JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

ARTICLE

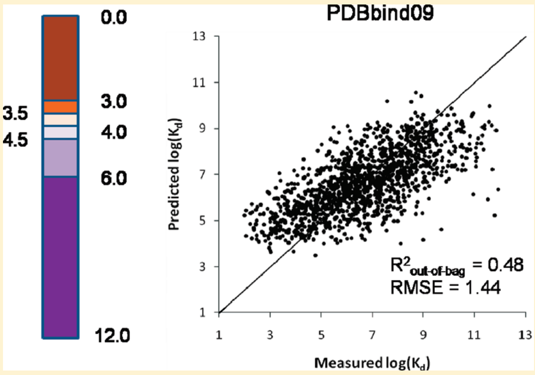
pubs.acs.org/jcim

Global Free Energy Scoring Functions Based on Distance-Dependent Atom-Type Pair Descriptors

Christian Kramer* and Peter Gedeck

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

Supporting Information



ABSTRACT: Scoring functions for protein−ligand docking have received much attention in the past two decades. In many cases, remarkable success has been demonstrated in predicting the correct geometry of interaction. On independent test sets, however, the predicted binding energies or scores correlate only slightly with the observed free energies of binding.

In this study, we analyze how well free energies of binding can be predicted on the basis of crystal structures using traditional QSAR techniques in a proteochemometric approach. We introduce a new set of protein−ligand interaction descriptors on the basis of distance-binned Crippen-like atom type pairs. A subset of the publicly available PDBbind09-CN refined set (MW < 900 g/mol, #*P* < 2, ndon + nacc < 20; $N = 1387$) is being used as data set. It is demonstrated how simple, yet surprisingly good, scoring functions can be generated for the whole diverse database ($R^2_{\text{out-of-bag}}$ = 0.48, R_p = 0.69, RMSE = 1.44, MUE = 1.14) and individual protein family subsets. This performance is significantly better than the performance of almost all other scoring functions published that have been validated on a test set as large and diverse as the PDBbind refined set.

We also find that on some protein families surprisingly good scoring functions can be obtained using simple ligand-only descriptors like log*S*, log*P*, and molecular weight. The ligand−descriptor based scoring function equals or even outperforms commonly used scoring functions, highlighting the need for better scoring functions. We demonstrate how the observed performance depends on the validation strategy, and we outline a general validation protocol for future free energy scoring functions.

INTRODUCTION

The free energy of binding is a pivotal criterion for compound selection and advancement in drug discovery and related areas. Free energies of binding are used to make assumptions about efficacy, selectivity, metabolism, and toxicity of drug candidates. In theoretical systems of biology and personalized medicine, the binding free energy is important for predicting differences between individuals with their genotypes. Accordingly, much research has been dedicated to predicting the free energy of binding.[1] Until now, this has resulted in a number of different approaches, from classical docking and various scoring functions[2−5] to MM-GB/PBSA,[6] quantum mechanics-based methods,[7,8] free energy perturbation,[9−11] and thermodynamic integration.[12−14] Unfortunately, docking and scoring techniques are often only successful in predicting the correct geometry of interaction; a fact that has been documented extensively in the literature.[15−21] This comes as no surprise because at least the older commercial scoring functions were trained to give the correct geometry, not the correct free energy. Nevertheless, scoring functions slightly correlate with free energy of binding, making them useful for early enrichment in virtual screening.[22] Among more complex approaches, MM-GB/PBSA is one of the cheapest sampling methods for which success has been demonstrated in the scientific literature.[12,23,24] However, the computational complexity of MM-GB/PBSA makes large scale applications very costly, and to the best of our knowledge, there is no study available for a set as large and diverse as the PDBbind09-CN (henceforth called PDBbind09) refined set[25] of different protein families among each other. Thermodynamic integration and other computationally very expensive methods should give very good, physically sound predictions of free energies. However, their complexity permits calculation for larger data sets, and usually they only yield relative predictions within chemical series.

While many publications in the past have been dedicated to describing scoring functions that perform well on predicting the correct geometries, there are only a few studies that deal with predicting the absolute free energy of binding of diverse data sets.[26,27] In this study, we specifically focus on the scoring problem, i.e., predicting the correct free energy of binding. We decouple the pose finding problem from the energy prediction problem by using the PDBbind09 database,[21] which provides access to correct geometries of protein−ligand interaction and measured binding constants. This approach is scientifically useful because it reduces the complexity of the docking/scoring

Received: December 3, 2010
Published: February 22, 2011

problem. Once it is feasible to predict the free energies of binding for known complexes, it should be straightforward to (I) rank different poses according to the highest energy predicted and (II) extend the methodology to hypothetical weak complexes with unfavorable interactions.

Scoring Functions Overview. Scoring functions can be divided into three major groups:[28] force-field based scoring functions,[29] knowledge-based scoring functions,[30−32] and empirical scoring functions.[33−40] Force-field based scoring functions use standard force-field terms to calculate an enthalpic energy of interaction for each pose. In some cases, desolvation effects are added via implicit solvent models. However, these energies usually neither correlate with the best pose nor with the free energy of binding, and therefore, they have become lately out of fashion. Knowledge-based scoring functions are based on a statistical analysis of protein−ligand crystal structures. The most prominent approach is to convert distance-dependent frequencies of atom−atom pairs into probabilities and scores. Empirical scoring functions are based on collections of force-field terms and terms representing entropic contributions (e.g., solvent accessible surface area, polar surface area, and amino-acid dependent contributions to entropies). Depending on the target, empirical scoring functions and knowledge-based scoring functions perform differently, and there is no single superior scoring function.[15]

Challenges in Parameter Fitting. The parameters in knowledge-based scoring functions are directly derived from crystal structures. Empirical scoring functions in contrast are fitted to a set of observed crystal structures. Because the parameters of both methods are depending on the underlying data set, they require validation with an independent data set to estimate the true predictive quality. Fitting the parameters of empirical scoring functions that contain angle- and distance-dependent terms is highly complex because they are interconnected. In addition to the fitting algorithm, the training data set is of huge importance to the quality of the scoring function. In the QSAR community, it is well-known that the quality of the model steadily increases with the number of samples used for training.[41]

QSAR Type Fitting. In the last 40 years, the domain of QSAR as a ligand-based drug discovery approach has evolved in parallel to the structure-based docking/scoring approaches.[42] In QSAR models, all relevant properties are encoded in descriptors that are analyzed and summarized in a model. In the QSAR community, a lot of knowledge has accumulated about model fitting, i.e., fitting algorithms, overtraining, and model validation.

Wikberg et al. coined the term of proteochemometrics for approaches that use both protein and ligand for affinity prediction.[43] In this sense, all global QSAR-type scoring functions are proteochemometric approaches. In the past years, a number of publications describe the use of protein−ligand descriptors to derive models of free energy of binding.

In 2004, Embrechts and co-workers published a scoring function based on distance-binned counts of 17 different Sybyl atom types.[44] They have used sets of 61 and 105 complexes to train the models. On the independent validation sets of 6 and 10 complexes, their models reached maximum R^2 values of 0.6 and 0.64, respectively.

In 2006, Tropsha et al. published a model based on ENTess descriptors, which are counts of protein−ligand atom quadruplets, and the electronegativity in the case of F, Cl, Br, I, P, and metals.[45] They have used 240 complexes with druglike compounds to train a k-nearest neighbor (kNN) model and 24 randomly selected compounds for independent validation. With their best model on the independent validation sets, they get values of R^2 up to 0.83. For the best model out of the 4510 models built with different numbers of nearest neighbors, descriptors included and specific test/training set splits, they get a q^2 of 0.66 on the training set.

In 2008, Klebe and co-workers published SFCScore, a PLS model based on ligand, protein, and ligand−protein interaction descriptors calculated for 855 partially proprietary complexes collected in the scoring function consortium.[40] On the subset of the PDBbind04 database ($N = 919$), which has not been part of the training set, SFCscore reaches $R^2 = 0.29$ ($R_{\mathrm{pearson}} = 0.54$).

In 2008, Artemenko published a scoring function based on MLR or a neural network and distance-dependent pair counts of AMBER atom types within the protein and the ligand. On the 10-fold independent crossvalidation of the training set consisting of 288 complexes, the scoring function reaches an RMSE of 1.79 log K units.[46]

In 2010, Ballester and Mitchell published RF-Score, a random forest model based on sums of protein−ligand element pair counts in a radius of 12 Å around each ligand atom.[47] This model has been built and evaluated based on the PDBbind07 database, containing 1300 protein−ligand complexes. In the out-of-bag validation, the model reaches $R^2 = 0.49$ and RMSE = 1.52. On the validation set that has been selected in a kind of space-filling manner,[21] the model reaches $R^2 = 0.60$ and RMSE = 1.58. However, we have recently shown that this model strongly benefits from the clustering of protein families. In leave-cluster-out crossvalidation on the whole PDBbind, RFscore reaches RMSE = 1.59 and an average $R^2 = 0.21$ within the families, and $R^2 = 0.29$ overall.[48]

In 2010, Breneman and co-workers published a scoring function based on property-encoded shape distribution (PESD) descriptors and support-vector machines.[49] Simulating different scenarios, they used 278 and 977 training/independent validation samples from the PDBbind05 database, where in order to increase the size of the data set both K_i and K_d values have been used. On the independent validation sets of 977 and 278 samples, they achieve model performances of $R^2 = 0.33$ ($R_{\mathrm{pearson}} = 0.57$), MUE = 1.36, RMSE = 1.74, and $R^2 = 0.41$ ($R_{\mathrm{pearson}} = 0.64$), MUE = 1.45, RMSE = 1.86, respectively. This study is similar to our study considering data set size and validation strategies.

In this publication, we introduce a new descriptor set that describes protein−ligand interactions. The set allows fitting local and global scoring functions for protein−ligand docking using QSAR methodologies. The QSAR approach allows directly making use of the large number of protein−ligand crystal structures complemented with K_d data that are published in databases such as PDBbind[50] or BindingMOAD.[51] With our models, we show that scoring functions are able to model at least half the free energy of the binding ($R^2 = 0.48$) of a diverse set of protein−ligand complexes.

Validation Strategies. It is usually very hard to compare different scoring functions according to their correlation with the free energy of binding for three reasons:

1.) Sampling algorithm: Different docking programs use different sampling algorithms, and the performance of some scoring functions strongly depends on the exact geometry used. If data sets with decoys of unknown geometry are used for the docking study, different docking programs might use different poses, of which it is not clear which should be the best pose. Here, we avoid the sampling problem by directly using the crystal structure.

2.) Training/validation set: Most of the scoring functions have been developed on the basis of publicly available data. However, it is often impossible to identify the subsets used for training and validation. For correct comparisons, however, the same validation set has to be used for every scoring function.

Ideally, the validation set is very large and includes all available information. Therefore, independent crossvalidation or bagging has to be used. Because it is not clear in all cases which complex has been part of which fitting procedure and we cannot refit individual scoring functions on the basis of subsets of all data available, it is very hard to compare existing scoring functions to one another. Ideally, the publisher of every new scoring function clearly states which complexes have been used for fitting and which crossvalidation or bagging strategy has been used to make results comparable.

3.) Family fitting effects: In principle there are two different scenarios when applying docking/scoring methods to a project. In scenario one, protein−ligand crystal structure and binding data for complexes from the same protein family are available. In this case, the docking/scoring functions can be tailored to a specific protein family. Here, one would either generate family specific local models or global models with local information available in the training set. The corresponding validation for both cases would be either crossvalidation or bagging. Scenario two is the situation where no local information is available. The corresponding validation strategy is leave-cluster-out validation.[48] Here, all members from one specific protein family have to be left out from the training and be used for validation. The second validation approach eliminates family specific parametrization, which cannot be done for completely new protein targets. We have described this family effect in a previous publication.[48] To our knowledge, only Breneman and co-workers have recently applied a methodology similar to the leave-cluster-out crossvalidation for validating their scoring function on the basis of PESD descriptors.[49] We recommend applying this method routinely for assessing the power of novel scoring functions.

METHODS

Descriptors. Atom types for each ligand atom are assigned using RDkit[52] and SMARTS[53] strings that code for atom types similar to the Crippen atom typing scheme[54] (one extra carbon atom type, eleven extra hydrogen atom types, two extra nitrogen atom types, and one extra oxygen atom type; SMARTS assignment is given in the Supporting Information). The Crippen atom typing scheme has been very useful for QSPR models, for example, for the prediction of distribution coefficients, but to our knowledge it has never before been used for scoring functions. The atom typing scheme applied distinguishes 29 atom types for carbon, 17 atom types for nitrogen, 14 atom types for oxygen, 16 atom types for hydrogen, 3 atom types for sulfur, and one atom type each for F, P, Cl, Br, and I. Overall there are 84 atom types for the ligand atoms. For each standard protein atom, the atom type has been determined using the same SMARTS definitions. For the proteins, additional atom types are Ca, Mg, Mn, Zn, and H_2O, giving overall 39 atom types for the protein atoms. Because only approximately half of the ligand atom types occur in natural amino acids, there are fewer atom types for proteins than for ligands. Table 1 shows exemplary the SMARTS patterns for all carbon atom types, including the log*P* increments and the assignment of donor/acceptor/neutral. The priority of the SMARTS definitions goes from top to bottom, i.e., a carbon atom not fitting any SMARTS pattern will get the atom type CS.

Table 1. Atom Typing SMARTS List for Carbon[a]

ID	SMARTS	log*P* incr	don/acc/neu
C1	[CH4]	0.1441	N
C1	[CH3]C	0.1441	N
C1	[CH2](C)C	0.1441	N
C2	[CH](C)(C)C	0	N
C2	[C](C)(C)(C)C	0	N
C2	[CH3][N,O,S,F,Cl,Br,I]	−0.2035	N
C3	[CH2 × 4][N,O,S,F,Cl,Br,I]	−0.2035	N
C28*	[C](=[O,N])[O,N]	−0.2051	N
C4	[CH1 × 4][N,O,S,F,Cl,Br,I]	−0.2051	N
C4	[CH0 × 4][N,O,S,F,Cl,Br,I]	−0.2051	N
C5	[C]=[!C;A;!#1]	−0.2783	N
C6	[C;A]=C	0.1551	N
C7	[CX2]#[A;!#1]	0.0017	N
C8	[CH3]c	0.08452	N
C9	[CH3]a	−0.1444	N
C10	[CH2 × 4]a	−0.0516	N
C11	[CHX4]a	0.1193	N
C12	[CH0 × 4]a	−0.0967	N
C13	[cH0]-[A;!C;!N;!O;!S;!F;!Cl;!Br;!I;!H]	−0.5443	N
C14	[c][#9]	0	N
C15	[c][#17]	0.245	N
C16	[c][#35]	0.198	N
C17	[c][#53]	0	N
C18	[cH]	0.1581	N
C19	[c](:a)(:a):a	0.2955	N
C20	[c](:a)(:a)−a	0.2713	N
C21	[c](:a)(:a)−C	0.136	N
C22	[c](:a)(:a)−N	0.4619	N
C23	[c](:a)(:a)−O	0.5437	N
C24	[c](:a)(:a)−S	0.1893	N
C25	[c](:a)(:a)=[C,N,O]	−0.8186	N
C26	[C](=C)(a)[A;!#1]	0.264	N
C26	[C](=C)(c)a	0.264	N
C26	[C](=C)a	0.264	N
C26	[C]=c	0.264	N
C27	[CX4][A;!C;!N;!O;!S;!F;!Cl;!Br;!I;!#1]	0.2148	N
CS	[#6]	0.08129	N

[a] The full SMARTS definition for all elements can be found in the Supporting Information. The star(*) denotes the atom types different from the standard Crippen types.

Distance-dependent protein ligand atom type pairs (ddPLATp) have been calculated for upper bin thresholds of 3.0, 3.5, 4.0, 4.5, 6.0, and 12.0 Å, ignoring all pairs with a distance larger than 12 Å. We chose the upper threshold of 12 Å because Ballester and Mitchell have recently shown that they get surprisingly good scoring functions with 12 Å as upper cutoff for counting element pairs.[47] We chose smaller bin sizes for close interactions because we expected differences for direct interactions to be larger in close proximities of

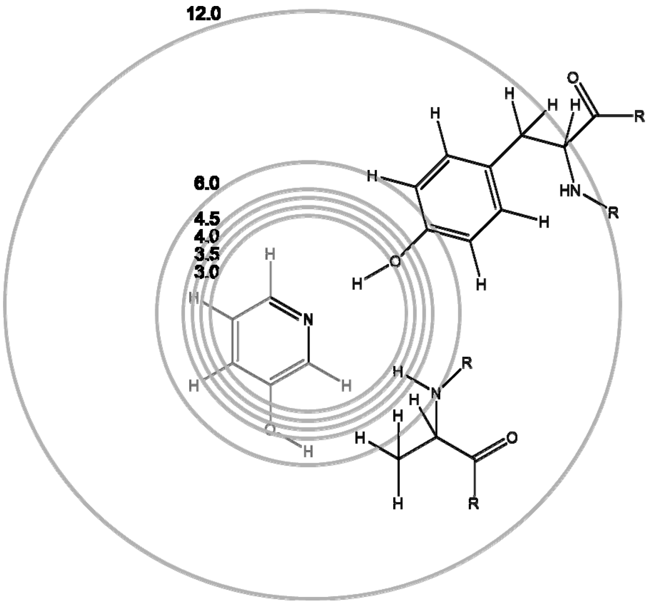


Figure 1. Distance-binning scheme for protein—ligand atom type pair descriptors.

the interacting atoms (as one would expect from classic Lennard—Jones potentials and electrostatic potentials). Binning has the advantage over distance-dependent functions that the parameters can directly be determined in the fitting procedure, while for standard distance-dependent functions, a functional has to be fit in a much less straightforward way. We overall chose six bins as a trade-off between resolution and overall number of descriptors. Each descriptor counts the occurrence of a certain ligand atom type/protein atom type pair with a distance between two bin thresholds. This results in 84(ligand) × 39(protein) × 6(number of bins) = 19656 descriptors. As a further descriptor, all ligand atoms of a certain type and all protein atoms in the binding site (distance to any ligand atom <4 Å) of a certain type have been counted. That results in 84 + 39 = 123 descriptors. In order to compare this approach with a simpler one, we calculated distance-dependent protein—ligand element pairs (ddPLEp) omitting hydrogen. Here, we summed all ligand atom—protein atom pairs for different elements within the boundaries of the thresholds given above. This results in 9 × 10 × 6 = 540 descriptors. Many of the descriptors generated are sparse, which means that they have entries different from 0 on few samples only. All descriptors with less than 5 entries different from zero were removed prior to model building. This reduced the number of atom-type pair-based descriptors by approximately 50%. The descriptor calculation for distance-dependent atom type pairs is schematically shown in Figure 1.

Additionally, we calculated atom and bond counts, the log*S*/log*P* and Lipinski violations[55] available within the molecular operating environment (MOE).[56] The difference between conformational energy of the bound conformation minus the conformational energy of the relaxed conformation for each ligand according to the OPLS2001 force field[57] is used as a further descriptor to capture the effect of ligand strain on binding.

Lastly, we compare the binding energies measured to the docking scores (score in place) from Glide (Glide XP),[38] GOLD[2] (ChemScore,[58] GoldScore, ASP[37]), and XScore (HMScore, HPScore, HSScore).[59] ASP and Chemscore did not work on two samples where the ligand is very small, and there seem to be no H-bond donors in the binding site (1e4h, 1y4z). Where possible, a local optimization has been enabled for the commercial scoring functions. Note that these scoring functions, especially XScore, were built on the basis of part of the crystal structures present in the PDBbind09 database, so validation with these descriptors is not completely independent.

Complex Selection/Exclusion and Preparation. For our calculations, we use the PDBbind09-CN refined set[21] with 1741 entries. We removed all complexes with ligands of a molecular weight larger than 900 (mostly polypeptides), with more than 20 donors and acceptors (mostly polyglycosides) and those with more than one P atom (mostly NADPH and ATP/ADP). The protein complexes were prepared using Schrodinger's protein preparation wizard[60] with all options switched on, i.e., adding hydrogens, assigning disulfide bonds, removing waters further away than 5 Å from the ligand, adjusting charges, capping termini, adding missing side-chains using prime, optimizing H-bond clusters, and doing a short minimization using OPLS2005[57] on the whole protein. Cases where the protein preparation wizard produced an error were individually inspected and either corrected or dropped. All ligands as given in the SD-Files in the PDBbind09 database were additionally read by RDkit,[52] and cases where RDkit reported an error were either corrected or removed. This left an overall number of 1387 complexes for the models.

We decided not to use the PDBbind07 core set[21] for validation because this gives highly overoptimistic performance estimates, as we recently showed.[48]

Mathematical Methods. We use our own implementation of bagged stepwise multiple linear regression (MLR) with the correct F-value that takes into account the size of the descriptor pool available for selection.[61] This approach has two mechanisms to avoid overfitting: First, it stops including descriptors if they do not give better predictions than 95% of random descriptors would give. Second, it uses bagging and has an individual test set for each linear regression sample, thus offering an unbiased estimate of the true model predictivity. We use 50 independent bagging samples and randomly selected 25% of the data set as test samples for each model. The advantage of this approach is a well adjusted, conservative descriptor selection with very few descriptors in the final model, which makes it easily interpretable.

As a second approach, we used partial least-squares (PLS)[62] with 7-fold crossvalidation without descriptor selection. We used the "one-standard-error" rule described by Hastie et al.[63] to determine the final number of latent components, i.e., we picked the most parsimonious model within one standard error of the minimum. Such a rule acknowledges the fact that the trade-off curve is estimated with error and, hence, takes a conservative approach. Compared to stepwise MLR, this approach has the advantage that it also includes sparse descriptors that have different values from 0 on few samples only. For the bagging MLR, it is very unlikely that a highly sparse descriptor improves the overall performance so much that it passes the F-test because it can only act on few examples, but its significance is measured by its impact on all the samples. The PLS is calculated in *R*[64] using the plsgenomics library.[65] In all inspected cases, the MLR model with descriptor selection was better than the PLS model in terms of R^2 on the validation set. Additionally, the MLR model is easier to interpret because it contains fewer descriptors. We therefore will not present results for the PLS model here.

We used three different validation scenarios to simulate various data availability situations. The first scenario is the standard cross-validation/bagging scenario, where random subsets are left out of the training procedure and predicted afterward. This corresponds to a situation where some local information, i.e., related crystal complexes, are available for the complexes to be predicted.

The second scenario is the leave-cluster-out scenario. Here, all members of a specific protein family are completely left out of the training. Then the whole family is predicted on the basis of a model trained on biochemically unrelated data. This scenario corresponds to the starting situation of a drug discovery project where no information about closely related compounds and targets is available.

In the third scenario, we train and validate purely local models, i.e., models that contain members from one family only, in both training and validation set. Here, we present the results for the out-of-bag test set predictions.

We prefer to use the described modeling strategies over leaving out a completely independent validation set for several reasons. First, there is no such thing as a completely different publicly available validation set. Second and more importantly, the predictive power of small independent test sets is most likely not more significant than the crossvalidation or the out-of-bag test set. Because we do not use descriptor selection combined with crossvalidation, there is no source for traditional overfitting. The MLR models based on descriptor selection are validated by out-of-bag validation, where the out-of-bag set is never involved in descriptor selection. The only effect that could be seen from an external test or validation set is how the model performs on a completely different set of proteins from different families. This is modeled in the LCOcv scenario.

We intentionally do not use a nonlinear method such as random forests or support-vector machines. Although it would technically be possible, this would make the models less interpretable.

Measures of Quality. We use the root mean squared error (RMSE), the mean unsigned error (MUE), and the predictive R^2 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. For evaluation of the commercial scoring functions, we use the square of Pearson's correlation coefficient R (R^2_{pearson}). RMSE and MUE measure the absolute accuracy of the prediction, i.e., how well the experimental value is reproduced by the model. R^2_{pearson} assesses how well models predict the relative order within a data set. A model that has a high RMSE/MUE (bad) and a high R^2_{pearson} (good) can still be useful to prioritise compounds in virtual screening. The predictive R^2 in contrast assesses how close the predictions are to the absolute target value. Parallel shifted predictions get a low predictive R^2.

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

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

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

$$R^2_{\text{pearson}} = \frac{\sum_{i=1}^{n} (y_{i,\,\text{meas}} - \overline{y_{\text{meas}}})(y_{i,\,\text{pred}} - \overline{y_{\text{pred}}})}{\sqrt{\sum_{i=1}^{n} (y_{i,\,\text{meas}} - \overline{y_{\text{meas}}})^2} \sqrt{\sum_{i=1}^{n} (y_{i,\,\text{pred}} - \overline{y_{\text{pred}}})^2}}$$

Table 2. Correlation of Common Scoring Functions with the Activities from the PDBbind09 Database

	ASP	ChemScore	GOLDScore	GlideXP	HMScore	HPScore	HSScore
R^2_{pearson}	0.253	0.250	0.181	0.100	0.362	0.374	0.366

Table 3. Bagged MLR Results for the Whole Set

	bagged MLR		
descriptor set	R^2	RMSE	MUE
ddPLATp	**0.45**	**1.47**	**1.18**
ddPLEp	0.41	1.52	1.22
MOE	0.34	1.62	1.29
ddPLATp + MOEcounts	**0.48**	**1.44**	**1.14**

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

RESULTS

Scores of GOLD (Chemscore, ASP, GoldScore), GlideXP and XScore (HMScore, HPScore and HSScore) have been calculated using scoring in place. The correlation coefficients of scores calculated with the K_d's measured are summarized in Table 2.

It does not make sense to calculate an RMSE or an MUE here because most of the scoring functions do not intend to exactly predict the free energy of binding and only give a value for ranking. Note that some of the scoring functions, especially all the XScore scoring functions (HMScore, HPScore, HSScore), have been parametrized on part of the PDBbind09 data set, so the correlation coefficients are probably overoptimistic.

Scoring functions based on each single descriptor set have been calculated for the whole data set of 1387 complexes. The results are summarized in Table 3.

The MLR model using the distance-dependent atom type pair count descriptor set yields a model with R^2 = 0.45 and RMSE = 1.47. The MLR model using the distance-dependent element pair count descriptors (which do not distinguish between different atom types) gives slightly worse predictions with R^2 = 0.41 and RMSE = 1.52. The MLR model based on standard MOE descriptors for the ligand, omitting the protein completely, gives a model with R^2 = 0.34 and RMSE = 1.62. This model is mainly based on molecular weight, log*S*, and Slog*P*.

Additionally, we tested the performance of adding the ddPLATp and the MOE descriptor set (without dbPLEp because this would introduce redundancies) to the best performing set. With the combined set and proper validation, R^2 = 0.48, RMSE = 1.44, and MUE = 1.14 is reached on the set of 1387 complexes from the PDBbind database. To our knowledge, this is better than anything else published before for a global scoring function rigorously validated on such a large and diverse data set. A plot of predicted versus experimental values is shown in Figure 2.

We have recently shown that leave-cluster-out crossvalidation (LCOcv) is necessary for scoring functions derived on diverse collections on protein−ligand complexes and fitted using flexible nonlinear algorithms such as random forests. The effect of LCOcv on the assessment of the predictive quality of the scoring functions has been tested. The results of the MLR model based on the ddPLATb + MOE descriptors validated using LCOcv, standard bagging, and standard bagging based on family information only (local models) are summarized in Table 4. On average, 12 descriptors enter each single MLR model based on the global data set, and 1 to 3 descriptors enter each single MLR model based on the local descriptors.

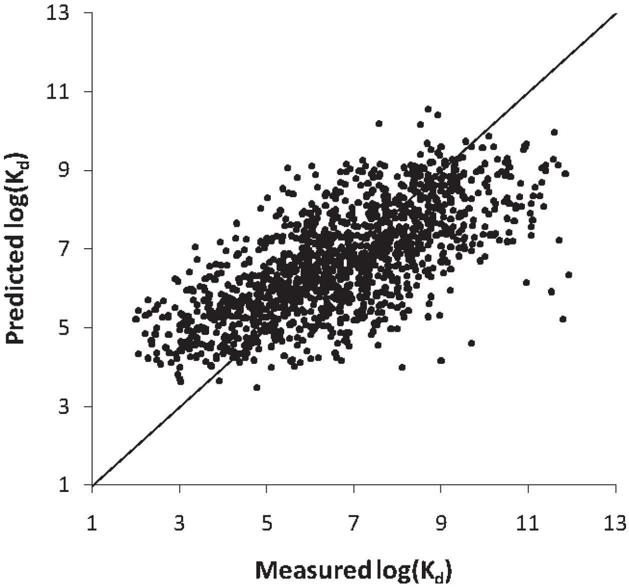


Figure 2. Predicted versus experimental values, ddPLATb + MOE descriptors model.

The overall RMSE of the LCOcv is 1.57, while the overall RMSE of the standard bagging approach is 1.44. The R^2 reduces from 0.47 to 0.38 when going from bagging to LCOcv. In the bagging model, 27% of the data are predicted within 0.5 log K_d, 51% within 1.0 log K_d and 71% within 1.5 log K_d. In the LCOcv model, 26% of the data are predicted within 0.5 log K_d, 46% within 1.0 log K_d and 66% within 1.5 log K_d. The overall performance of the local models is in between the two approaches. However, for most of the local models, there is no correlation. The overall statistics of the local models look good because the variance within the individual families is lower than the overall variance, and most complexes are predicted with the average activity within the family. The average R^2 in the clusters for both LCOcv and bagging is lower than the overall R^2 because the variance of the overall set is larger than the variance of each single set. A plot of the correlation coefficients and RMSEs obtained with LCOcv versus the correlation coefficients and MUEs obtained with bagging is shown in Figure 3.

RMSE, R^2, and MUE for the individual groups are similar, with the exception of thrombin and factor Xa. Here the predictions from the bagging model are better than the predictions from the LCO model, and the values for R^2 decrease from 0.39 to 0.29 and

Table 4. Results for Each Protein Family after Leave-Cluster-Out Crossvalidation, Standard Bagging, and for Local Models

		leave-cluster-out CV			bagging			local		
protein family	*N*	R^2_{pearson}	RMSE	MUE	R^2	RMSE	MUE	R^2	RMSE	MUE
HIV protease	159	0.02	1.61	1.26	0.17	1.38	1.09	0.21	1.35	1.07
trypsin	69	0.47	1.04	0.84	0.58	0.94	0.76	0.59	0.90	0.74
carbonic anhydrase	50	0.39	1.52	1.32	0.48	1.40	1.23	0.08	1.82	1.52
thrombin	56	0.29	2.25	1.97	0.39	1.23	0.95	0.31	1.27	0.97
PTP1B (protein tyrosine phosphatase)	30	0.71	1.10	0.94	0.57	0.96	0.80	0.55	0.70	0.55
factor Xa	32	0.12	1.88	1.55	0.27	1.27	1.01	−0.07	1.21	1.05
urokinase	28	0.38	1.41	1.12	0.30	1.33	1.07	0.28	1.22	1.04
different similar transporters	15	0.26	0.96	0.82	0.22	0.90	0.75	−0.19	1.12	0.85
c-AMP dependent kinase (PKA)	17	0.29	1.44	1.17	0.26	1.43	1.15	−0.08	1.34	1.20
beta-glucosidase	17	0.23	1.15	0.97	0.14	0.87	0.77	−0.12	0.84	0.67
antibodies	8	0.56	1.81	1.31	0.76	1.59	1.19	−0.41	2.16	1.76
casein kinase II	15	0.33	0.83	0.68	0.39	0.73	0.56	0.14	0.62	0.51
ribonuclease	8	0.02	0.97	0.83	0.25	0.82	0.63	−0.32	0.56	0.52
thermolysin	14	0.57	1.00	0.68	0.45	1.08	0.79	−0.16	1.54	1.24
CDK2 kinase	10	0.71	1.20	1.00	0.80	1.17	0.99	−0.26	1.74	1.53
glutamate receptor 2	13	0.02	1.25	1.11	0.09	1.27	1.11	−0.20	0.86	0.69
P38 kinase	12	0.50	0.72	0.56	0.43	0.73	0.60	−0.21	1.03	0.95
beta-secretase 1	10	0.93	1.41	1.18	0.93	1.34	1.10	0.72	1.43	1.04
tRNA-guanine transglycosylase	12	0.00	1.15	0.96	0.05	1.22	0.97	−0.18	1.06	0.85
endothiapepsin	5	0.95	1.37	1.22	0.97	1.48	1.24	−0.56	1.82	1.47
alpha-mannosidase 2	10	0.05	1.66	1.30	0.13	1.75	1.37	−0.32	1.67	1.38
carboxypeptidase A	10	0.24	2.21	1.58	0.12	2.20	1.64	−0.03	2.06	1.47
penicillopepsin	10	0.39	1.58	1.40	0.28	1.60	1.44	−0.17	1.93	1.65
all clusters with 4−9 complexes	293	0.37	1.56	1.25	0.43	1.48	1.20	0.28	1.66	1.32
all clusters with 2−3 complexes	265	0.38	1.52	1.23	0.40	1.49	1.22	0.34	1.56	1.25
singletons	219	0.24	1.75	1.42	0.25	1.73	1.39	0.15	1.83	1.50
overall	1387	0.38	1.57	1.25	0.48	1.44	1.14	0.41	1.53	1.21

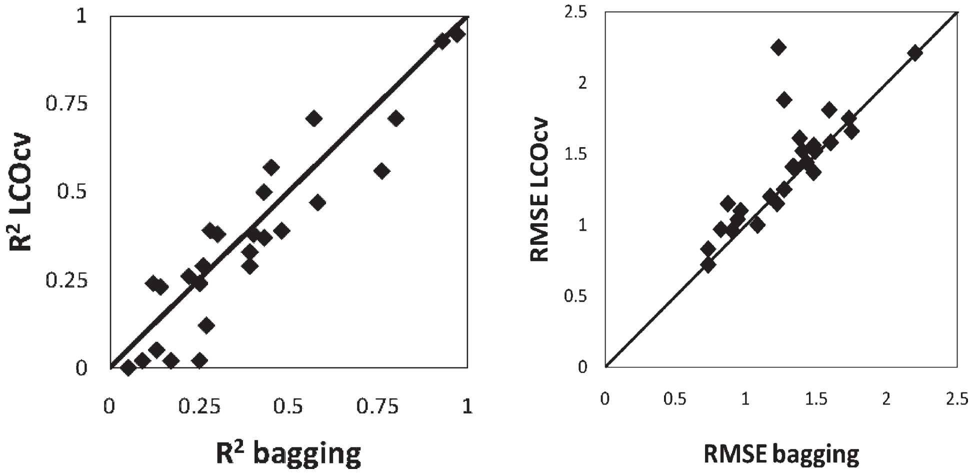


Figure 3. Correlation coefficients and RMSEs obtained with bagging versus correlation coefficients and RMSEs obtained with LCOcv.

from 0.27 to 0.12. In all other families, the performance remains more or less the same. The performances of the scoring functions on the local sets vary strongly. While the performance on the HIV protease family is rather bad, the performance of the local scoring functions on the second largest family, trypsin, is rather good with R^2 = 0.59. Here the solubility (log$S$), a single descriptor from MOE, is responsible for nearly all the correlation. The third set, carbonic anhydrase, is an example where the predictions of the local model (R^2 = 0.08) are worse than the predictions from a global model (R^2 = 0.39/0.48).

For most families, the predictions from the global model are better than predictions obtained from a local model, when measured in terms of R^2. That actually is the observation one hopes for: generalization and insight from data outside of the family domain. R^2 is the measure of choice here because the MUE and the RMSE of local models mainly depend on the variance of the test set, while R^2 measures the correlation that can be used in virtual screening. The predictions for groups containing families with two to nine members with the LCO validation strategy are better than the test set prediction obtained from a local model, although the local model training set contains members from the same protein family. This illustrates the effect of the balance between increasing training set size versus increasing number of similar samples in the training set. The difference in prediction performance on all singletons directly illustrates the effect of different training set size because for them there are no similar proteins in the training set, but the training set size of the LCO validation is five to six times larger than the training set size of the local model.

The fourteen most frequently occurring descriptors within all the 50 MLR models for the combined set are listed in Table 5.

The 10 largest errors of the predictions for the stepwise MLR model with all descriptors (effectively 10−15 descriptors used in each MLR model) are shown in Figure 4.

None of the largest outliers (maybe besides the ligand in 6CPA) is druglike. Most of them are natural products, metabolites thereof, or closely related compounds. Most of the complexes with the largest prediction error have either very high or very low measured K_d values. Zinc and iron each occur in three of the strongest outliers. This might be due to the fact that crystal structures with natural products are biased toward zinc- and iron-containing binding sites. The same applies for phosphates and phosphonates. Natural product binding can be very specific and supported by whole protein entropic factors, where hydrogens far away from the binding site are stabilized upon binding.[66] In the protein data bank,[67] the ligand of 2VC7 is a protonated tautomer of the ligand shown in Figure 4. We decided to do the calculations on the basis of the tautomer shown in Figure 4 because this chemically makes more sense, and it is the tautomer in the PDBbind database.

DISCUSSION

We have presented a method to predict the free energy of binding with novel global scoring functions for protein−ligand complexes. The best scoring function presented has a correlation coefficient R^2 of 0.48, an RMSE of 1.44, and a MUE of 1.14 on the independent test set. This is much better than the correlation of commercially available scoring functions with the PDBbind09 refined set and, to the best of our knowledge, better than any other scoring function published and stringently evaluated on the whole PDBbind09 database.

The scoring function is fitted like a QSAR model: All molecular interactions are coded in descriptors, and these are subjected to a QSAR fitting procedure. This procedure has the advantage that all the QSAR fitting methodology developed in the last 40 years can directly be applied to the scoring problem. Because a lot of descriptors are being considered (up to 10.099 in the largest model), the models have to safely be guarded against overtraining. In our modeling approach, this is ensured by a specific strategy: For each single model of the bagged stepwise MLR approach with descriptor selection 25% of the data set are completely left out of the training. After the training of each single model the corresponding hold-out test set is predicted. This is done 50 times, and the hold-out test set predictions for all samples are averaged. Thus, it is ensured that a test set sample is never involved in descriptor selection and therefore independent.

We have introduced a new set of descriptors on the basis of the Crippen-like atom type pairs. Using distance-dependent atom type pair count descriptors, a scoring function is obtained that has R^2 = 0.45, RMSE = 1.47, and MUE = 1.18 on the bagging test set in a MLR model. This set has been compared to a simpler set of descriptors on the basis of distance-dependent element pair counts. This model has a significantly (P < 0.05 according to a t test on the absolute errors) worse performance with R^2 of 0.41, RMSE of 1.52, and MUE of 1.22, showing that the atom type differentiation brings more relevant details into the model.

Table 5. Most Important Descriptors[a]

descriptor	occurrence (out of 50)	coefficient	meaning
Ptype#.H1	50	+++	number of H1 hydrogens (aliphatic) in the binding pocket
weight	47	+++	molecular weight
LO3.PH3.6.0	46	−−	number of ligand−O3 (ether oxygen): protein H3 (hydrogen connected to aromatic carbon) atom pairs between 4.5 and 6.0 Å
Ptype#.C18	37	+++	number of C18 carbons (aromatic carbon carrying one hydrogen) in the binding pocket
LCl.PC23.4.0	36	++	number of ligand−chlorine: protein C23 (tyrosine aromatic carbon to hydroxy oxygen) atom pairs between 3.5 and 4.0 Å
b_rotN	36	−−−	number of rotatable bonds in the ligand
LC25.PN10.3.5	29	++	number of ligand−C25 (aromatic carbon connected to C,N,O): protein N10 (charged histidine guanidinium nitrogen) atom pairs between 3.0 and 3.5 Å
LC24.PN16.6.0	28	++	number of ligand−C24 (aromatic carbon connected to nonaromatic sulfur): protein N16 (tryptophan aromatic nitrogen) atom pairs between 4.5 and 6.0 Å
LN1.PH6.6.0	26	++	number of ligand−N1 (aliphatic uncharged nitrogen carrying at least one hydrogen): protein H6 (tyrosine phenoxy hydrogen) atom pairs between 4.5 and 6.0 Å
LO4.PN2.3.5	19	+	number of ligand−O4(ether oxygen with at least one aromatic neighbor): protein−N2 (backbone-nitrogen) atom pairs between 3.0 and 3.5 Å
LN2.PH10.6.0	14	+	number of ligand−N2(uncharged nitrogen connected to nonaromatic heavy atoms and one hydrogen): proteinH10 (amide hydrogen) atom pairs between 4.5 and 6.0 Å
LC2.PH10.4.0	13	+	number of ligand−C2 (tertiary or quarternary aliphatic carbon): protein H10 (amide hydrogen) atom pairs between 3.5 and 4.0 Å
LH13.PC28.3.0	13	+	number of ligand−H13 (hydrogen on noncharged aromatic nitrogen): protein C28 (amide or carboxy carbon) atom pairs up to 3.0 Å
LC23.PH10.3.5	12	+	number of ligand−C23 (aromatic carbon connected to hydroxy oxygen): protein H10 (amide hydrogen) atom pairs between 3.0 and 3.5 Å

[a] The coefficient column indicates whether the descriptor has a positive or a negative impact on the free energy observed (++±--: abs (standardized coefficient) > 0.3, +±-: abs (standardized coefficient) 0.1−0.3, ±: standardized coefficient <0.1.

Despite its simplicity, the latter model is still better in predicting free energies of binding on the PDBbind database than all commercially available scoring functions tested, and it is, to the best of our knowledge, among the best scoring functions published. Thus, we suggest considering this function as a lower benchmark that any new and more complex scoring function has to beat.

The best model in terms of performance measures is obtained on the basis of all descriptors with R^2 = 0.48, RMSE = 1.44, and MUE = 1.14 and consists on average of 12 descriptors per submodel. Nearly half of the variance in binding energy can be explained using the presented model. Taking into account that the data set is very heterogeneous with data from different laboratories, measured using different assays with different buffers at different temperatures and pH values, the upper limit of performance is probably much lower than R^2 = 1.0.

Analysis of the Ligand-Only Model. Interestingly, a not-too-bad model can also be obtained using simple ligand-based MOE descriptors only. These give R^2 = 0.34, RMSE = 1.62, and MUE = 1.29. The most important descriptors from this set are molecular weight, Slog*P* and log*S*. Previously, it has been shown a couple of times that the free energy of binding is indeed slightly correlated with the three descriptors.[68−70] This observation makes sense because the free energy of binding is derived from the difference between the free energy in the bound and in the unbound state, and changing the energy of the unbound state (i.e., the free energy in solution log*S*) affects the free energy of binding. Log*P* and molecular weight are also correlated with the solubility.

We tested the performance of the MOE descriptors on the local models. For six out of the 23 protein families, models with correlation coefficients R^2 > 0.25 can be obtained using one simple MOE descriptor only. These results are summarized in Table 6.

These are peculiar correlations, but in the case of trypsin, PTP1B, and beta-secretase1, they are clearly significant (lower confidence interval level R^2 > 0.3). The descriptors clearly do not represent the full physical basis behind the binding event. Nevertheless, they are significantly correlated with the free energy of binding, and the correlations suggest that there might be very simple explanations for the differences in binding affinities. Maybe the most important message from this observation is that these data sets taken on their own cannot help to distinguish between good and bad scoring functions because any scoring function can probably be trained to work on these data sets in isolation.

The local models trained on the whole descriptor set are worse than the models trained on the ligand-only descriptors because there are so many descriptors from which to to choose. For data sets with around 10 samples, no reasonable models can be generated on the basis of such a large set of descriptors because the probability that one out of ∼10.000 random descriptors is highly correlated with the target data is so high that "real" correlations are noised-out, and most of the submodels only consist of a constant.

The families where the MOE scoring function works best are a subgroup of the families where the best scoring function presented works best in terms of R^2. There are some families where none of the scoring functions works well. Among others, these include HIV-protease, beta-glucosidase, and glutamate receptor

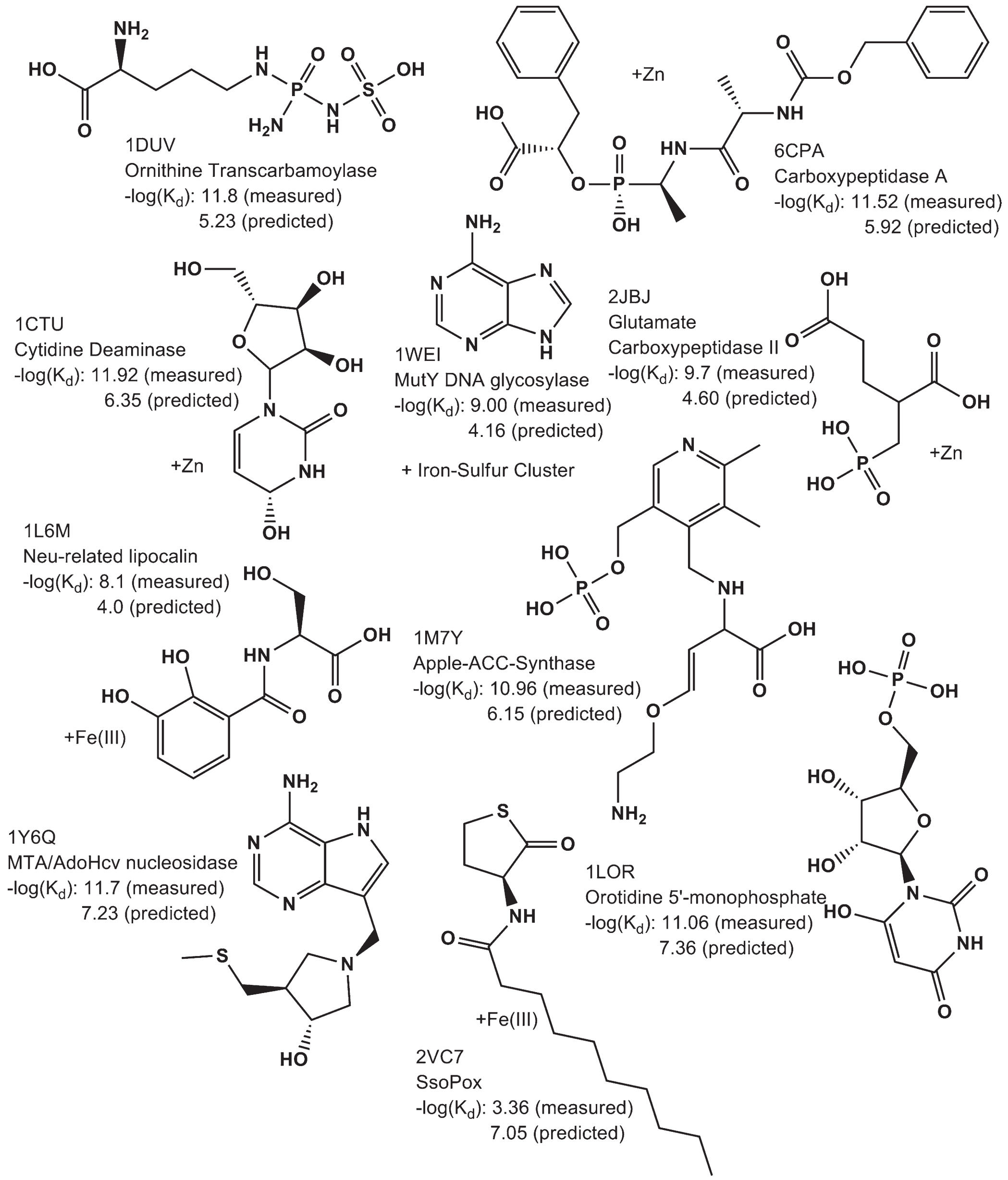


Figure 4. Strongest outliers.

2. We plan to investigate these failures further in the future, but going into the details of the problems on these receptors is beyond the scope of this manuscript.

The MOE model is significantly worse than the best model presented, considering the performances on the whole data set (upper 99% confidence interval limit[71] of the MOE model, $R^2 = 0.39$; lower 99% confidence interval limit of the best model, $R^2 = 0.43$). Adding the information contained in the distance-dependent descriptors clearly contributes to the prediction quality. However, the MOE model is as good as or better than all the commercial models we have tested. This is probably due to the facts that this data set includes ligands with molecular weight up to 900 g/mol and commercial scoring functions are trained to predict correct geometries and not necessarily the free energies of binding.

Descriptor Interpretation. The descriptors we have introduced here are straightforward to interpret: The number of protein aliphatic hydrogens in the binding pocket is the most important descriptor. This and the fourth most important descriptor, number of protein aromatic carbons in the binding pocket, stand for hydrophobicity of the binding pocket. Molecular weight is the second most important descriptor, and the most important descriptor of the MOE only model. The third most important descriptor is the number of pairs of ligand ether oxygen atoms, with protein hydrogens connected to aromatic carbons. Ether oxygen placed above an aromatic ring gets a high count in this descriptor. The interaction of ether oxygens with aromatic rings contributes negatively to the binding affinity, probably because the hydrogen bond to water molecules is stronger and the aromatic rings disrupt the water network around ethers and ether-containing substructures. This descriptor has the highest counts for sugar ligands, which contain at least two ethers. The meaning of the descriptor can be rationalized in the following way: It has been shown that in aqueous solution sugars

Table 6. Performance of the MOE Descriptor Set on Local Models with $R^2 > 0.25^a$

family	N	main descriptor	R^2	RMSE	lower 95% confidence level	upper 95% confidence level
trypsin	69	log$S$	0.51	0.99	0.35	0.67
PTP1B	30	log$S$	0.61	0.65	0.41	0.81
urokinase	28	weight	0.45	1.06	0.20	0.70
thermolysin	14	b_rotN	0.40	1.12	0.06	0.74
CDK2 kinase	10	weight	0.29	1.31	−0.10	0.68
beta-secretase 1	10	weight	0.89	0.88	0.78	0.99

[a] Results given for the out-of-bag test set.

are surrounded by a hydration shell that has stable coordination of water molecules up to the third shell,[72] indicating a high energy gain from hydration. Aromatic systems, on the other hand, are prototypical for aromatic and hydrophobic interactions, which are opposite to interactions with complex networks of water molecules. The sixth most important descriptor, number of rotatable bonds, also has a negative sign. This might account for the loss of entropy upon binding. The other most important descriptors describe specific interactions. LCl.PC23:4.0, for example, is the count of all pairs of chlorine in the ligand and C23 carbon (aromatic) in the protein in between 3.5 and 4.0 Å. This and many other specific interactions have been described in detail in a recent review by Bissantz et al.[73] The model presented here can complement the interactions described with average binding affinity contributions. However, most of the interactions have only a very small contribution to the overall binding affinity, and the most significant contributions found here are those based on hydrophobic interactions.

Only real crystal complexes have been used for fitting the scoring functions presented here. Complexes with very low affinity are not part of the PDBbind09 database. In other words, this data set is lacking negative examples or decoys. Thus, the scoring functions presented here and all other scoring functions parametrized solely on the PDBbind or a similar data set lack negative examples, and their predictions on complexes with unknown geometry have to be treated with caution. Nevertheless, the PDBbind data set represents a very useful set for fitting and testing scoring functions because presently it is the largest publicly available collection of diverse protein−ligand structures with activity data. As such, it is a prototype of a global validation set for scoring functions. Only a scoring function that performs well on the PDBbind database can be expected to perform reliably on a new diverse data set with unknown binding geometries.

Performance Dependence on Validation Strategy. The performance measured strongly depends on the kind of validation carried out. Here, we have presented three different kinds of validation: standard crossvalidation/bagging, leave-cluster-out crossvalidation, and "local model only". Each validation strategy simulates different data availability situations and gives (slightly) different results. The best performance measures are obtained using the standard crossvalidation/bagging approaches. Leave-cluster-out crossvalidation simulates a real drug discovery situation with a completely new target. In some cases (trypsin, factor Xa, beta-glucosidase), the RMSE and MUE obtained with LCOcv are worse. The correlation coefficients within the protein families, validated using LCOcv, are highly correlated with the results obtained with standard bagging (R^2 = 0.85). This shows that the predictions for some targets are parallel shifted and can still be used in applications where only the relative performance is important, such as in virtual screening. Local models based on the descriptors presented are worse than the global models in nearly all cases. Only for the two largest families, HIV-protease (159 samples) and trypsin (69 samples), the correlation coefficient of the local model is clearly better than the correlation coefficient obtained with LCOcv. There is probably a minimum data set size necessary to build local models. Already for carbonic anhydrase (50 samples), the correlation coefficient obtained with the global model (R^2 = 0.39) is better than the correlation coefficient obtained with the local model (R^2 = 0.08).

Largest Outliers. The largest errors of prediction are summarized in Figure 4. The good news for this scoring function is that none of the ligands (besides maybe the ligand of 6CPA) from the largest outliers are druglike. We have decided to apply very soft filters on removing compounds in order to get as much information as possible about the performance of this scoring function. We have only removed compounds with a molecular weight above 900 g/mol, with more than one P atom and with more than twenty donors + acceptors in order to remove highly redundant natural ligands like NADPH and ADP/ATP. Nevertheless, most of the largest outliers are natural products, metabolites, or closely related compounds. Most complexes with the largest prediction errors have either very high or very low K_d values. In eight out of nine cases with natural ligands, the binding energy is highly underestimated. It is not unlikely that evolution has tuned proteins that bind one specific natural ligand to recognize exactly this ligand and gain binding efficiency from entropic effects that only occur upon binding of the exact ligand. This has for example been shown for the biotin/streptavidin complex, which shows highly cooperative binding effects.[74,75] Complex structural entropic effects beyond hydrophobic/hydrophilic surfaces can probably not be mapped by the descriptors we introduced and need more complex calculations that sample the conformational space of the bound complex.

We further find that zinc and iron in the binding pocket and phosphorus in the ligand occur more frequently than would be expected in the 10 largest outliers. We have used a specific atom type each for zinc and iron in the protein and one atom type for phosphorus in the ligand. Either the description provided by the descriptors used is not detailed enough or phosphorus is frequently found in natural ligands that bind to specific proteins with complex entropic effects and zinc in the binding site.

Comparison with Other Proteochemometric Scoring Functions. In 2004, Embrechts et al. published a related method for deriving scoring functions.[44] They used distance-binned Sybyl atom type pairs in a QSAR-like fitting procedure to predict free energies of binding. With the kernel-PLS method, they got an R^2 of 0.45 without descriptor selection and an R^2 of nearly 0.60 after descriptor selection on the randomly selected external test set of 6 and 10 compounds, respectively. Unfortunately the

Table 7. Comparison of the Performance on the Independent Validation Set of Several Scoring Functions with Lower and Upper Confidence Interval Boundaries (95% Significance Level)

	validation set size	RMSE	R^2	95% confidence interval
Embrechts et al. 2004[44]	6; 10	not given	0.45; 0.60	[0.00,0.91]; [0.28,0.92]
Tropsha et al. 2006[45]	24	not given	0.66	[0.46,0.86]
Klebe et al. 2008[40]	919	1.80−1.89	0.29	[0.24,0.34]
Artemenko 2008[46]	288	1.79	not given	not possible to calculate
Breneman et al. 2010[49]	977; 278	1.76; 1.86	0.33; 0.41	[0.28,0.38]; [0.32,0.50]
this study	1378	1.44	0.48	[0.44,0.52]

external test set is very small: the 95% confidence interval[63] for N = 6 and R^2= 0.45 is from 0.0 to 0.91. The 95% confidence interval for N = 10 and R^2=0.60 is from 0.28 to 0.92. In the scoring function published here, we get R^2 = 0.48 on the 1387 test set complexes (95% confidence interval: 0.44 to 0.52).

In 2008, Artemenko has published a scoring function on the basis of neural networks and distance dependent protein−ligand AMBER atom type descriptors and some more physicochemical descriptors.[46] On the independent test sets, the best model reaches an RMSE of 1.79 (here: 1.44). The model, based on MOE descriptors only, that we have presented here reaches an RMSE of 1.62. The best model published by Artemenko has a significantly better performance (RMSE = 1.45) on the internal crossvalidation test set, which has been involved in descriptor selection. This illustrates the problems of overtraining associated with descriptor selection. The other reason for the difference on the performance might come from the training set size and the atom typing scheme.

In 2008, Sotriffer, Klebe, and co-workers published the SFCscore scoring function.[40] It is based on MLR and PLS models and descriptors selected out of a pool of 66 descriptors. They used ligand-only descriptors and descriptors describing the interaction between ligand and protein atoms, classified as hydrophobic, polar, and aromatic atoms for most cases. The training set is a collection of in-house structures from different pharmaceutical companies and 290 complexes selected from the PDB. All scoring functions generated are validated on a subset of the PDBbind04 database, which has not been part of the training set. On the independent validation set, the scoring functions reach a Pearson's correlation coefficient of 0.49−0.54, corresponding to a R^2_{pearson} between 0.24 and 0.29 and an RMSE of 1.80−1.89 on the first 800 complexes of the validation set. We tried to compare our scoring function directly to SFCscore. This was not straightforward because the PDBbind04-CN database was not online at the time we made the comparison. We therefore calculated the performance of our scoring function published here on all complexes of the PDBbind09 database that have been published before 2004 (N = 670). Our scoring function has an RMSE of 1.46 and a R^2 of 0.48 on that subset.

The significance of the performance measured strongly depends on the size of the validation set. A fair comparison of different approaches must take into account the size of the validation sets and assume that the validation sets have been chosen randomly. In Table 7, we list the performances on the independent validation set of different proteochemometric scoring functions published and the upper and lower 95% confidence levels for the performances.[71]

In terms of R^2, confidence levels, and RMSE on the independent validation set as listed in Table 7, the scoring function presented in this study is clearly among the best or even the best proteochemometric scoring function published so far. The two proteochemometric scoring functions published in 2004 and 2006 unfortunately have been validated on a very small test set only, which makes them nearly incomparable to other scoring functions. With 10 or 24 validation set compounds, the range of the confidence intervals is very wide (10: $\Delta R^2 \sim 0.64$; 24: $\Delta R^2 \sim 0.40$), and even with 278 compounds, there is still a range of $\Delta R^2 \sim 0.18$ between the lower and the upper confidence threshold. Table 7 shows that it is necessary to have large validation sets to be able to compare different models thoroughly.

The modeling approach presented here can easily be extended using different descriptor sets and an ever larger set of protein−ligand complexes with binding data. Because the distance dependent element pair counts are probably one of the simplest descriptor sets one can calculate, similar to atom counts for ligand-only models, we suggest the performance of this model (R^2 = 0.41, RMSE = 1.52, and MUE = 1.22) as a lower benchmark limit for scoring functions fitted to the PDBbind database. Any scoring function employing more complex terms like atom types or distance- and angle-dependent functions must outperform this scoring function according to Occam's razor. Commercially available scoring functions perform quite badly on this data set. However, when comparing commercial scoring functions, one must take into consideration that they are built to find the correct geometry of druglike molecules, which is different from predicting the free energy of a wide range of molecules from 78 to 900 g/mol, where the binding geometry is known.

We hope that this model is not the end but rather one step further toward new and fruitful approaches of fitting scoring functions. As mentioned above, only scoring functions that give good predictions of free energies on the PDBbind data set (which currently is the largest collection of crystal structures plus binding data) can be expected to give good predictions on other diverse data sets. Once we are able to predict the correct free energy, we should also be able to correctly distinguish between different docking poses.

Summary and Outlook. We have presented a scoring function for free energies that has been trained and evaluated on the basis of the PDBbind09-CN database. The scoring function is fitted in a QSAR-like manner with descriptors coding for specific interactions. We have used distance-binned Crippen-like atom type pair counts that are straightforward to interpret and have never before been used in scoring functions. Using a standard bagging/crossvalidation validation strategy gives a scoring function with R^2 = 0.48 (R_p = 0.69), RMSE = 1.44, and MUE = 1.14 for the PDBbind09 database. Using a leave-cluster-out validation strategy, the resulting scoring function has R^2 = 0.38, RMSE = 1.57, and MUE = 1.25 for the PDBbind09 database. When applying this scoring function on a new target, the performance expected is RMSE = 1.57. This can easily be improved by retraining the scoring function once data for the new target is available. The results of the standard bagging/crossvalidation are much better than the commercially available scoring functions we have tested and, to our knowledge, are as good as or even better than the best other scoring functions published and validated on the PDBbind database.

A couple of improvements are possible for this scoring function. These include descriptors representing angle- and distance-dependent terms, energy calculations based on a higher level of theory, and entropy terms obtained from MM-sampling. The scoring function probably does not perform well on ligands with more than one phosphorus atom or more than 20 donors and acceptors, since we removed these compounds from the training set in advance. The most severe limitation of this scoring function (and most other scoring functions), however, is that it has been trained on positive examples only, i.e., really existing complexes. The scoring function presented here might be able to detect ligands that do not make complexes, but it cannot be expected to do so. Including decoys in the training step poses additional scientifically interesting problems, which we want to investigate, but these are clearly beyond the scope of this paper. Generally, inclusion of decoys into the training phase is probably one of the most important improvements toward the practical usability of this and all other scoring functions based solely on the PDBbind database.

In its current state, the scoring function presented is useful for judging docking poses and for virtual screening, where low correlations with the real free energies of binding are sufficient. Additionally, it is useful for estimating an upper limit of binding free energy, which we will elaborate in an upcoming paper. There we also analyze the descriptors used here and simplifications thereof in depth and show how they can be used in drug design.

■ ASSOCIATED CONTENT

Supporting Information. File containing the SMARTS strings for Crippen typing. File containing the coefficients for the final model.This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: Christian.Kramer@novartis.com.

■ ACKNOWLEDGMENT

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

■ REFERENCES

(1) 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.

(2) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–748.

(3) Stroganov, O. V.; Novikov, F. N.; Stroylov, V. S.; Kulkov, V.; Chilov, G. G. Lead finder: An approach to improve accuracy of protein−ligand docking, binding energy estimation, and virtual screening. *J. Chem. Inf. Model* **2008**, *48*, 2371–2385.

(4) Hecht, D.; Fogel, G. B. Computational intelligence methods for docking scores. *Curr. Comput.-Aided Drug Des.* **2009**, *5*, 56–68.

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

(6) Gohlke, H.; Case, D. A. Converging free energy estimates: MM-PB(GB)SA studies on the protein−protein complex Ras-Raf. *J. Comput. Chem.* **2004**, *25*, 238–250.

(7) Raha, K.; Merz, K. M. Large-scale validation of a quantum mechanics based scoring function: Predicting the binding affinity and the binding mode of a diverse set of protein−ligand complexes. *J. Med. Chem.* **2005**, *48*, 4558–4575.

(8) Raha, K.; Peters, M. B.; Wang, B.; Yu, N.; Wollacott, A. M.; Westerhoff, L. M.; Merz, K. M., Jr. The role of quantum mechanics in structure-based drug design. *Drug Disc. Today* **2007**, *12*, 725–731.

(9) Jorgensen, W. L. Efficient drug lead discovery and optimization. *Acc. Chem. Res.* **2009**, *42*, 724–733.

(10) Knight, J. L.; Brooks, C. L. Lambda-dynamics free energy simulation methods. *J. Comput. Chem.* **2009**, *30*, 1692–1700.

(11) Marrone, T. J.; Briggs, J. M.; McCammon, J. A. Structure-based drug design: Computational advances. *Annu. Rev. Pharmacol. Toxicol.* **1997**, *37*, 71–90.

(12) Gilson, M. K.; Zhou, H. Calculation of protein−ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.

(13) Ytreberg, F. M.; Swendsen, R. H.; Zuckerman, D. M. Comparison of free energy methods for molecular systems. *J. Chem. Phys.* **2006**, *125*, 184114.

(14) Kim, R.; Skolnick, J. Assessment of programs for ligand binding affinity prediction. *J. Comput. Chem.* **2008**, *29*, 1316–1331.

(15) Warren, G. L.; Andrews, C. W.; Capelli, A.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; 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.

(16) 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.

(17) Coupez, B.; Lewis, R. A. Docking and scoring: Theoretically easy, practically impossible? *Curr. Med. Chem.* **2006**, *13*, 2995–3003.

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

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

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

(21) 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.

(22) Jansen, J. M.; Martin, E. J. Target-biased scoring approaches and expert systems in structure-based virtual screening. *Curr. Opin. Chem. Biol.* **2004**, *8*, 359–364.

(23) Li, Y.; Liu, Z.; Wang, R. Test MM-PB/SA on true conformational ensembles of protein−ligand complexes. *J. Chem. Inf. Model* **2010**, *50*, 1682–1692.

(24) Brown, S. P.; Muchmore, S. W. Large-scale application of high-throughput molecular mechanics with Poisson-Boltzmann surface area for routine physics-based scoring of protein−ligand complexes. *J. Med. Chem.* **2009**, *52*, 3159–3165.

(25) 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.

(26) Zhang, C.; Liu, S.; Zhu, Q.; Zhou, Y. A knowledge-based energy function for protein−ligand, protein−protein, and protein−DNA complexes. *J. Med. Chem.* **2005**, *48*, 2325–2335.

(27) Cozzini, P.; Fornabaio, M.; Marabotti, A.; Abraham, D. J.; Kellogg, G. E.; Mozzarelli, A. Simple, intuitive calculations of free energy of binding for protein−ligand complexes. 1. Models without explicit constrained water. *J. Med. Chem.* **2002**, *45*, 2469–2483.

(28) 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.

(29) Englebienne, P.; Moitessier, N. Docking ligands into flexible and solvated macromolecules. 5. Force-field-based prediction of binding affinities of ligands to proteins. *J. Chem. Inf. Model* **2009**, *49*, 2564–2571.

(30) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337–356.

(31) Velec, H. F. G.; Gohlke, H.; Klebe, G. DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.* **2005**, *48*, 6296–6303.

(32) Muegge, I. PMF scoring revisited. *J. Med. Chem.* **2006**, *49*, 5895–5902.

(33) Pham, T. A.; Jain, A. N. Customizing scoring functions for docking. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 269–286.

(34) Catana, C.; Stouten, P. F. W. Novel, customizable scoring functions, parameterized using N-PLS, for structure-based drug discovery. *J. Chem. Inf. Model* **2007**, *47*, 85–91.

(35) Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, B. S.; Johnson, A. P. eHiTS: An innovative approach to the docking and scoring function problems. *Curr. Protein Pept. Sci.* **2006**, *7*, 421–435.

(36) Moustakas, D. T.; Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijmans, N.; Rizzo, R. C. Development and validation of a modular, extensible docking program: DOCK 5. *J. Comput. Aided-Mol. Des.* **2006**, *20*, 601–619.

(37) Mooij, W. T. M.; Verdonk, M. L. General and targeted statistical potentials for protein−ligand interactions. *Proteins* **2005**, *61*, 272–287.

(38) 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.

(39) Hartmann, C.; Antes, I.; Lengauer, T. Docking and scoring with alternative side-chain conformations. *Proteins* **2009**, *74*, 712–726.

(40) Sotriffer, C. A.; Sanschagrin, P.; Matter, H.; Klebe, G. SFCscore: Scoring functions for affinity prediction of protein-ligand complexes. *Proteins* **2008**, *73*, 395–419.

(41) Gedeck, P.; Rohde, B.; Bartels, C. QSAR: How good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *J. Chem. Inf. Model* **2006**, *46*, 1924–1936.

(42) Gedeck, P.; Kramer, C.; Ertl, P. Computational Analysis of Structure−Activity Relationships. In *Progress in Medicinal Chemistry*; Lawton, G.; Witty, D. R.; Eds.; Elsevier: Amsterdam, The Netherlands, 2010; Vol. 49, pp 113−160.

(43) Lapinsh, M.; Prusis, P.; Gutcaits, A.; Lundstedt, T.; Wikberg, J. E. S. Development of proteo-chemometrics: A novel technology for the analysis of drug-receptor interactions. *Biochim. Biophys. Acta* **2001**, *1525*, 180–190.

(44) Deng, W.; Breneman, C.; Embrechts, M. J. Predicting protein−ligand binding affinities using novel geometrical descriptors and machine-learning methods. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 699–703.

(45) Zhang, S.; Golbraikh, A.; Tropsha, A. Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein−ligand interfaces. *J. Med. Chem.* **2006**, *49*, 2713–2724.

(46) Artemenko, N. Distance dependent scoring function for describing protein−ligand intermolecular interactions. *J. Chem. Inf. Model* **2008**, *48*, 569–574.

(47) 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.

(48) 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.

(49) Das, S.; Krein, M. P.; Breneman, C. M. Binding affinity prediction with property-encoded shape distribution signatures. *J. Chem. Inf. Model* **2010**, *50*, 298–308.

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

(51) Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (mother of all databases). *Proteins* **2005**, *60*, 333–340.

(52) *RDKit: Open-Source Cheminformatics*, version Q3-2010. http://www.rdkit.org.

(53) *SMARTS Theory Manual*; Daylight Chemical Information Systems, Inc: Santa Fe, NM. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed October 2010).

(54) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure−activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21–35.

(55) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3–26.

(56) *Molecular Operating Environment (MOE)*, version 2009.10; Chemical Computing Group: Montreal, Canada, 2010.

(57) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(58) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein−ligand docking using GOLD. *Proteins* **2003**, *52*, 609–623.

(59) Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11–26.

(60) *Maestro*, version 9; Schrodinger: NewYork, 2010.

(61) 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.

(62) Wold, H. In *Multivariate Analysis*; Academic Press: New York, 1966; pp 391−420.

(63) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*; Springer Series in Statistics; 2nd ed.; Springer: New York, 2009; section 7.10, p 244.

(64) *The R Project for Statistical Computing*, R, version 2.10, 2010. http://www.r-project.org.

(65) Boulesteix, A.; Lambert-Lacroix, S.; Peyre, J.; Strimmer, K. *plsgenomics R Package. PLS Analyses for Genomics*, 2010. http://www.r-project.org.

(66) Williams, D. H.; O'Brien, D. P.; Sandercock, A. M.; Stephens, E. Order Changes within receptor systems upon ligand binding: Receptor tightening/oligomerisation and the interpretation of binding parameters. *J. Mol. Biol.* **2004**, *340*, 373–383.

(67) *RCSB Protein Data Bank*. http://www.rcsb.org/ (accessed January 7, 2011).

(68) Abad-Zapatero, C.; Metz, J. T. Ligand efficiency indices as guideposts for drug discovery. *Drug Discov. Today* **2005**, *10*, 464–469.

(69) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004**, *9*, 430–431.

(70) Reynolds, C. H.; Tounge, B. A.; Bembenek, S. D. Ligand binding efficiency: Trends, physical basis, and implications. *J. Med. Chem.* **2008**, *51*, 2432–2438.

(71) Zar, J. H. *Biostatistical Analysis*; Prentice Hall International: Upper Saddle River, NJ, 2004; pp 43−45.

(72) Liu, Q.; Schmidt, R. K.; Teo, B.; Karplus, P. A.; Brady, J. W. Molecular dynamis studies on the hydration of α, α-trehalose. *J. Am. Chem. Soc.* **1997**, *119*, 7851–7862.

(73) Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal chemist's guide to molecular interactions. *J. Med. Chem.* **2010**, *53*, 5061–5084.

(74) Sano, T.; Cantor, C. R. Cooperative biotin binding by streptavidin. Electrophoretic behavior and subunit association of streptavidin in the presence of 6 M urea. *J. Biol. Chem.* **1990**, *265*, 3369–3373.

(75) Hyre, D. E. Cooperative hydrogen bond interactions in the streptavidin−biotin system. *Protein Sci.* **2006**, *15*, 459–467.

NOTE ADDED AFTER ASAP PUBLICATION

This paper was published ASAP on February 22, 2011, with an incorrect version of Figure 4. The corrected version was published ASAP on February 25, 2011.