Optimizing the Signal-to-Noise Ratio of Scoring Functions for Protein−Ligand Docking

Markus H. J. Seifert*

4SC AG, Am Klopferspitz 19a, D-82152 Planegg-Martinsried, Germany

Received September 17, 2007

Empirical scoring functions provide estimates of the free energy of protein−ligand binding in situations when atomic-scale simulations are intractable, for example, in virtual high-throughput screening. Currently, such scoring functions are often inaccurate, and further improvements are complicated by the lack of reliable training data, the complex interplay between scoring functions and docking algorithms, and an inconsistent statistical treatment of positive and negative training data. In comparison to various other performance measures of scoring functions, "analysis of variance" provides a well-behaved objective function for optimization, which focuses on the signal-to-noise ratio of ligand−decoy discrimination. In combination with a large database of ligands and decoys, an in situ optimization of scoring function parameters was able to generate improved, target-specific scoring functions for three different proteins of pharmaceutical interest: cyclin-dependent kinase 2, the estrogen receptor, and cyclooxygenase-2. Statistical analysis of the improvements observed in "receiver-operating characteristic" curves showed that the optimized scoring functions achieved a significantly (between $p < 0.0001$ and $p < 0.05$) higher enrichment of true ligands. A scaffold dependence of the resulting binding modes was observed, which is discussed in conjunction with the rigid receptor hypothesis commonly made in protein−ligand docking. In summary, the approach described here represents a well-adapted statistical method for setting up scoring functions.

INTRODUCTION

The in silico prediction of protein−ligand binding affinity represents one of the major challenges in the structure-based design of small molecules as bioactive compounds. Simulation methods promise to provide unbiased estimates of free binding energy albeit at high computational costs.[1−4] Therefore, structure-based design and in particular virtual high-throughput screening (vHTS) typically rely on empirical or knowledge-based scoring functions.[5] These scoring functions are derived from experimental data using statistical methods, for example, regression analysis, which is extensively used in quantitative structure−activity relationship (QSAR) studies.[6−8] Scoring functions typically compromise accuracy for speed, leading to low correlations with experimental data and a high rate of false positives during virtual screening, but nevertheless an enrichment of true ligands is achieved in the top end of ranking lists sorted by score.[9,10] Recently, progress has been made in generating targeted scoring functions: for example, Hetenyi et al. combined an established scoring function with a ligand-based QSAR model which predicted the affinities of bulky, flexible ligands with a correlation coefficient of $R^2 \sim 0.8$.[11] Antes et al. applied a "design of experiment" strategy to tune empirical scoring functions to kinases and ATPases.[12] Similarly, Salo et al.[13] and Andersson et al.[14] used experimental design strategies for optimizing protein−ligand docking parameters. Another very promising approach has been proposed by Pham and Jain, who used negative training data, that is, the knowledge not only of binding molecules but additionally of nonbinding molecules, to optimize the generic Surflex scoring function.[15] They used a mean squared error function for optimization that, however, treats ligands and decoys inconsistently. Furthermore, it is desirable to use not only random decoys but additionally decoy molecules, which are physically similar, but topologically different to the ligands.[16]

The main obstacles for developing improved scoring functions are as follows: (i) Conventional statistical modeling relies on consistent and extensive experimental data of protein−ligand binding affinity. The publicly available data, however, is mostly inconsistent with respect to assay formats and—as a result—the types of affinity measures; for example, both K_i and K_d values are present in the PDBbind database, but typically in a mutually exclusive manner.[17] Due to large experimental variations—for example, between different assay formats[18]—such data are less suitable for building theoretical models. (ii) A complex interplay exists between scoring functions and the protein−ligand docking process: for example, the scoring function and its parameters strongly influence the predicted binding mode of a ligand; therefore, changes in the parameters may have unexpected effects on binding modes and thereby on the final score value. (iii) A method for a consistent statistical treatment of positive and negative training data—as provided by sets of ligands and decoys—remains to be identified. (iv) Conventionally, improved scoring functions are analyzed with respect to the enrichment of true ligands, but less attention has been paid to a thorough assessment of its statistical significance.

Neglecting the actual affinity value and focusing on the discrimination between ligands and nonbinding decoy molecules alleviates the problem of experimental uncertainties.[19] In practical applications, for example, during the hit identification phase of drug discovery, a reliable discrimination between ligands and decoys offers basically the same benefits

* Phone: +49-89-700763-0. Fax: +49-89-700763-29. E-mail: markus.seifert@4sc.com.

10.1021/ci700345n CCC: $40.75 © 2008 American Chemical Society
Published on Web 02/23/2008

as those of binding affinity prediction. The step from affinity prediction to molecule classification, however, renders conventional linear regression analysis less suitable, and the question remains which method is best for analysis and optimization of scoring functions in this setting.

In principle, a scoring function produces a specific distribution of ligand scores, and another one for decoy scores. In order to identify the true ligands from a background of decoys by the help of a scoring function, these distributions have to be separated along the score axis. The underlying problem therefore is to optimize the scoring function such that these distributions either move away from each other or get sharper, both of which are able to provide a better separation. Conventional methods for achieving this, for example, linear discriminant analysis (LDA),[20,21] are not suitable due to the above-mentioned interdependence between scoring function parameters, binding modes of docked ligands, and the calculated score value. In particular, the close coupling between protein−ligand docking procedures and score calculation methods creates a different situation compared to standard QSAR, where the descriptors are independent of the fitted regression parameters. Therefore, docking and scoring cannot be treated separately, and an "in situ"—that is, within the framework of a given protein−ligand docking program—optimization is necessary. In this study, a suitable objective function for in situ optimization of scoring functions is identified and applied for improving—without a loss of generality—the well-known Böhm scoring function.[22,23] Thereby, targeted scoring functions are generated for three target proteins of pharmaceutical interest, namely, cyclin dependent kinase 2 (CDK2), estrogen receptor (ER), and cyclooxygenase 2 (COX2).

It is shown that an objective function derived from classical "analysis of variance" (ANOVA)[24,25] achieves a significant improvement in the enrichment of true ligands for all three targets under consideration. Notably, this result is obtained by using only a local optimization. This simple, but effective, concept for optimizing scoring functions provides a novel paradigm which favorably augments the established QSAR approach for setting up scoring functions in situations where, for example, experimental data are flawed by large measurement errors and complex systems—for example, protein−ligand docking programs—have to be optimized. Additionally, it is generally applicable to other problems where the separation of two or more ensembles has to be maximized.

EXPERIMENTAL SECTION

Identification of Suitable Objective Functions. The effect of various modifications of score distributions on a set of parameters which quantify the distance between distributions has been simulated using Octave:[26] ANOVA parameters η^2 and F,[25] the Kruskal−Wallis (KW) rank-sum,[27] the Kolmogoroff−Smirnov (KS) distance,[28,29] the maximally achievable Matthews correlations coefficient (maxMCC),[30] the Boltzmann-enhanced discrimination receiver-operating characteristic (BEDROC) using $\alpha = 160.9$,[31] the area under the accumulation curve (AUAC),[31] the enrichment factor for the top scoring molecules within a ranking list of 1% size of the full database (EF),[31] and the Kullback−Leibler divergence (KL).[32] For a detailed description of the parameters, please refer to Supporting Information S1. The octave simulation utilized several octave subroutines written by Kurt Hornik:[33] anova.m, kolmogorov_smirnov_test_2.m, and kruskal_wallis_test.m. The simulation comprised 10^6 decoy molecules and 10^3 ligands. A second simulation was performed with the same setup, but using 10^3 decoy molecules and 10^3 ligands (data not shown). For each of the molecules, a random score with a Gaussian distribution—without a loss of generality—was generated according to desired transformation of the score distribution.

Using these theoretical scores, a "receiver-operating characteristic" (ROC) curve[34] and all of the parameters mentioned above were calculated. This was repeated 20 times for each of the 50 points along the desired trajectory. Therefore, this procedure is able to assess the mean and the standard deviation of all the parameters along the trajectory. Four different trajectories were evaluated: (i) shifting the mean score of the ligands ($\bar{S}_L = 0 \rightarrow -6$) away from the decoys ($\bar{S}_D = 0$) with the standard deviation σ of both being constant ($\sigma_L = 1$, $\sigma_D = 2$), (ii) shifting away the mean score of one-half of the ligands ($\bar{S}_{L,1} = 0 \rightarrow -10$) while the other half and the decoys maintain a constant mean score and standard deviation ($\bar{S}_{L,2} = 0$, $\bar{S}_D = 0$, $\sigma_L = 1$, $\sigma_D = 2$), (iii) reducing the standard deviation of the scores of the decoy molecules ($\sigma_D = 10 \rightarrow 2$, $\bar{S}_D = 0$, $\bar{S}_L = -6$, $\sigma_L = 1$), and (iv) increasing the variance of the scores of the ligands ($\sigma_L = 0.2 \rightarrow 6$, $\bar{S}_L = 0$, $\bar{S}_D = 0$, $\sigma_D = 1$).

Generation of Targeted Scoring Functions. Targeted scoring functions were generated using ProPose[35,36] and—without a loss of generality—an extended version of the Böhm scoring function.[22,23] A cation−π interaction term was added to the original Böhm scoring function, and an ad hoc parameter for this term was set to −0.7. The geometry of the interaction is similar to the aromatic ring center−aromatic ring interaction which favors a positioning of the cation above the aromatic ring.[35] The other scoring function parameters were set to standard values (dimensionless):[35] hydrogen bonding (h_don - lone_pair) −4.7, electrostatic interaction (pos_charge - neg_charge) −2.0, aromatic interaction (aro_center - aro_ring) −0.7, hydrophobic interaction −0.1, atomic clashes 0.25, and close contacts −0.1. Therefore, seven parameters have to be optimized, resulting in a seven-dimensional search space. In order to be as close as possible to real-world applications, only the top-scoring pose was considered for calculating the final score of protein−ligand complex.

Targeted scoring functions were generated for cyclin dependent kinase 2 (PDB ID 1CKP), estrogen receptor (PDB ID 3ERT), and cyclooxygenase 2 (PDB ID 1CX2). The active sites of the target proteins were defined as described in the CCDC-Astex set of protein structures.[37] In the 1CKP structure, the glycol covering the active site was removed. PDB entry 1CKP does not contain the full structure of purvalanol B.[38] The full structure has been modeled; however, only the published part of the structure is given in the figures, in order to provide an indication of the original ligand position within the CDK2 active site. PrepD was used to create ProPose target description files for docking.[35] The ligand and decoy databases for the three targets were extracted from the "directory of useful decoys" (DUD).[16] Additionally, a set of 5000 random molecules from the 4SC in-house database was docked into the target proteins.[19]

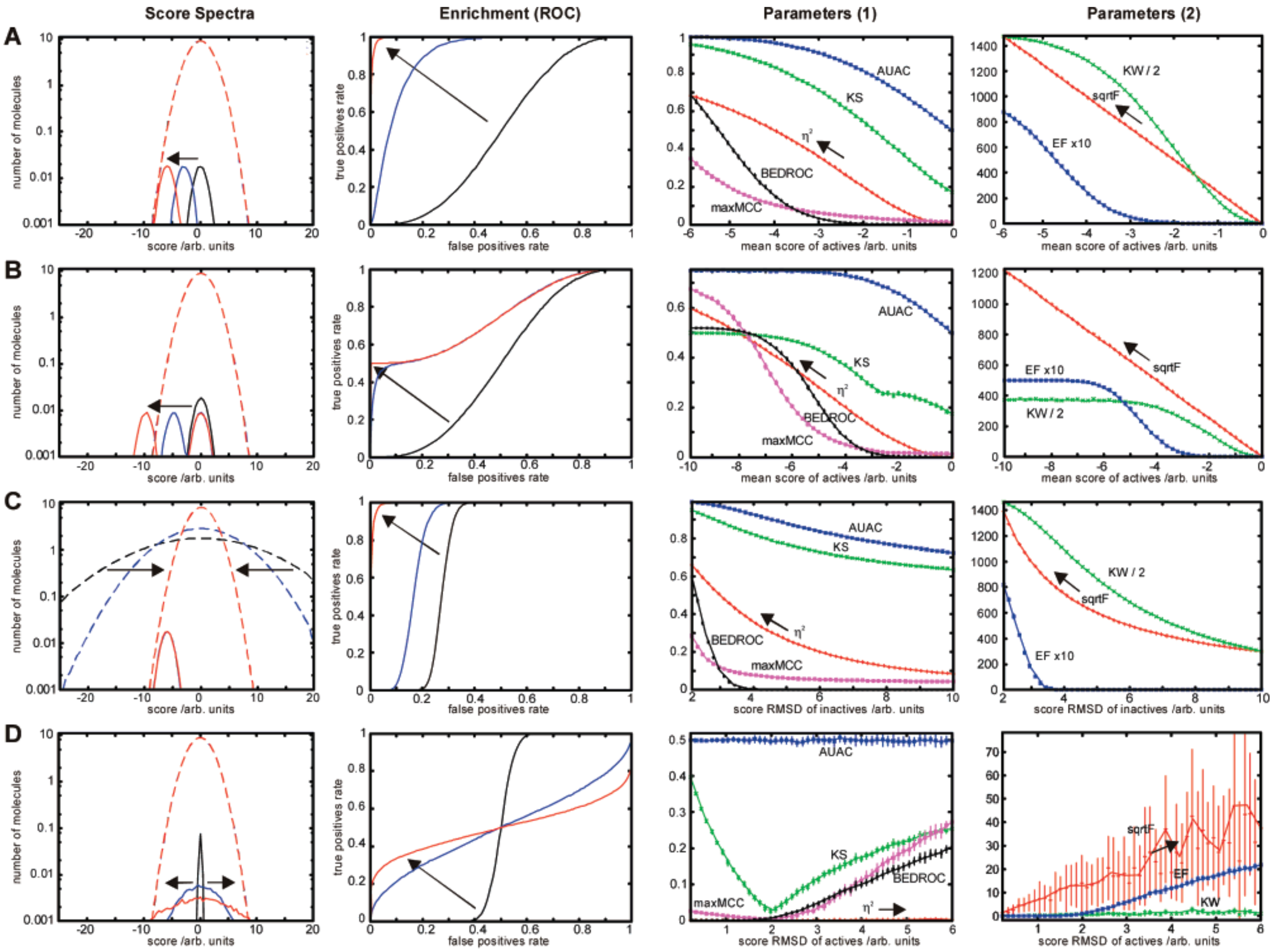

Figure 1. The effect of modifying the score distribution (score spectra) of ligands (solid lines) and decoys (dashed lines) on the enrichment of ligands in lists sorted by score (ROC) and on various statistical parameters. A more negative score indicates a higher predicted binding affinity. Instead of plotting F directly, the square root of F is depicted (sqrtF). (A) Shifting the mean score of ligands away from the mean score of decoys leads to a pronounced improvement in enrichment and an increase in all of the calculated parameters, which is detectable early on in some cases (η^2, F, AUAC, KS, and KW) or starts relatively late (BEDROC, maxMCC, and EF). KW, KS, and AUAC show a saturation behavior. (B) Moving the mean score of one-half of the ligands, while the other half and the decoys maintain a constant mean score, improves the enrichment to some extent, which is also reflected in the parameters. In this case, all parameters show a saturation behavior except the ANOVA parameters η^2 and F. (C) Lowering the standard deviation of decoy scores improves the enrichment, which is also visible in the parameters, but BEDROC, maxMCC, and EF start to raise rather late. (D) Increasing the standard deviation of ligand scores leads to an artificial enrichment since now some ligands are shifted to lower scores due to the increased noise in the score. BEDROC, maxMCC, and EF provide a gradient in the direction of larger standard deviation, whereas η^2, AUAC, and KW remain unaffected. KS shows a minimum at the point of equal standard deviation of ligand and decoy scores. F exhibits only a very weak gradient with respect to its noise level.

A taboo search algorithm (steepest ascent, mildest descent), which locally searches the seven-dimensional search space, was applied. The optimization algorithms searched the parameter space using relative step sizes (typically ±20%), thereby avoiding a change of sign and zero values. The algorithm comprised the following steps:

1. A random subset of active and inactive molecules—each of size ≤ 100—was selected from ligands and decoys (or alternatively from ligands and random molecules).

2. The two selected subsets of active and inactive molecules were docked into the target.

3. Initial ANOVA parameters F and η^2 were determined from the docking scores of active and inactive molecules. Since F and η^2 do not provide the direction of the difference of the mean values, they have to be multiplied by sign($\bar{S}_D - \bar{S}_L$).

4. The gradient of F in all $7 \times 2 = 14$ directions was determined by variation of the scoring functions parameters and performing 14 docking runs for active and inactive molecules with subsequent ANOVA calculations.

5. A step in the direction of the steepest ascent (if not possible, in the direction of the mildest descent) of F was taken, provided that the target state had not been visited before (taboo search). If that state had been visited before, the second-best state was chosen, and so forth. All visited states in parameter space were stored.

6. Steps 4 and 5 were repeated until a predefined improvement of F was found (typically $F > 70$ for the subsets of active and inactive molecules) or the maximum number of iterations was reached (typically 16).

7. Using the newly identified scoring function parameters, the full set of ligands, decoys, and random molecules was redocked and final ANOVA parameters were calculated.

For all of the targets, the algorithm was applied to optimize the discrimination between (a) ligands and decoys (LD) and (b) ligands and random molecules (LR) in independent runs.

In order to examine the long-term behavior of the optimization algorithm, the CDK2 LD optimization run was continued for 53 iterations.

Assessment of Statistical Significance. The ability to enrich ligands in the top end of a score-sorted list of molecules is commonly summarized in ROC curves, which plot the achieved true-positive rate (TPR) for any given false-positive rate (FPR).[39] The statistical evaluation of the ROC plots was performed by ROCKIT 0.9b and 1.1b.[40] ROCKIT calculates maximum-likelihood estimates of the parameters of a bivariate binormal model for data from two correlated diagnostic tests and thereby estimates the binormal ROC curves implied by those data and their correlation.[40] Additionally, it computes the statistical significance of the difference between two ROC curve estimates according to statistical tests. In this study, the area test—a univariate *z*-score test of the difference between the areas under the two ROC curves—and the TFP test—a univariate *z*-score test of the difference between the true positive fractions (TPFs) on the two ROC curves at a selected false-positive fraction—were applied. Since ROCKIT is not able to handle more than ~10 000 data points per condition, the number of the COX2 decoy molecules was reduced for this analysis from 12 464 to 9295 by random selection.

Sufficiently large external data sets for validation purposes are difficult to find, but for CDK2, such a data set is available. Therefore, external validation was performed using the CDK2 data set of Bradley et al., which contains 17 550 compounds including 306 compounds active on CDK2.[41] The molecules were ionized using Sybyl,[42] and 3D structures were generated by Corina.[43] All compounds with undefined stereochemistry were removed. Compounds of activity bin “50”, that is, with questionable activity, were discarded. In order to allow for the assessment of statistical significance, only scaffolds with more than 10 active compounds were considered (scaffolds 0, 5, 6, and 10). Scaffold 6, however, had no defined stereochemistry and was therefore neglected. Ligands were defined from the set of active molecules (activity bin “100”). Decoys were defined from the inactive molecules resulting from similarity searches and synthetic approaches (“simscreen” and “synscreen”). Background (random) molecules were defined from the inactive molecules derived from a general screening library (“divscreen”). This procedure resulted in an external validation data set containing in total 132 ligands, 771 decoys, and 8383 pseudo-random background molecules. These molecules were docked into PDB structure 1CKP using the initial and the optimized sets of parameters. The results were analyzed using ROCKIT.

RESULTS AND DISCUSSION

Objective Function for Optimization. In order to distinguish between true ligands and decoy molecules by the help of protein—ligand docking, it is necessary that the score distributions of ligands and decoys—as calculated by the scoring function used for docking—are separated as far as possible on the score axis. There are various methods to quantify such a separation (see Supporting Information S1). The optimal quantification method and therefore best objective function for optimizing scoring functions was identified by a survey of various statistical parameters (Figure 1). Four different modifications of the score distributions were examined, and their effect on the enrichment of ligands at the top end of score-sorted lists and on several statistical parameters is evaluated. Without going into the details of statistics (see Supporting Information S1), the following observations were made: (i) EF, BEDROC, and maxMCC respond to shifts of the mean scores relatively late (Figure 1A—C). This is detrimental in particular for optimization runs which start from zero discrimination between ligands and decoys. Additionally, they are prone to providing a gradient in the direction of a large variance of the ligand scores (Figure 1D). Such an increased variance may be caused, for example, by one ligand scaffold shifting to lower scores and another one shifting to higher scores, with no net effect on the mean value of the scores. This is a different situation compared to Figure 1B, where a net effect is indeed present. A significantly different mean value of the ligand scores, however, is the only reliable indication that the scoring function is able to provide a model that is applicable to all known ligand scaffolds. Scoring functions that show a preference for specific scaffolds may be useful—to some extent—in practical applications of virtual screening, but their overall performance is likely to be unstable. Therefore “early recognition” metrics like EF and BEDROC are not suitable as objective functions for optimizing scoring functions. (ii) KS shows a nonmonotonous behavior, rendering it unsuitable (Figure 1D). (iii) AUAC and KW get saturated relatively early when only one-half of the ligands shifts (Figure 1B). Due to the transformation of scores into ranks, which is the basis of all nonparametric methods including KW, BEDROC, AUAC, and EF, the gradient for optimization will drop as soon as the top ranks get populated with some ligands. (iv) The ANOVA parameters *F* and η^2, in contrast, are monotonous and exhibit a large dynamic range with only a weak tendency—relative to the noise level—of *F* to provide a gradient in the direction of potential artifacts. Notably, these parameters work even for non-normally distributed data (Figure 1B). Double-peaked, that is, non-normally distributed, data preclude testing the statistical significance by comparison to the *F* distribution. This is, however, not intended here, and hence, the calculation of *F* and η^2 is possible for any kind of distribution. (v) KL has been discarded in this

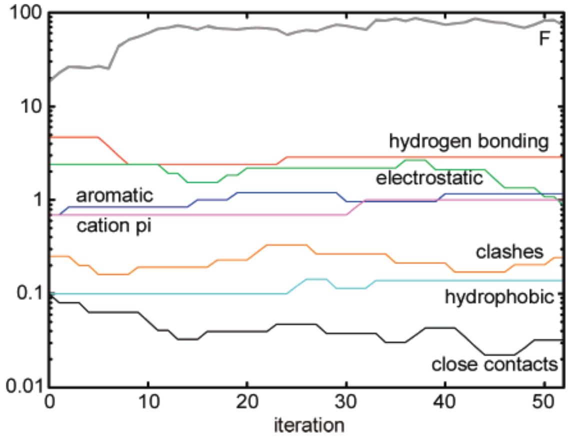

Figure 2. The value of the objective function *F* and the absolute values of the scoring function parameters in course of the CDK2 optimization run for ligand—decoy discrimination. Downscaling of the hydrogen-bonding parameter, the clash penalties, and the weight of close contacts early on in the optimization run resulted in a pronounced increase of *F*.

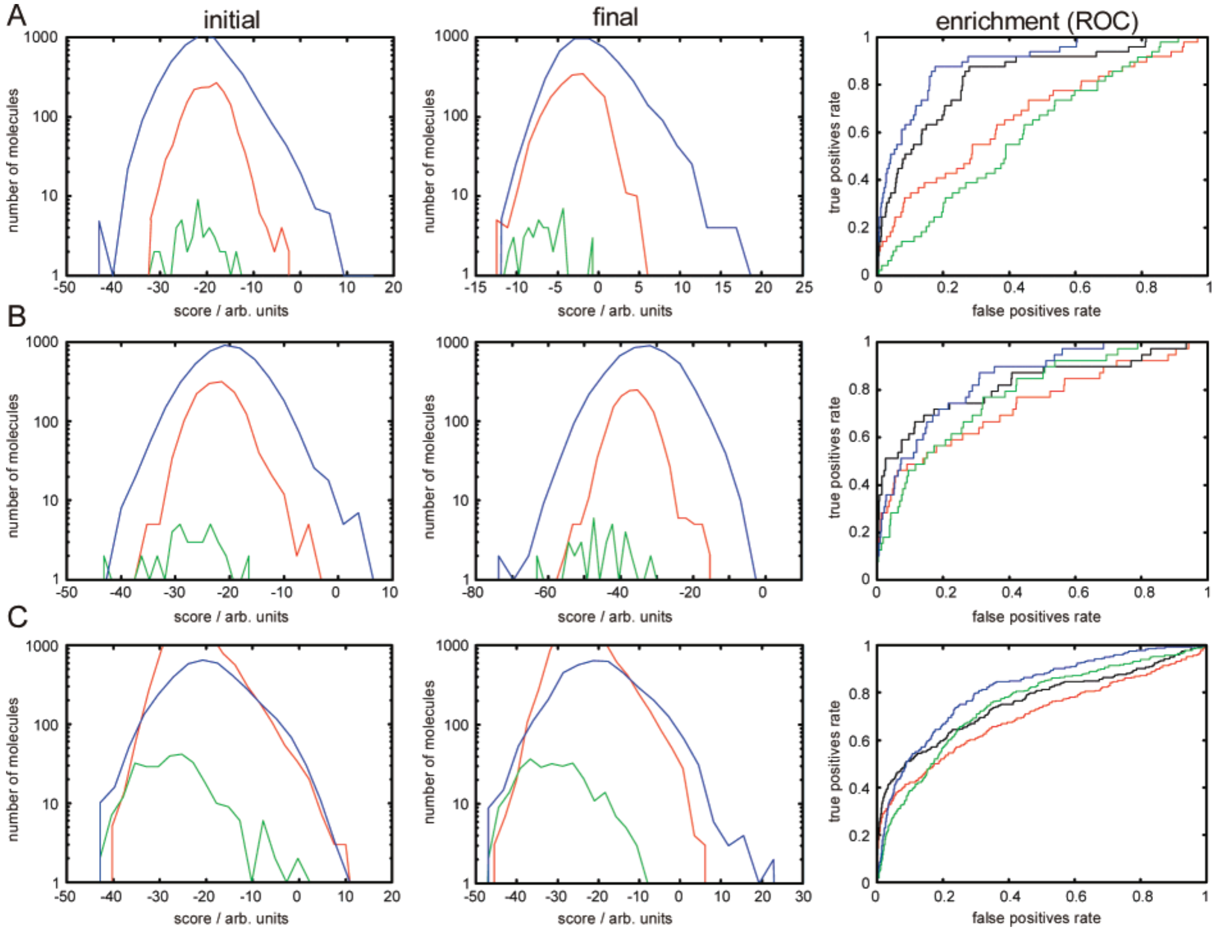

Figure 3. Initial score spectra (left column), final score spectra (middle column), and enrichment curves (right column) resulting from scoring function optimization for CDK2 (A), ER (B), and COX2 (C). The score spectra depict the score distribution of ligands (green), decoys (red), and random molecules (blue). The enrichment curves—here shown as receiver-operating curves (ROC)—of initial (red) and final (black) ligand−decoy discrimination and initial (green) and final (blue) ligand−random-molecule discrimination are plotted. These curves were derived from the ligand−decoy optimization runs.

survey due to the irregularities of the logarithm that occur for empirical score distributions containing nulls.

An objective function for optimizing the separation of ligand and decoy score distributions has to provide a large dynamic range, that is, a strong gradient over the full range of possible score distributions.[44] The best candidates fulfilling these criteria are the ANOVA parameters F and η^2, and to some extent AUAC. Due to their sound statistical foundation, the ANOVA parameters, and F in particular, were chosen as the objective functions for optimizing scoring functions in protein−ligand docking. In principle, η^2 measures the discriminatory power of the scoring function, whereas F is related to the signal-to-noise ratio of the scoring function.[19,45] F has the additional advantage that its value increases with the number of observations, that is, the number of successful docking calculations. This exerts an additional pressure to dock as many molecules as possible, thereby preventing the optimization routine to discard inopportune molecules, which would lead to a strong bias of the results.

Generation of Targeted Scoring Functions. Using the signal-to-noise ratio F as an objective function, the Böhm scoring function[22,23]—without a loss of generality—was optimized independently for three docking targets of pharmaceutical interest: CDK2, ER, and COX2. Böhm's scoring function is prototypical for an empirical scoring function of the form Score $= \sum a_i P_i$, with scoring function parameters a_i and corresponding properties P_i of the best binding mode identified by protein−ligand docking. It has been selected for this investigation due to its low complexity, which limits the dimensionality of the parameter space. The optimization was carried out in situ—that is, within the framework of ProPose[35]—by continuously redocking and scoring a subset of ligands and decoys. Two independent optimization runs were performed for each of the targets, one optimizing LD and the other one optimizing LR discrimination. A representative course of the parameters during optimization is shown in Figure 2. This optimization run involved 95 608 docking calculations in total. The local search in the parameter space quickly identified new sets of scoring function parameters with an improved signal-to-noise ratio F for docking into CDK2. The overall results of all of the LD optimization runs are summarized in Table 1. Similar improvements were obtained using optimization of LR discrimination (data not shown). The scoring functions parameters resulting from these optimization runs, however, were different from those cited in Table 1, indicating the presence of multiple local maxima of F.

As expected, the resulting scoring function parameters are highly specific for the respective target. Due to the purely empirical approach, the parameters are not transferable to any other target, scoring function, or docking program. The optimization method itself, however, can be applied to any

Table 1. Summary of Target-Specific Optimization Runs for Ligand−Decoy Discrimination

parameter	initial	CDK2	ER	COX2
hydrogen bonds	−4.70	−2.89	−3.76	−5.64
electrostatic	−2.00	−2.65	−0.82	−0.82
aromatic	−0.70	−0.97	−0.56	−0.84
cation π	−0.70	−1.01	−0.81	−0.56
hydrophobic	−0.10	−0.14	−0.10	−0.06
clashes	0.25	0.21	0.19	0.52
close contacts	−0.10	−0.04	−0.14	−0.12
N_{actives}[a]		49	39	282
$N_{\mathrm{inactives}}$[b]		1766	1434	11582
$N_{\mathrm{actives,optim}}$[c]		49	39	100
$N_{\mathrm{inactives,optim}}$[d]		100	100	100
iterations[e]		53	16	16
F_{initial} (LD)[f]		229	520	1705
F_{final} (LD)[g]		825	1167	3530
F_{initial} (LR)[h]		167	1606	787
F_{final} (LR)[i]		2621	2559	1436
F_{initial} (optim)[j]		17	29	29
F_{final} (optim)[k]		87	74	89
EF_{initial} (LD)[l]		3	6.5	6.3
EF_{final} (LD)[m]		5.3	10	8.3
EF_{initial} (LR)[n]		1.6	5.5	4.5
EF_{final} (LR)[o]		8.3	7	5.9

[a] Total number of ligands. [b] Total number of decoys. [c] Number of ligands docked for optimization. [d] Number of decoys used during optimization. [e] Number of iterations during optimization. [f] Initial F computed for ligands versus decoys. [g] Final F computed for ligands versus decoys. [h] Initial F computed for ligands versus random molecules. [i] Final F computed for ligands versus random molecules. [j] Initial F computed for ligands versus decoys used for optimization. [k] Final F computed for ligands versus decoys used for optimization. [l] Initial enrichment factor at 5% database size computed for ligands versus decoys. [m] Final enrichment factor at 5% database size computed for ligands versus decoys. [n] Initial enrichment factor at 5% database size computed for ligands versus random molecules. [o] Final enrichment factor at 5% database size computed for ligands versus random molecules.

target, scoring function, or docking program due to its generic approach. For example, when using ProPose, the downscaling of the hydrogen-bonding parameter has the strongest impact on CDK2 screening results (see Table 1). This can be attributed not only to a change in score calculation but additionally to a strong impact of this particular parameter on base fragment placement, which is usually dominated by hydrogen bonding. But for COX2, this parameter is up-scaled considerably in order to provide optimized results. Although the parameter values obviously do not generalize to different targets, it has to be noted that this does not impose real problems on the virtual screening workflow: the optimization is performed only once in advance of virtual screening for ligands of a specific target, and any docking calculation during virtual screenings later on will benefit from the optimized set of parameters.

The improvements achieved are notable: for example, for CDK2, the enrichment of ligands using a background of random molecules increased from $\mathrm{EF}_5 = 1.6$ to $\mathrm{EF}_5 = 8.3$, which means that the poor initial scoring function has been tuned into a much more predictive one. In this case, the signal-to-noise ratio F is improved by a factor of nearly 16. The change in enrichment is less dramatic for ER and COX2 (ER: $\mathrm{EF}_5 = 5.5$ to $\mathrm{EF}_5 = 7.0$; COX2: $\mathrm{EF}_5 = 4.5$ to $\mathrm{EF}_5 = 5.9$), which is in agreement with the change in F (ER: factor of 1.6; COX2: factor of 1.8): due to the definition of F, the relative change in enrichment is approximately the square root of the relative change in F (see Table 1). Although the CDK2 results are outstanding in this set of targets, the improvements for ER and COX2 are still highly respectable for a local search algorithm and the strict setup used, that is, the rigid protein definition, considering only the top-scoring pose, and using a large number of diverse ligands and decoys, which does not allow for artificially good results by overfitting of the data. In general, the results are in-line with preliminary results obtained for another target using another optimization method.[1] Similar improvements are found for using a background of decoy molecules (see Table 1). Using the same set of ligands and decoys from DUD in combination with DOCK 3.5.53 and a physics-based scoring function, Huang et al. reported enrichment factors of $\mathrm{EF}_{20} = 1.4$, 1.3, and 3.3 for CDK2, ER, and COX2, respectively.[16] The initial Böhm scoring function here achieved considerably better results for CDK2 ($\mathrm{EF}_{20} = 3.0$) and ER ($\mathrm{EF}_{20} = 2.8$), but inferior results for COX2 ($\mathrm{EF}_{20} = 2.5$). The optimized scoring functions further improved the results for CDK2 ($\mathrm{EF}_{20} = 3.2$) and ER ($\mathrm{EF}_{20} = 3.5$) and achieved comparable results for COX2 ($\mathrm{EF}_{20} = 2.9$).

Remarkably, both LD and LR discrimination are improved, although only ligand−decoy discrimination was considered during the LD optimization runs. Therefore, the optimized scoring function parameters can be generalized to a different reference database, providing a first cross-validation of the method. Additionally, the improvements gained for the much smaller subset of ligands and decoys used during optimization translate well into what is observed later for the full set of molecules. For example, F calculated for the subsets increased by factors of 5.1, 2.6, and 3.1 during optimization for CDK2, ER, and COX2, respectively. Taking into account the full set of molecules, F improved by factors of 3.6, 2.2, and 2.0 for CDK2, ER, and COX2, respectively. Although there is some loss of improvement, this is clearly paid-off by the large reduction in computing power needed (between 92% and 98%). The ability to generalize from a small subset to a much larger database of molecules provides a second cross-validation, indicating that the method indeed learns general patterns from the training data.

The increase in signal-to-noise ratio is visible in the score spectra, which plot the number of molecules versus their score value (Figure 3). For CDK2, all scores are offset to higher scores, usually indicating less affine molecules (Figure 3A). However, the decoy scores and in particular the ligand scores shift to lower scores relative to the scores of random molecules. This effect leads to a pronounced improvement in the discrimination of ligands, decoys, and random molecules, which is clearly reflected in the ROC curves. The score distributions are not fully separated after the optimization procedure, which may be attributed to the limitations of the local search algorithm and the rigid receptor docking (see Discussion later on). For the estrogen receptor, the overall variance of the scores increased, but the ligand scores were shifted toward lower scores relative to the decoy scores (Figure 3B). Even this small shift is sufficient to induce a considerable improvement of the enrichment curve, for both LD and LR discrimination. The scores resulting from COX2 docking do not change their range, but ligand and decoy scores are moved toward the lower end of the random molecules score distribution (Figure 3C). The enrichment

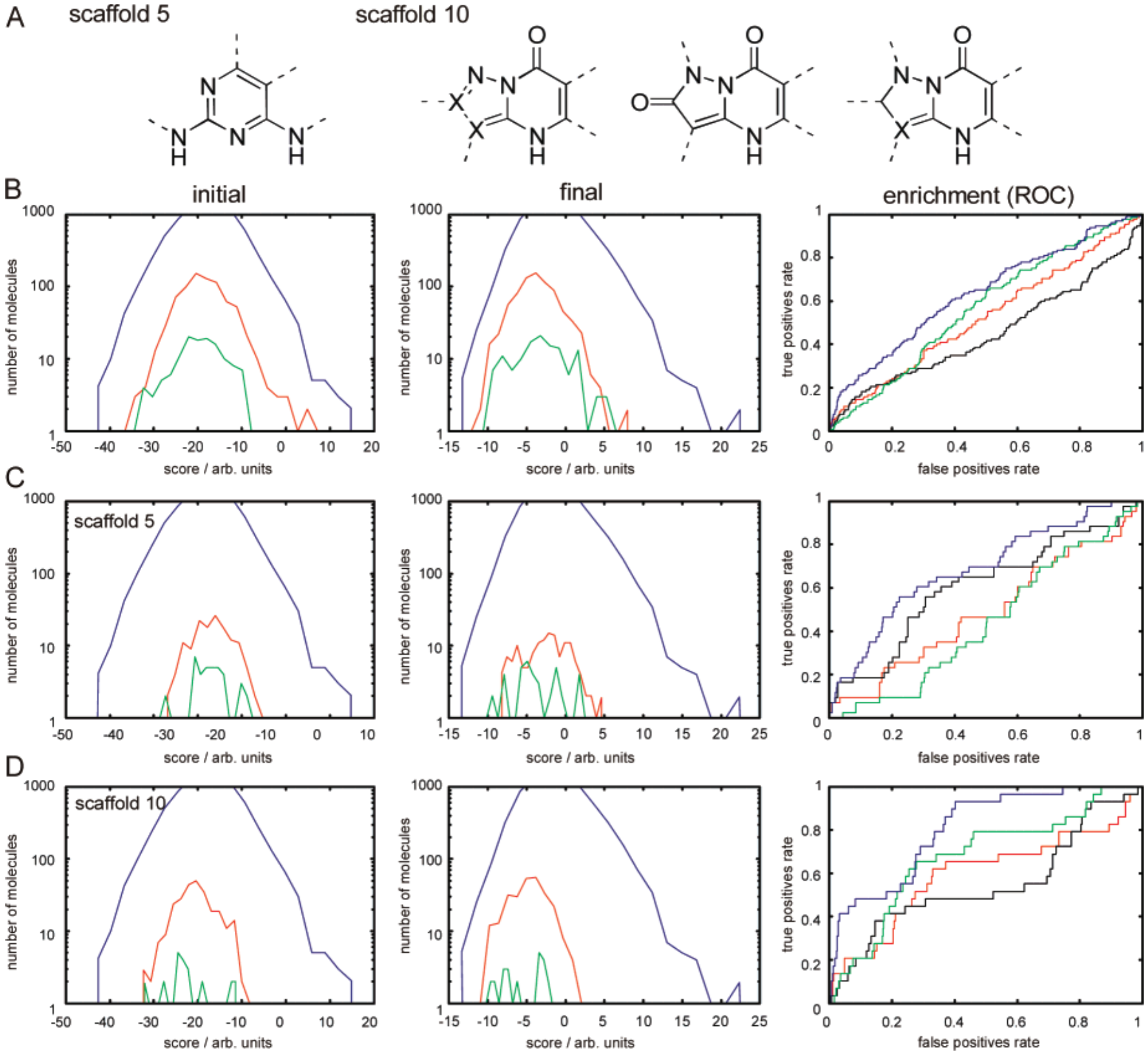

Figure 4. Results from the external validation. Scaffolds 5 and 10 from the Bradley data set are illustrated (A). Initial score spectra (left column), final score spectra (middle column), and enrichment curves (right column) for the Bradley CDK2 data set are shown for all scaffolds under consideration (A), for scaffold 5 (B), and for scaffold 10 (C). The score spectra depict the score distribution of ligands (green), decoys (red), and random molecules (blue). The enrichment curves—here shown as receiver-operating curves (ROC)—of initial (red) and final (black) ligand—decoy discrimination and initial (green) and final (blue) ligand—random-molecule discrimination are plotted.

of true ligands compared to decoys and random molecules is clearly improved, as visible in the ROC curves.

Comparing the results for CDK2, ER, and COX2, no simple rule can be given for how the training data influences the efficiency of the algorithm. But in general, the performance of the algorithm and the robustness of the results both depend on the size and composition of the training data set. Since the optimization method relies on score distributions, the stability of these determines the stability of the results. A previous study showed that score distributions are rather stable if the differences of the distributions of simple molecular descriptors like molecular weight are small.[19] Additionally, ligands are likely to be similar to each other with respect to such descriptors—as long as induced fit and solubility anchors are neglected—due to their ability to bind to the same active site. Decoys are similar to ligands by definition.[16] Therefore, score distributions of ligands and decoys are relatively stable (compare, for example, Figures 3A and 4B), as long as the sampling error of the difference in mean scores $\sim \sqrt{\sigma_L^2/n_L+\sigma_D^2/n_D}$ is controlled. This sampling error is caused by using only a subset of ligands and decoys for optimization. When random molecules are used as counterparts of the ligands during optimization, stability of the results can be achieved by additionally matching the distributions of simple molecular descriptors between different training sets. Finally, the size of the training data sets (>100 molecules) used in this study for learning only a few scoring function parameters (e.g., seven for the modified Böhm scoring function) does not allow for overfitting the training data.

Statistical Analysis of Docking Scores. The assessment of the statistical significance of the improvements in the ROC plots is not straightforward due to the paired nature of the data; that is, the same set of molecules is used for docking calculations with both the initial and the optimized scoring function.[39,46] Therefore, the correlation between the scores calculated with both scoring functions has to be considered when assessing the significance. This correlation can be quite large; for example, the COX2 ligands showed an effective correlation coefficient of 0.82 between the scores calculated with the initial and the optimized scoring function. For detailed analysis, a parametric approach was performed

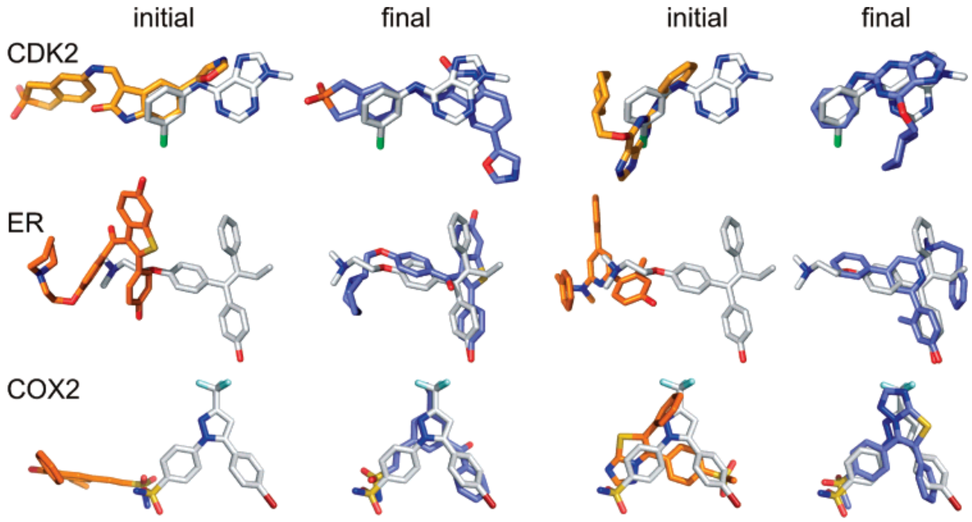

Figure 5. Examples of improved binding mode reproduction using target-specifically optimized scoring functions. The binding modes resulting from the initial and optimized scoring functions are shown in orange and blue, respectively. Two examples for each target are given. For CDK2, the hydrogen bonding to the protein backbone clearly improved. ER inhibitors, which did not dock into the active site initially, were positioned correctly using the optimized parameters. The binding mode of COX2 inhibitors was improved considerably.

because the score distribution does not deviate considerably from a normal distribution, as visible in the score spectra. ROCKIT[40,47] analysis showed that AURC improved significantly ($p < 0.0001$) for CDK2 ligands and decoys, CDK2 ligands and random molecules, COX2 ligands and decoys, and COX2 ligands and random molecules. The improvement of AURC for ER ligands and decoys was not significant, but for ER ligands and random molecules, a significant increase was found ($p < 0.05$). Even more important, the true-positive rate at a given false-positive rate was significantly improved for all targets: TPR at a 5% FPR, which is approximately comparable to the enrichment factor $\text{EF}_{5\%}$, improved by 0.15 ± 0.14 (95% confidence interval, $p < 0.01$), 0.42 ± 0.13 ($p < 0.0001$), 0.08 ± 0.04 ($p < 0.0001$), 0.08 ± 0.05 ($p < 0.001$), 0.16 ± 0.15 ($p < 0.05$), and 0.18 ± 0.18 ($p < 0.05$), for CDK2 ligands and decoys, CDK2 ligands and random molecules, COX2 ligands and decoys, COX2 ligands and random molecules, ER ligands and decoys, and ER ligands and random molecules, respectively. Full data regarding the fit of the ROC curves can be found in Supporting Information S2, including plots of the fitted ROC curves. The errors associated with the estimated improvements of TPR are relatively high due to the propagation of errors where the absolute uncertainties in the individual fitted ROC curves have to be summed up. But it has to be emphasized that this thorough statistical assessment clearly demonstrated the presence of significant improvements.

External validation is indispensable for assessing the generalizability of the optimized parameter set and for preventing overfitting of the training data. Using the independent and sufficiently large CDK2 data set of Bradley et al.,[41] the effect of the optimized CDK2 parameters was examined. Figure 4 shows that improvements are visible for all scaffolds under consideration (see Methods). Ligands and decoys are clearly shifted toward better scores with respect to the distribution of the scores of the background "random" molecules, which are actually derived from a generic screening library. This effect is similar to what has been observed for the training data set (see Figure 3A). The improvement in ligand−decoy discrimination, however, was marginal in this case. The statistical significance of the differences in the ROC curves was analyzed using ROCKIT (see Table 2). This analysis revealed significant ($p = 0.0023$ ... $p < 0.0001$) improvements in the ROC area as well as in $\text{TPF}_{5\%}$ for ligand−random molecule discrimination, which is in line with the observed changes in the score spectra. Again, ligand−decoy discrimination did not profit significantly from the optimized parameters. However, the most

Table 2. Summary of the External Validation of the Optimized CDK2 Scoring Function

scaffolds	0, 5, 10	5	10
N_{actives}[a]	132	43	29
$N_{\text{inactives}}$[b]	771	143	269
N_{random}[c]	8383	8383	8360
initial ROC parameters			
A_{LD}[d]	0.53	0.50	0.55
TPF_{LD}[e]	0.09	0.09	0.16
A_{LR}[f]	**0.57**	**0.44**	**0.66**
TPF_{LR}[g]	**0.05**	**0.009**	**0.14**
final ROC parameters			
A_{LD}[d]	0.46	0.58	0.55
TPF_{LD}[e]	0.1	0.12	0.16
A_{LR}[f]	**0.64**	**0.68**	**0.81**
TPF_{LR}[g]	**0.17**	**0.21**	**0.38**
significance			
$p(A_{\text{LD}})$[h]	0.0092	0.034	0.47
$p(\text{TPF}_{\text{LD}})$[i]	0.38	0.23	0.49
$p(A_{\text{LR}})$[j]	0.0023	<0.0001	0.0003
$p(\text{TPF}_{\text{LR}})$[k]	<0.0001	<0.0001	0.006

[a] Number of active molecules, i.e., ligands. [b] Number of inactive molecules, i.e., decoys. [c] Number of background molecules. [d] Area under ROC curve for ligand−decoy comparison. [e] True positive rate at 5% false positive rate for ligand−decoy comparison. [f] Area under ROC curve for ligand−random comparison. [g] True positive rate at 5% false positive rate for ligand−random comparison. [h] One-tailed significance level for difference in A_{LD}. [i] One-tailed significance level for difference in TPF_{LD}. [j] One-tailed significance level for difference in A_{LR}. [k] One-tailed significance level for difference in TPF_{LR}. Significant improvements obtained by the optimized parameters are shown in boldface.

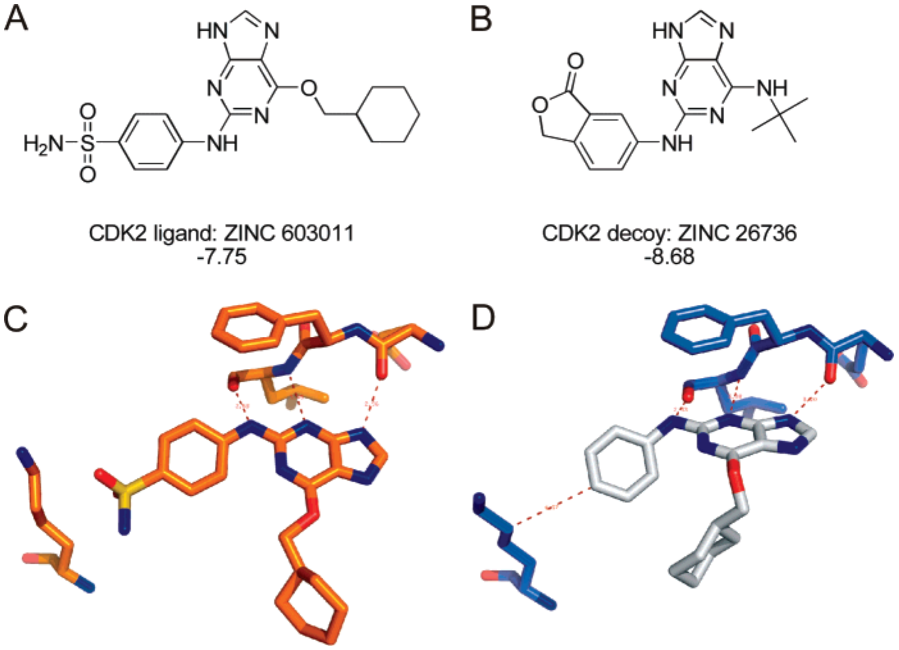

Figure 6. An example of a CDK2 ligand and decoy pair where discrimination by docking failed. The CDK2 ligand (A) receives a slightly worse score than the decoy molecule (B), with the difference being negligible compared to the overall distribution (see Figure 3A, middle column). The failure to discriminate between both molecules results from the rigid receptor hypothesis used during docking. (C) The experimental binding mode of ZINC-603011 (PDB ID 2C6O) indicates an interaction between the sulfonamide moiety and lysine 364 shown on the left-hand side. (D) Using the optimized parameters, the core structure (ZINC-3814462) of both the ligand and decoy is docked well into the active site of CDK2, as derived from PDB ID 1CKP. However, due to a different conformation of the lysine and a shift of the docked molecules of ∼0.8 Å relative to the experimental binding mode, a clash occurs when the sulfonamide is added to the core structure during incremental construction. This leads to bad binding modes for both the ligand and decoy (data not shown) and, thereby, to an indifferent score. A flexible treatment of the protein is likely to result in avoidance of such cases.

important issue for virtual screening performed in order to identify hit molecules is indeed ligand−random molecule separation. Ligand−decoy discrimination, in contrast, gets more important during lead identification and optimization. Therefore, early virtual screening efforts in the course of a drug discovery project will certainly profit from the optimization procedure described here.

In summary, the results derived from a small training data set (∼150 molecules) generalize sufficiently not only to large databases of ligands/decoys (∼1800 molecules) and ligands/random molecules (∼5000 molecules) but additionally to a fully independent validation data set (∼9200 molecules). Furthermore, new scaffolds can be identified which are not present in the training data: for example, scaffold 10 of the Bradley data set was not present in the small subset of DUD used for training the CDK2 scoring function. In fact, there is a molecule in the full DUD data set for CDK2 (ZINC-02239389), which belongs to this scaffold but is actually earmarked as “decoy”. Due to the subset selection, this molecule was not part of the training data. Therefore, scaffold 10 was not known to the optimization algorithm, but nevertheless the optimized CDK2 scoring function improved the enrichment of this scaffold in the external validation data set (Figure 4D).

Analysis of Binding Modes. Notably, the optimization of the signal-to-noise ratio of scoring functions resulted in improved binding mode reproduction for some ligands (Figure 5). A close inspection of the binding modes revealed that this improvement depends on the scaffold of the ligands: for example, none of the staurosporine analogs was docked correctly into the CDK2 active site derived from PDB 1CKP using any of the scoring functions. This is in contrast to, for example, the indolin-2-one scaffold where 10 out of 12 molecules were docked successfully using the optimized scoring function (initial scoring function: 8 out of 12). In total, 30 out of 49 ligands were docked successfully into CDK2. Similar observations were made for ER and COX2: for example, the tamoxifen-like and the chromane scaffolds were not docked correctly into the ER active site with any of the scoring functions, while all of the other scaffolds were either docked equally well with both initial and optimized scoring functions (e.g., pyridine/pyrimidine core structure; initial: 9 out of 14; optimized: 11 out of 14) or the binding modes were clearly improved using the optimized scoring function (e.g., hydroxy-naphtalene scaffold; initial: 5 out of 11; optimized: 8 out of 11). In total, 22 out of 39 ER ligands were docked with reasonable binding modes. For COX2, 186 molecules received a score using both the initial and the optimized scoring function. Difficulties arose from the nonselective COX2 inhibitors like diclofenac or ketoprofene, where the experimental binding modes could not be reproduced using the current active-site definition. Regarding the selective COX2 inhibitors, some scaffolds, for example, those containing the imidazole (initial: 19 out of 32; optimized: 28 out of 32) and pyrazole (initial: 23 out of 40; optimized: 24 out of 40) core structures, were docked very well, in contrast to, for example, the thiazole scaffold (initial: 4 out of 14; optimized: 4 out of 14). In total, 102 out of 186 binding modes were reproduced well. For one scaffold—the imidazole scaffold—statistical significance was achieved (McNemar χ^2 test, $p < 0.05$), showing that this scaffold clearly benefited from using the optimized scoring function. The observed scaffold dependence explains the overlap of the score distributions, which is present even when the optimized scoring function is used: the ligands whose native binding modes are not reproduced sufficiently cannot be expected to receive a better scoring on average. Additionally, this implies that it is necessary to use a large number of molecules and scaffolds for investigating and optimizing the performance of scoring functions.

The differences between the scaffolds are clearly an artifact caused by the rigid receptor hypothesis, which is commonly made for fast docking calculations. For example, the ligand−decoy pair shown in Figure 6 cannot be distinguished on the basis of the calculated scores. Although the purine core structure is docked reasonably well into the CDK2 active site, the addition of the benzenesulfonamide or isobenzofuranone moiety during the incremental construction, which is carried out by ProPose, leads to a steric clash with lysine-364. The lack of protein flexibility prevents a suitable binding mode to be identified by docking. The application of algorithms, which support protein flexibility, is likely to alleviate this problem, however, with additional computational costs. Since scoring functions are applied mainly for virtual screening during the hit identification phase of drug discovery, the additional costs may currently prevent the large-scale application of protein flexibility in such a setting. Therefore, it is more reasonable to optimize scoring functions in such a way that the results can be imported directly into the virtual screening workflow, for example, by using a target-specific “in situ” optimization, as is proposed here. In addition to protein conformation, the protonation of the ligand and protein and correct ligand stereochemistry are likely to be equally important. Although the favorable side effect of improved binding modes depends on the scaffold of the ligands, it may inspire a new “reverse” paradigm for improving protein−ligand docking: up to now, the primary focus of protein−ligand docking software was first to reproduce the binding modes and then to predict the binding affinity. Under the premise, however, that the protein−ligand docking software is in principle able to find a nativelike binding mode, it may be possible to focus primarily on scoring and thereby force the top-scoring binding modes to get closer to the native binding mode.

The local optimization algorithm presented here is able to provide significant improvements; the identification of the global optimum, however, is likely to allow for a more meaningful interpretation of the resulting scoring parameters in terms of physicochemical properties. Although being only seven-dimensional, the parameter space is already large enough to complicate the search for the global optimum considerably. Future research will address this important issue.

CONCLUSIONS

In summary, we have shown that target-specifically optimized scoring functions can be generated using ligand−decoy data and an objective function derived from analysis of variance. This objective function F, which is related to the signal-to-noise ratio of the scoring function, has a favorable dynamic range, favors an increased difference of mean scores for ligands and decoys, resists an artificial increase of noise level, exerts a pressure for successfully docking as many molecules as possible, and provides a sound statistical basis. An in situ optimization algorithm has been described, which basically treats the protein−ligand docking software and its implemented scoring function as a black box, thereby taking into account all implementation-specific details of the scoring function and additionally making the procedure generally applicable, for example, for ligand−ligand alignment. Using this algorithm and a large database of ligands and decoys, Böhm’s scoring function, which was chosen here for sake of simplicity, was optimized for three targets of pharmaceutical interest. A significant increase in the enrichment of ligands and in the signal-to-noise ratio was obtained for all three targets under stringent conditions, that is, using only top-scoring poses, a setup as close as possible to real world virtual screening applications, and a sound statistical treatment. It was argued that the rigid receptor hypothesis, which is commonly used for fast docking algorithms, leads to scaffold dependence of the results. The approach described here—although being purely empirical—will open novel routes for designing more effective scoring functions for pharmaceutical research.

ABBREVIATIONS

ANOVA, analysis of variance; AUAC, area under accumulation curve; AURC, area under ROC curve; BEDROC, Boltzmann-enhanced discrimination of receiver-operating characteristic; CDK2, cyclin-dependent kinase 2; COX2, cyclooxygenase 2; DUD, directory of useful decoys; ER, estrogen receptor; FNR, false negative rate; FPR, false positive rate; KL, Kullback−Leibler divergence; KS, Kolmogoroff−Smirnov; KW, Kruskal−Wallis; LDA, linear discriminant analysis; LD, ligand−decoy; LR, ligand−random molecule; MCC, Matthews correlation coefficient; p, significance level; PDB, protein data bank; QSAR, quantitative structure−activity relationships; ROC, receiver-operating characteristic; $\bar{S}_L$, mean score of the ligands; σ_L, standard deviation of ligand scores; $\bar{S}_D$, mean score of the decoys; σ_D, standard deviation of decoy scores; TNR, true negative rate; TPR, true positive rate; vHTS, virtual high-throughput screening.

ACKNOWLEDGMENT

The author is indebted to Daniel Vitt and Bernd Kramer for unwavering support, and to Andrea Aschenbrenner for proofreading the manuscript.

Supporting Information Available: The statistical parameters (S1) and the results of the ROCKIT analysis (S2) are described in more detail. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Seifert, M. H.; Kraus, J.; Kramer, B. Virtual high-throughput screening of molecular databases. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 298.

(2) Michel, J.; Verdonk, M. L.; Essex, J. W. Protein-ligand binding affinity predictions by implicit solvent simulations: a tool for lead optimization. *J. Med. Chem.* **2006**, *49*, 7427−7439.

(3) Foloppe, N.; Hubbard, R. Towards predictive ligand design with free-energy based computational methods. *Curr. Med. Chem.* **2006**, *13*, 3583−3608.

(4) Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M. Validation and use of the MM-PBSA approach for drug discovery. *J. Med. Chem.* **2005**, *48*, 4040−4048.

(5) Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem., Int. Ed. Engl.* **2002**, *41*, 2644−2676.

(6) Hansch, C. On the structure of medicinal chemistry. *J. Med. Chem.* **1976**, *19*, 1−6.

(7) Crippen, G. M. Distance geometry approach to rationalizing binding data. *J. Med. Chem.* **1979**, *22*, 988−997.

(8) Topliss, J. G.; Edwards, R. P. Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.

(9) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; 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.

(10) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., III. Assessing Scoring Functions for Protein-Ligand Interactions. *J. Med. Chem.* **2004**, *47*, 3032−3047.

(11) Hetenyi, C.; Paragi, G.; Maran, U.; Timar, Z.; Karelson, M.; Penke, B. Combination of a modified scoring function with two-dimensional descriptors for calculation of binding affinities of bulky, flexible ligands to proteins. *J. Am. Chem. Soc.* **2006**, *128*, 1233−1239.

(12) Antes, I.; Merkwirth. C.; Lengauer, T. POEM: Parameter Optimization using Ensemble Methods: application to target specific scoring functions. *J. Chem. Inf. Model.* **2005**, *45*, 1291−1302.

(13) Salo, J.-K.; Yliniemelä, A.; Taskinen, J. Parameter refinement for molecular docking. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 832−839.

(14) Andersson, C. D.; Thysell, E.; Lindström, A.; Bylesjö, M.; Raubacher, F.; Linusson, A. A multivariate approach to investigate docking parameters' effects on docking performance. *J. Chem. Inf. Model.* **2007**, *47*, 1673−1687.

(15) Pham, T. A.; Jain, A. N. Parameter estimation for scoring protein-ligand interactions using negative training data. *J. Med. Chem.* **2006**, *49*, 5856−5868.

(16) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.

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

(18) Kashem, M. A.; Nelson, R. M.; Yingling, J. D.; Pullen, S. S.; Prokopowicz, A. S., III; Jones, J. W.; Wolak, J. P.; Rogers, G. R.; Morelock, M. M.; Snow, R. J.; Homon, C. A.; Jakes, S. Three mechanistically distinct kinase assays compared: Measurement of intrinsic ATPase activity identified the most comprehensive set of ITK inhibitors. *J. Biomol. Screening* **2007**, *12*, 70−83.

(19) Seifert, M. H. Assessing the discriminatory power of scoring functions for virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 1456−1465.

(20) Fisher, R. A. The use of multiple measurements in taxonomic problems. *Eugenics* **1936**, *7*, 179−188.

(21) Hastie, T.; Tibshirani, R.; Friedman, J. H. *The Elements of Statistical Learning: Data mining, inference, and prediction*; Springer-Verlag: Heidelberg, Germany, 2001.

(22) Böhm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.

(23) Böhm, H. J. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309−323.

(24) Fisher, R. A. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinburgh* **1918**, *52*, 399−433.

(25) Bortz, J. *Statistik*, 6th ed.; Springer Medizin Verlag: Heidelberg, Germany, 2005.

(26) *Octav*e, version 2.1.40; Eaton, J. W. University of Wisconsin, Department of Chemical Engineering: Madison, WI, 2002. http://www.gnu.org/software/octave/ (accessed Dec 6, 2007).

(27) Kruskal, W. H.; Wallis, A. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583−621.

(28) Wilkie, A. D. Measures for comparing scoring systems. In *Credit Scoring and Credit Control*; Thomas, L. C., Crook, J. N., Edelman, D. B., Eds.; Clarendon Press: Oxford, 1992; pp 123−138.

(29) Kolmogoroff, A. *Grundbegriffe der Wahrscheinlichkeitsrechnung*; Springer: Berlin, 1933 (Reprint: Springer, Berlin, 1973).

(30) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442−451.

(31) Truchon, J. F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488−508.

(32) Kullback, S.; Leibler, R. A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79−86.

(33) E-mail: Kurt.Hornik@ci.tuwien.ac.at.

(34) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*, 2534−2547.

(35) Seifert, M. H.; Schmitt, F.; Herz, T.; Kramer, B. ProPose: a docking engine based on a fully configurable protein-ligand interaction model. *J. Mol. Model.* **2004**, *10*, 342−357.

(36) Seifert, M. H. ProPose: steered virtual screening by simultaneous protein-ligand docking and ligand-ligand alignment. *J. Chem. Inf. Model.* **2005**, *45*, 449−460.

(37) Nissink, J. W.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein-ligand interaction. *Proteins* **2002**, *49*, 457−471.

(38) Gray, N. S.; Wodicka, L.; Thunnissen, A. M.; Norman, T. C.; Kwon, S.; Espinoza, F. H.; Morgan, D. O.; Barnes, G.; LeClerc, S.; Meijer, L.; Kim, S. H.; Lockhart, D. J.; Schultz, P. G. Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science* **1998**, *281*, 533−538.

(39) Zweig, M. H.; Campbell, G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* **1993**, *39*, 561−577.

(40) *ROCKIT*, version 1.1b; Kurt Rossmann Laboratories for Radiological Image Research, University of Chicago: Chicago, IL, 2007. http://www-radiology.uchicago.edu/krl/roc_soft6.htm (accessed Dec 6, 2007).

(41) Bradley, E. K.; Miller, J. L.; Saiah, E.; Grootenhuis, P. D. Informative library design as an efficient strategy to identify and optimize leads: application to cyclin-dependent kinase 2 antagonists. *J. Med. Chem.* **2003**, *46*, 4360−4364.

(42) *Sybyl*, version 7.0; Tripos Inc.: St. Louis, MO, 2006.

(43) *Corina*, version 2.4; Molecular Networks GmbH: Erlangen, Germany, 1999.

(44) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics* **2000**, *16*, 412−424

(45) Due to the definition of $F = {\sigma_{\mathrm{signal}}}^2/{\sigma_{\mathrm{noise}}}^2$ as the ratio of signal variance to noise variance, F is basically the square of the signal-to-noise ratio. The square root of F is related to the area under the ROC curve (AURC), which itself is linked to other measures of distributional differences (Bamber, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psychol.* **1975**, *12*, 387−415.). For example, assuming that both ligand and decoy scores are independent normal variates an almost maximum likelihood estimate of AURC is given by $\Phi((\bar{s}_L - \bar{s}_D)/\sqrt{\sigma_L^2 + \sigma_D^2})$ where Φ represents the normal cumulative distribution function and $\bar{s}_L$, $\bar{s}_D$, σ_L^2, and σ_D^2 denote the sample means and variances for ligand and decoy scores (Reiser, B.; Guttman, I. Statistical Inference for P(Y<X): the normal case. *Technometrics* **1986**, *28*, 253−257. Faragi, D.; Reiser, B. Estimation of the area under the ROC curve. *Statist. Med.* **2002**, *21*, 3093−3106.). Therefore AURC is independent of the number of ligands and decoys, but its dynamic range is of course limited. In that sense AURC is similar to the discriminatory power η^2 (also known as explained variance or squared correlation ratio) derived from ANOVA. Limited dynamic range and independence of sample sizes are useful features for the assessment of different methods, but for optimization purposes they are not optimal. The F value resulting from ANOVA, in contrast, can be written as $\sqrt{F} \sim (\bar{s}_L - \bar{s}_D)/\sqrt{(n_L - 1)\sigma_L^2 + (n_D - 1)\sigma_D^2}$ which is dependent on the number of ligands and decoys and has an unlimited dynamic range (see Supporting Information S1). Assuming further $n_L = n_D = n$ results in AURC $= \Phi(\sqrt{F/n})$.

(46) Hanley, J. A.; McNeil, B. J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **1983**, *148*, 839−843.

(47) Metz, C. E.; Herman, B. A.; Shen, J.-H. Maximum-likelihood estimation of ROC curves from continuously-distributed data. *Stat. Med.* **1998**, *17*, 1033.

CI700345N