaining variance). RMSE and MUE measure the absolute accuracy of the prediction, i.e. how well the experimental value is reproduced by the model. In contrast to Pearson's R^2 the predictive R^2 assesses how close the predictions are to the absolute target value. Parallel shifted predictions get a low predictive R^2 and a high Pearson's R^2. The two rank correlation coefficients $R_{Spearman}$ and $\tau_{Kendall}$ measure the relative predictions within the data set. $R_{Spearman}$ assumes equal distances between the measured values. $\tau_{Kendall}$ is less parametric than $R_{Spearman}$, only using the relative ranks

$$\mathrm{MUE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_{i,pred} - y_{i,meas}\right|$$

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

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{N}\left(y_{i,preds} - y_{i,means}\right)^2}{\sum_{i=1}^{N}\left(y_{i,means} - \bar{y}\right)^2}\right); \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$R_{Pearson} = \frac{\sum_{i=1}^{N}\left(y_{i,meas} - \bar{y}_{meas}\right)\left(y_{i,pred} - \bar{y}_{pred}\right)}{\sqrt{\sum_{i=1}^{N}\left(y_{i,meas} - \bar{y}_{meas}\right)^2}\sqrt{\sum_{i=1}^{N}\left(y_{i,pred} - \bar{y}_{pred}\right)^2}}$$

$$R_{Spearman} = \frac{\sum_{i=1}^{N}\left(rnk(y_{i,meas}) - rnk(\bar{y})\right)\left(rnk(y_{i,pred}) - rnk(\bar{y})\right)}{\sqrt{\sum_{i=1}^{N}\left(rnk(y_{i,meas}) - rnk(\bar{y})\right)^2}\sqrt{\sum_{i=1}^{N}\left(rnk(y_{i,pred}) - rnk(\bar{y})\right)^2}}$$

$$\tau_{kendall} = \frac{\sum_{i=1}^{n-1}\sum_{j>1}^{n} order_{ij}}{\sqrt{\sum_{i=1}^{n-1}\sum_{j>1}^{n}\left|order_{ij}\right| + equals_{ij,meas}}\sqrt{\sum_{i=1}^{n-1}\sum_{j>1}^{n}\left|order_{ij}\right| + equals_{ij,pred}}}$$

With

$$order_{ij} = \begin{cases} 1 & if\left(y_{i,meas} - y_{j,meas}\right)\left(y_{i,pred} - y_{j,pred}\right) > 0 \\ -1 & if\left(y_{i,meas} - y_{j,meas}\right)\left(y_{i,pred} - y_{j,pred}\right) < 0 \\ 0 & else \end{cases}$$

and

$$equals_{ij,x} = \begin{cases} 1 & if \quad y_{i,x} = y_{j,x} \\ 0 & else \end{cases}$$

Here, N is the number of instances, $y_{i,pred}$ is the predicted binding energy, and $y_{i,meas}$ is the measured binding energy.

Table 1. Performance of the Standard Bagged MLR Model and the Pruned Model

	bagged MLR model			pruned bagged MLR model		
	R^2	RMSE	MUE	R^2	RMSE	MUE
training set	0.62	1.38	1.11	0.57	1.47	1.17
test set	0.52	1.54	1.22	0.55	1.49	1.19

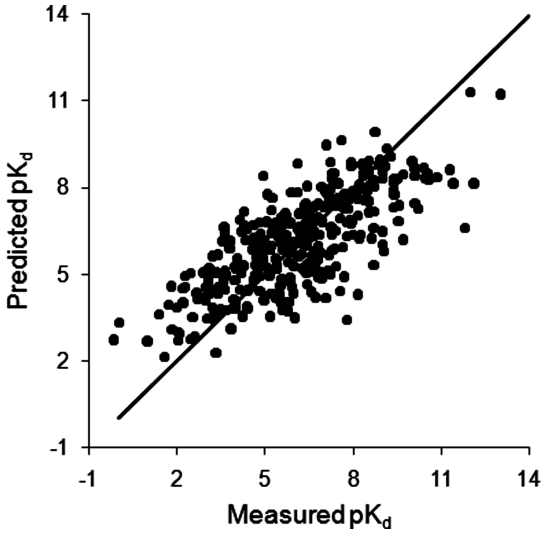


Figure 1. Out-of-bag predicted versus measured pK_d, complexity-reduced bagged MLR model.

RESULTS

Training a model with the combined set of the distance-dependent atom type pair counts, the distance-dependent element counts and the MOE descriptors we obtained a model with the out-of-bag test set performance of R^2 = 0.52, RMSE = 1.54, and MUE = 1.21. 33 descriptors have entered at least one of the bagging models, but only three descriptors have entered more than 50% of the bagging models. Training a second model with the three descriptors only gives a model with test set performance of R^2 = 0.55, $R_{Pearson}$ = 0.74, $R_{Spearman}$ = 0.73, $\tau_{Kendall}$ = 0.53, RMSE = 1.49, MUE = 1.19, and M_{ed}AE = 1.05. Interestingly the R^2 for the training set of the first model is R^2 = 0.62, while the R^2 for the training set of the complexity-reduced model is R^2 = 0.57. So the performance of the pruned model in terms of R^2 is better, and the training set and the test set prediction qualities become very similar. This supports the view that the model is very robust. The performances are summarized in Table 1.

A plot of predicted versus measured pK_d is shown in Figure 1.

The pruned model is based on three descriptors. The first and most important descriptor in terms of correlation with the observed binding affinity is the square root of the count of all ligand[C,N,F,S,Cl,Br,I]-protein[C,N,O,S] element pairs in a distance between 4.5 and 6 Å. We found this descriptor when we observed that the ligand_carbon-protein_carbon count in the distance of 4.5 to 6 Å had a correlation of 0.41 with the binding affinity. Being surprised by the fact that the carbon atom type does not seem to make a difference, we investigated whether or not the element type makes a difference. We found that adding the counts for all other protein heavy atoms in the 4.5 to 6 Å range and the counts for the ligand N, F, S, Cl, Br, and I atoms even improves the correlation. Adding the counts for ligand O and P decreases the correlation. Figure 2 shows a plot of the binding affinity versus the sqrt(count(L[C,N,F,S,Cl,Br,I]-P[C,N,O,S])) descriptor and a histogram of the distribution of the values. This descriptor enters the overall equation with a standardized coefficient of 1.59.

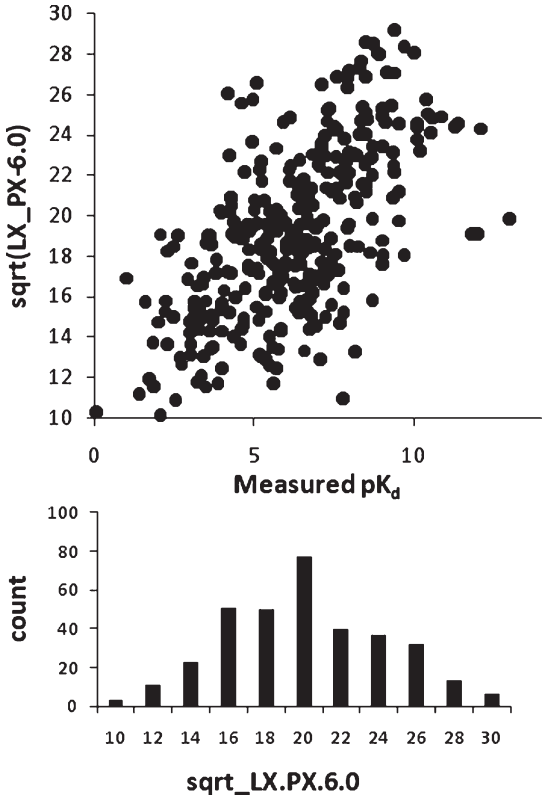


Figure 2. (a) pK_d versus sqrt(count(L[C,N,F,S,Cl,Br,I]-P[C,N,O,S])), $R^2_{Pearson}$ = 0.45 and (b) distribution of sqrt_LX.PX.6.0.

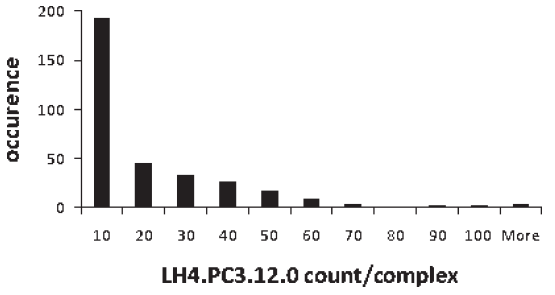


Figure 3. Distribution of LH4.PC3.12.0.

The second most important descriptor is the count of all ligand hydroxy hydrogens in a distance of 6.0 to 12.0 Å to aliphatic protein carbon atoms bound to two nonaromatic heavy atoms with at least one noncarbon heavy atom among them (possible H-bond donors/acceptors). This protein atom type occurs in methionine, proline, glycine, serine, cysteine, lysine, and arginine. Around 47% of all the ligands have a value different from zero on this descriptor. It has a negative coefficient of −0.77 on the standardized set, meaning that the above-mentioned amino acids in the binding pocket contribute unfavorably to free

dx.doi.org/10.1021/ci200030h |J. Chem. Inf. Model. 2011, 51, 2139–2145

Table 2. Standardized and Direct Coefficients of the Pruned Model

descriptor	coefficient standardized	coefficient direct
sqrt_LX.PX.6.0	1.59	0.383
LH4.PC3.12.0	−0.77	−0.038
LS1.PC19.6.0	0.47	1.232
constant		−0.695

energy of binding predicted if the ligand carries hydroxy hydrogens. The distribution of values on this descriptor is shown in Figure 3.

The third most important descriptor is the count of ligand uncharged aliphatic sulfur in a distance of 4.5 to 6.0 Å to ligand aromatic carbon atoms with three aromatic neighbors. This atom type occurs in the proteins only for the two carbon atoms that share both aromatic rings in tryptophan. Although there are only twelve complexes where this descriptor has a value different from zero, it is significant enough to be selected in most bagging models. Without this descriptor the performance drops to R^2 = 0.51 and RMSE = 1.55.

The standardized and direct coefficients of the pruned model are listed in Table 2.

DISCUSSION

In this article, we have presented a scoring function for the free energy of binding that is based on three interpretable descriptors. The scoring function is able to predict 55% of the total variance of the CSAR-NRC HiQ data set in out-of-bag validation with R^2 = 0.55, RMSE = 1.49, and MUE = 1.19.

The model is based on three simple descriptors: The first descriptor is the square root of the count of all ligand heavy atoms (excluding oxygen and phosphorus) − protein heavy atom pairs in a distance of 4.5 to 6 Å. This descriptor is highly unspecific, since it does depend neither on atom type, nor on orientation or even element type. It increases with the number of bound heavy atoms and the buriedness of the ligand − deeply buried ligand atoms have more protein neighbors than surface bound ligand atoms. This descriptor is very different from molecular weight (MWt): although MWt is correlated with sqrt_LX.PX.6.0 with R^2 = 0.7, the correlation of MWt with the binding affinity is only R^2 = 0.25. The correlation of sqrt_LX.PX.6.0 with the binding affinity is R^2 = 0.45. This is a huge difference and to our knowledge this counts representing a combination of ligand size (corrected by oxygen and phosphorus) and buriedness has not been observed before. It also needs to be noted that oxygen and phosphorus (mostly appearing as phosphate in this data set) are not counted. We assume that this is due to the fact that both can undergo many hydrogen bonds with water and are thus very hydrophilic and contribute unfavorably to the binding affinity, if they are transferred from water to the binding site.

After submitting the predictions for the CSAR-NRC HiQ benchmark we found that it is not necessary to take the square root; the performances of the models with and without square root on this descriptor are very similar. In order to publish the model we used for the submission of our scores, we decided to keep the square root in this manuscript.

The second descriptor is a count of all pairs of ligand hydroxy hydrogens in a distance of 6.0 to 12.0 Å to aliphatic protein carbon atoms which are bound to two nonaromatic heavy atoms with at least one noncarbon heavy atom among them. The protein carbon atom type occurs in methionine, proline, glycine, serine, cysteine, lysine, and arginine. The distance between the two atoms is quite large, so this descriptor does not code for direct interactions.

We tried to replace this descriptor with simpler descriptors and found that the count of hydroxy oxygens in the ligand and the count of the aliphatic hydroxy hydrogens can replace the LH4.PC3.12.0 descriptor while almost retaining the performance of the model. The coefficients yielded are nearly the same with −0.6 to −0.7 log units per hydroxy group. This gives a model with an out-of-bag R^2 of 0.53, compared to R^2 = 0.55 for the best model. The counts of above-mentioned carbon atoms in the binding pocket (<4 Å distance to any ligand atom) as well as the counts of the noncarbon protein heavy atoms in the binding pocket do not enter the model. We also tried to replace this descriptor with ligand donor- and acceptor counts. Here only the acceptor count(which is correlated with the hydroxy group count with R^2 = 0.67) was significant, but the model became worse (R^2 = 0.51) compared to the model based on the hydroxy count.

In order to publish the model that we have used to make the CSAR-NRC HiQ benchmark predictions we decided to concentrate on the pair descriptor. It has a negative coefficient which fits to the fact that hydrophilic parts prefer being in contact with water. If this descriptor is a pure count of ligand hydroxy atoms, it complements the first descriptor which is a count of all ligand atoms without oxygens. Again this descriptor is rather unspecific.

The third descriptor is highly specific. It is the count of the pairs of ligand uncharged nonaromatic sulfur atom - protein tryptophan aromatic atoms being part of both aromatic rings in a distance between 4.5 and 6 Å. This descriptor pops up for two different kinds of ligand atoms: Sulfonamides at the proximity of tryptophans with the lone pair of the carboxy oxygen in the plane of tryptophan and for three complexes (2epn, 1swk, 2c1q) where a nonaromatic rings containing sulfur is surrounded by several tryptophans. In Figure 3 the interactions are shown for 2pov and 1swk.

This is the only specific interaction that we were able to identify that significantly contributes to the observed binding energy. Every count of this interaction contributes 1.23 pK_d values to the predicted binding energy. If we exclude the two biotin complexes from the CSARdock data set, the descriptor still gets selected, however with a slightly lower coefficient (0.8 pK_d units/interaction count, compared to 1.2 if both biotin complexes are included in the training). Including this descriptor the error for 8 out of the 12 complexes having a nonzero value in this descriptor decreases, whereas the error for 4 of the 12 complexes increases. This descriptor also affects the prediction performance on the other complexes. Leaving out this desriptor, the median error for the complexes which have zero counts on this descriptor increases by 0.02 pK_d units.

We did not identify any descriptor that accounts for ligand−metal interactions, although this occurs in many more examples than for example the sulfur-tryptophan interaction. Probably the true interaction that causes the large gain of energy is not the interaction of the sulfur with the tryptophans but rather the interaction of some probably hydrophobic atoms or fragments close to the sulfur. The descriptor has a coefficient of 1.23, meaning that this interaction can contribute quite a lot to the binding energy predicted. Although this interaction only occurs in twelve of the 343 complexes, it can be identified having a significant contribution to the binding energy.

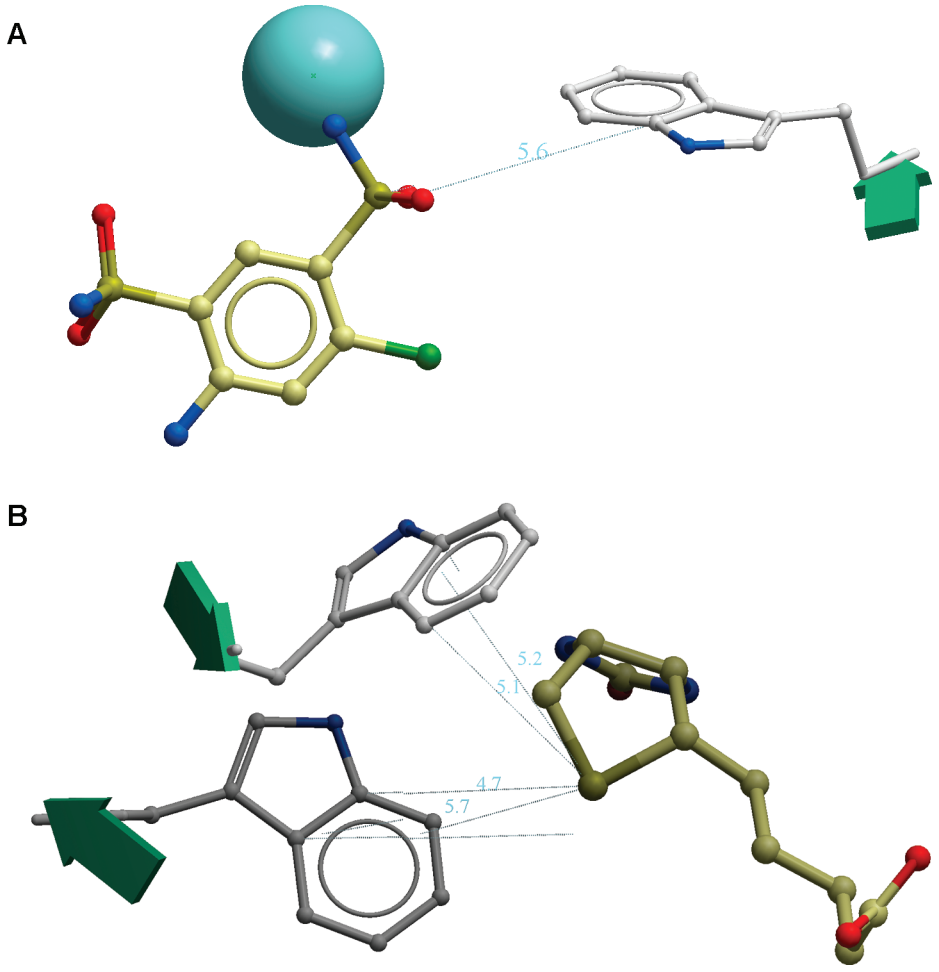


Figure 4. (a) Interaction of sulfonamide with tryptophan in 2pov (including the Zn^{2+} ion in the binding site) and (b) interaction of biotin with tryptophan of streptavidin in 1swk.

The constant of the model is rather small with −0.69. We seem to have found an equation that is additive and approaches a value close to zero for very small ligands. The predicted values should approach zero as the molecule gets smaller, since the binding energy is based on the difference between free energy in solution and free energy in the bound state. As compounds get smaller this energy should get smaller and a nonexisting hypothetical compound with zero on all descriptors that cannot interact should have the same energy (=zero) in both solution state and bound state.

In order to judge the quality of a scoring function it is necessary to know the experimental uncertainty to be able to compare the remaining variance to the variance one should expect from experimental data. For $\log P_{ow}$, which is the distribution coefficient between water and octanol a very simple standard measurement, it is reasonable to assume an experimental uncertainty between laboratories of 0.3−0.5 log units. For pK_d data, we do not know the experimental uncertainty. The median experimental uncertainty of 0.05 log units as given in the CSARdock database seems very overoptimistic, given that there are many more sources for uncertainties in biological measurements than in $\log P_{ow}$ measurements. (Note that $\log P_{ow}$ and pK_d are on the same energetic scale: $\log P_{ow}$ is the quotient of the free energy in octanol and in water, pK_d is the quotient of the free energy in water and in the binding pocket). In a data set such as the CSARdock data set unavoidably there are differences in temperature, buffer, air pressure, state of the biological material, other lab material, and so on for each measurement that all contribute to the measured energies. For public data of binding to biological targets the uncertainty should be larger than 0.3−0.5 log units, but we have no idea of how large this should be and how good our models can be in the light of this data.

The fitting procedure we have used delivers very conservative, well validated models. The initial model was based on four to six descriptors per submodel, whereas the complexity-reduced model consists of three descriptors. The performance on the training set (R^2 = 0.57) is very close to the performance on the test set (R^2 = 0.55), which is a further indicator for the robustness of the model. All test set predictions are obtained in an out-of-bag mode, which means that the test set predictions are really independent. In the pruning step descriptors which have been selected in less than 50% of the initial models are removed. This results in better models with increased performance.

Compared to other scoring functions the model presented here is simple. The fact that, despite or maybe even because of its simplicity, the model still performs so well, raises a couple of questions that we try to address here:

- Is it overtraining?

  This is rather unlikely. The final model consists of only three descriptors that have been selected in a conservative manner using the descriptor pool size adjusted F-value. For this model

it is crucial to use a modeling strategy that can prevent overtraining, since the ratio of descriptors available to samples is very large. The bagged stepwise multiple linear regression with descriptor pool size adjusted F-value and subsequent complexity reduction is one of the few techniques we are aware of which can deal with such a huge number of descriptors. Here this can be seen from the fact that the performance on the out-of-bag test set is almost the same as the performance on the training set, showing that this model is very robust.

- Why are there only three descriptors in the model? (We should know from theory that there are a lot of different interactions present in protein−ligand binding all having different contributions!)

This might be due to the modeling procedure. Since initially there are ∼6.400 descriptors for training the model the significance level (the F-value) is rather high. In order to be identified as a significant contributor a descriptor must have a higher correlation compared to the situation where for example only a set of 64 descriptors is available. Interactions that have a very weak contribution might not be identified. This situation however can be improved in the future with larger training data sets. Nevertheless the approach used here was able to identify a seldom but strong specific interaction that only occurs in twelve compounds.

- Why do distance- and angle-dependent functions not perform better?

We can only speculate. Maybe the effects of entropy-enthalpy compensation are so strong that the contributions of very specific interactions are damped. One should also keep in mind that in reality the bound state is not the single snapshot that is obtained with an X-ray structure but rather an ensemble of structures due to the dynamics at room temperature. Present angle- and distance-dependent scoring equations might not correctly capture the dynamic nature of ligand binding. On the other hand maybe the best angle- and distance-dependent terms have not yet been found or we might have to use higher-level methods for calculating interaction energies.

- Can this scoring function be used for real projects?

The model presented here is based on positive examples only. It can probably be used to estimate some upper limit for the free energy of binding. However there is no reason to assume that it could filter out nonbinders, because it has not been trained on nonbinders. The intention of the CSAR-NRC HiQ benchmark was not to provide a training set for practically usable scoring functions but rather to assess the power of predicting free energies of binding when the interaction geometry is known. For training real scoring functions, a set of decoys for each receptor is necessary and a training mechanism that can deal with hypothetical wrong interaction geometries. Maybe then the angle- and distance dependent terms become important again.

CONCLUSIONS

We have presented a scoring function for the CSAR-NRC HiQ benchmark data set that is capable of predicting 55% of the total variance of the free energies of binding measured. It is based on three interpretable descriptors and a very conservative well validated training method. The performance is very robust with $R^2 = 0.57$, RMSE = 1.47, and MUE = 1.17 for the training set and $R^2 = 0.55$, RMSE = 1.49, and MUE = 1.19 for the out-of-bag test set. This scoring function illustrates that it is possible to predict a high proportion of the free energy of binding with rather simple methods, given that the interaction structure is known. The two major descriptors discovered are very unspecific and represent counts of ligand (C,N,F,S,Cl,Br,I) − protein (C,N,O,S) element pairs in 4.5−6 Å distance and the count of all aliphatic ligand hydroxy groups weighted by their buriedness. The third descriptor codes for the interaction between ligand sulfur and protein tryptophan.

For real applications it would be necessary to complement the data set with decoys and to develop training routines that can take into account several structures for the decoys. Additionally for training future scoring functions it is critically necessary to know the experimental uncertainty associated with public K_d data in order to assess what performance can be expected.

AUTHOR INFORMATION

Corresponding Author

*E-mail: christian.kramer@novartis.com.

ACKNOWLEDGMENT

C.K. thanks the Novartis Institutes for BioMedical Research education office for a Presidential PostDoc Fellowship.

REFERENCES

(1) Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.* **2010**, *53*, 5061–5084.

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

(3) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.* **2004**, *47*, 3032–3047.

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

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

(6) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of Protein−Ligand Interactions. Docking and Scoring: Successes and Gaps. *J. Med. Chem.* **2006**, *49*, 5851–5855.

(7) Taylor, R.; Jewsbury, P.; Essex, J. A review of protein-small molecule docking methods. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 151–166.

(8) 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.

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

(10) 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.

(11) Kramer, C.; Gedeck, P. Global Free Energy Scoring Functions based on distance-dependent Atom-type Pair Descriptors. *J. Chem. Inf. Model.* **2011**, *51*, 707–720.

(12) Wildman, S. A; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.

(13) Kramer, C.; Beck, B.; Clark, T. A Surface-Integral Model for Log POW. *J. Chem. Inf. Model.* **2010**, *50*, 429–436.

(14) Kramer, C.; Tautermann, C. S.; Livingstone, D. J.; Salt, D. W.; Whitley, D. C.; Beck, B.; Clark, T. Sharpening the Toolbox of

Computational Chemistry: A New Approximation of Critical F-Values for Multiple Linear Regression. *J. Chem. Inf. Model.* **2009**, *49*, 28–34.

(15) Gauss, C. F. Theory of the motion of celestial bodies in a conical section of the sun's environment; Perthes und Besser: Hamburg, Germany, 1809.

(16) Legendre, A. M. Appendix: On the method of least squares. In *New methods for determining the orbits of comets with a supplement that contains several improvements on those methods and their application to two 1805 comets*. Ed.: Firmin Didot, Paris, 1805.