J Comput Aided Mol Des (2009) 23:633–644
DOI 10.1007/s10822-009-9276-1

Robust optimization of scoring functions for a target class

Markus H. J. Seifert

Received: 26 January 2009 / Accepted: 18 April 2009 / Published online: 27 May 2009

Abstract Target-specific optimization of scoring functions for protein–ligand docking is an effective method for significantly improving the discrimination of active and inactive molecules in virtual screening applications. Its applicability, however, is limited due to the narrow focus on, e.g., single protein structures. Using an ensemble of protein kinase structures, the publically available directory of useful decoys ligand dataset, and a novel multi-factorial optimization procedure, it is shown here that scoring functions can be tuned to multiple targets of a target class simultaneously. This leads to an improved robustness of the resulting scoring function parameters. Extensive validation experiments clearly demonstrate that (1) virtual screening performance for kinases improves significantly; (2) variations in database content affect this kind of machine-learning strategy to a lesser extent than binary QSAR models, and (3) the reweighting of interaction types is of particular importance for improved screening performance.

Keywords Virtual screening · Scoring function · Protein ligand docking · Anova · Target family · Kinase

Introduction

The modulation of protein kinases has proven to be an accessible way to selectively influence cellular processes using small molecules. Structure-based design has contributed significantly to understanding the startling complexity of the interactions between small molecules and their kinase targets [1]. These efforts resulted in vast amounts of structural data for both ligands and proteins. Systematic and efficient exploitation of this information for drug discovery has the potential to shorten timelines and to reduce the risk of failure. Virtual screening is one of the methods that profits most from such empirical approaches as long as ab initio simulations are prohibitively expensive with respect to computing power. Virtual screening methods are reviewed frequently, e.g., in [2, 3].

The benefit offered by virtual screening methods during the early drug discovery process is directly related to the predictivity of scoring functions that assess protein–ligand binding affinity, albeit some methods use molecular similarity as a measure for binding affinity. Affinity prediction based on the structure of the protein–ligand complex has the potential to provide an estimate unbiased by human perception of chemical similarity. The scoring of protein–ligand complexes still presents a major challenge [4] and large efforts are made to derive scoring functions with increased predictivity [5–11]. Several promising approaches aim at an automated refinement of scoring function parameters using, e.g., optimization of parameters with genetic algorithms [12] or random walking in conjunction with line optimization [13]. Introduction of additional "shims"—i.e., pharmacophore interaction points—is able to improve the affinity prediction as well [14, 15].

A target-specific optimization (TOP) approach that manipulates the score distributions of active (ligands) and inactive molecules (decoys) directly has been proposed recently [16]. TOP iteratively maximizes the separation of the score distributions by identifying suitable scoring function parameters. The separation of the distributions is

Electronic supplementary material The online version of this article (doi:10.1007/s10822-009-9276-1) contains supplementary material, which is available to authorized users.

M. H. J. Seifert (✉)
4SC AG, Am Klopferspitz 19A, 82152 Planegg-Martinsried, Germany
e-mail: markus.seifert@4sc.com; mhj.seifert@gmx.de

measured by the help of an analysis of variance (ANOVA). The increased separation directly leads to an improved enrichment of active molecules in virtual screening experiments. TOP was focused on the optimization towards a docking into a single protein structure. It is, however, desirable to generate optimized scoring functions with a broader applicability domain in order to reduce the need for tailoring a scoring function to each single target separately. This manuscript provides a novel multi-factorial optimization procedure that is able to tune a scoring function to a target family in a single run. In contrast to earlier approaches, this optimization procedure is now based on a two-factor analysis of variance which allows for differentiating the effects of proteins and ligands on the distribution of docking scores. Using the results of this analysis the scoring function parameters are iteratively tuned to finally provide a global optimum for the discrimination of active and inactive ligands irrespective of the particular target protein structure.

The general focus of this study is on improving protein–ligand docking as a virtual screening tool, therefore virtual screening performance is considered as the only relevant parameter for optimization in this context. The target family of kinases was selected here as test case due to its importance for drug discovery, and four well-studied kinase targets, CDK2, EGFR, p38α, and Src, were chosen as representatives. The outline of this study is as follows: (a) The initial ProPose [17, 18] scoring function was evaluated on large full data sets of ligands and decoys for various kinase targets. (b) Random subsets of ligands and decoys for each target were generated and used for subsequent optimization of the scoring function. (c) The optimization procedure DIRECT target-family specific optimization (D-TOP) iteratively refined the parameters of the scoring function by maximizing a multi-factorial objective function. (d) After convergence, the full data sets were re-docked using the optimized scoring function parameters. (e) An external validation was performed by applying the optimized scoring function to two kinase targets not present in the training data set. Finally, the method was compared to binary QSAR, an established machine learning method.

Experimental section

Data sets

The "directory of useful decoys" (DUD) [19] provides nine kinase data sets from which four have been chosen as training data for the D-TOP procedure: CDK2 and p38α of CMGC branch of the human kinome tree [20], and EGFR and Src from the TK branch. This setting enables intra-branch and inter-branch comparisons at the same time. CDK2 has been used as a target before in [16] where a sub-optimal performance of the original ProPose scoring function for this target was detected. For this reason, and to allow for a comparison, CDK2 was included as a target in this study as well. Two independent data sets from the TK branch, FGFR1 and VEGFR2, were utilized for external validation. The full training and external validation data sets comprised 34,690 and 7,664 small molecules, respectively. The bindingDB [21] was evaluated as a source of small molecules as well, but due to its smaller number of molecules per target it was not possible to clarify the influence of sampling issues (see supplementary information, Table S1). Therefore the DUD was the primary source of small molecules.

The DUD contains different tautomerization states for some of the ligands. No special processing was performed on these ligands. The homogeneity of the ligand data with respect to molecular descriptors and frequent substructures was investigated by comparing the DUD ligand data to ligands for the same targets retrieved from the bindingDB [21] and additionally to random compounds (see supplementary information S2 and S3). The D-TOP optimization procedure was applied only to a random subset of the ligand data in order to reduce computational costs. For each target 50 active and 50 inactive molecules were selected randomly. The same set of actives and inactives is docked into the different conformations of a protein. Different random subsets were generated for assessing the robustness of the results. The full ligand data set was used before and after the optimization procedure for calculating the relevant performance measures.

For all targets, protein structures were extracted from the protein data bank (PDB): for CDK2 PDB entries 2R3 M and 2R3O [22], for EGFR entry 1M17 [23], for p38α entries 1A9U [24] and 1WBS [25], and for Src entry 2SRC [26]; i.e., for two kinases, p38α and CDK2, two conformations were taken into account. This provides additional diversity of the structures of kinase active sites. The Src ADP–PNP bound conformation has been modified by rotating the Lys295 side chain dihedral angle between the γ and δ carbon by 90°, in order to allow for the docking of molecules that do not share the ATP-like binding mode. Two kinases from the DUD were used for external validation: FGFR1 (PDB 1AGW) [27] and VEGFR2 (PDB 2P2H) [28]. The active sites of these proteins were prepared for docking using standard procedures [17].

DIRECT target-family specific optimization

The D-TOP algorithm is an iterative procedure consisting of a basic optimization algorithm and a specific objective function: first, Dakota [29] provides the basic optimization algorithm which generates input parameters for the

protein–ligand docking scoring function, calls the docking control program, and reads the resulting value of the objective function. The docking control program initiates the docking of ligands and decoys into the respective target structures, waits until all docking calculations have finished, and subsequently computes the objective function based on the docking scores, which is returned to Dakota. ProPose, a protein–ligand docking software based on the incremental construction algorithm, is used as docking engine [17, 18].

Previous studies showed that the search is more effective when performed in a log-transformed parameter space [16]. Therefore Dakota and its optimization algorithm work in the log-space which is converted back into linear space for the docking calculations by the docking control program. The initial volume of parameter space is defined as follows (absolute values, linear space, lower bound … upper bound): hydrogen bonding 1…10, electrostatic 1…10, aromatic 0.1…1.6, cation-π 0.1…4, hydrophobic 0.01…1, clashes 0.01…1, and close contacts 0.01…1. Due to the log-transformation no changes of the sign of the parameter values are allowed.

Optimization algorithm

A priori it was clear that the algorithm should be deterministic, gradient-free, and able to handle expensive objective functions in a parameter space of low to medium dimensionality. This led to two promising candidates, both originally developed by D. R. Jones et al.: the "efficient global optimization" (EGO) [30] and the "dividing rectangles" (DIRECT) algorithm [31]. The EGO algorithm, however, was not able to achieve significant improvements even in the single target case; therefore the focus was on evaluating the DIRECT algorithm. The Dakota optimization toolkit [29] provides the NCSU DIRECT algorithm [32] which is shown to be effective in identifying scoring function parameters with improved discrimination of active and inactive molecules. DIRECT iteratively divides a given hypercube into smaller volumes while focusing more and more into regions with higher values of the objective function. The algorithm stops when the lower limit of the box volume (in our case 10^{-12}) is reached.

Objective function

The objective function is calculated using a multi-factorial ANOVA which is applied to the docking scores after preprocessing: for some molecules more than a single score value exists due to different tautomers or target protein conformations. In this the case the minimum score is used for calculating the objective function. For example, docking 50 actives and 50 inactives per protein structure leads to 600 docking calculation per iteration, the results of which are condensed to $\sim$400 scores (depending on the number of tautomers present in the subset) and subsequently analyzed by a two-factor ANOVA: the two factors are activity (factor A: active/inactive) and target protein (factor B: CDK2, EGFR, p38α, Src). The results are stored using effect encoding (see supplementary Table S4) [33].

Factor A, factor B, and the factor interaction A $\times$ B represent the independent variables (x), whereas the score is the dependent variable (y). The factor interaction variables are computed by multiplication of the factor A variable with each of the factor B variables. The independent variable x was extended by one constant column to account for the intercept term before a least squares fit to equation $y = bx$ is calculated. The F ratio with respect to the factor A (activity) is given by [33]:

$$F_{\mathrm{A}} = \frac{\left(R^2_{y,x_{\mathrm{A}}x_{\mathrm{B}}x_{\mathrm{A}\times\mathrm{B}}} - R^2_{y,x_{\mathrm{B}}x_{\mathrm{A}\times\mathrm{B}}}\right)\cdot(N - p\cdot q)}{\left(1 - R^2_{y,x_{\mathrm{A}}x_{\mathrm{B}}x_{\mathrm{A}\times\mathrm{B}}}\right)\cdot(p-1)}$$

with $R_{y,x_{\mathrm{A}}x_{\mathrm{B}}x_{\mathrm{A}\times\mathrm{B}}}$ being the multiple correlation coefficient of the dependent variable with all independent variables, $R_{y,x_{\mathrm{B}}x_{\mathrm{A}\times\mathrm{B}}}$ the multiple correlation coefficient of the dependent variable with all variables excluding x_{A}, N the total number of molecules, and p and q the number of factor grades. The regression weight b_{A} with respect to factor A is used to determine the direction of the effect, i.e., if active molecules score lower or higher than inactive molecules on average. The final objective function is given by:

$$\mathrm{OF} = \mathrm{sign}(b_{\mathrm{A}})\sqrt{F_{\mathrm{A}}}.$$

It has to be noted that this ANOVA-type objective function is used solely for optimization purposes and by no means for proving the statistical significance of the effects. The evaluation of statistical significance is described in the next paragraph.

Scoring function

The scoring function implemented in ProPose is a weighted sum of geometry scores G for different molecular interactions:

$$\begin{aligned}\mathrm{Score} = {} & a_0 + a_{\mathrm{hb}}G_{\mathrm{hb}} + a_{\mathrm{el}}G_{\mathrm{el}} + a_{\mathrm{aro}}G_{\mathrm{aro}} + a_{\mathrm{pi+}}G_{\mathrm{pi+}} \\ & + a_{\mathrm{hyd}}G_{\mathrm{hyd}} + a_{\mathrm{clash}}G_{\mathrm{clash}} + a_{\mathrm{close}}G_{\mathrm{close}}.\end{aligned}$$

It consists of a constant a_0, and terms G for hydrogen bonds (hb), electrostatic (el), aromatic (aro), cation-π (pi+), and hydrophobic (hyd) interactions accompanied by clash penalties (clash) and close contact rewards (close). Each of the geometry terms G is composed of two factors:

$$G_i = \sum_j g_{i,j} \cdot p_{i,j}.$$

For an interaction type i, the product of a geometry factor g with an indicator variable p is summed up over all possible interactions j, where p indicates the presence of a specific interaction. The details of the scoring function are described in Ref. [18].

Data analysis

After convergence of the optimization procedure the new scoring function parameters are evaluated on the full data set including all ligands. Histograms and "receiver operating characteristic" (ROC) curves are plotted and the "area under the ROC curve" (AUR) is calculated. Significance of the improvements as seen in the ROC curves was assessed using ROCKIT, which calculates maximum-likelihood estimates of the parameters of a bivariate model for data from two diagnostic tests and thereby estimates the binormal ROC curves implied by those data and their correlation [34]. External validation was performed using two kinase data sets from the DUD which were not used for optimization: FGFR1 and VEGFR2 (see section "Data sets").

In order to identify the reason for improved virtual screening performance, the contributions to the score were analyzed by normalizing the raw score contributions: as mentioned above, the ProPose score is a sum of factors where each factor is given by the product of *Parameter weight* $a \times$ *Geometry factor* $g \times$ *Presence indicator* p. The last term represents a Boolean variable which indicates the presence or absence of an interaction. This factorization is now used for evaluating the average influence of the new parameters on the formation and geometry of the interactions: first the raw contributions to the score are evaluated and averaged over all targets and all docked molecules from the full training data set. Second, the raw score contributions are divided by the corresponding parameter value. This results in a value that represents the average of the product *Geometry factor* $g \times$ *Presence indicator* p. Third, this value is divided by the average number of interactions per molecule of the respective type. The result now gives an indication of the average geometry factor of the interaction types and allows for analyzing the average contribution of binding mode geometry to the score before and after the optimization procedure.

Comparison to binary QSAR

Since the optimization method applied here strongly resembles a generic machine learning approach, it is interesting to compare this approach to established and computationally less intensive methods: the binary QSAR approach as implemented in MOE (version 2007.09, Chemical Computing Group, 1010 Sherbrooke St. W, Suite 910, Montreal, Canada) is a variant of a naïve Bayes classifier, which is known to be applicable for creating predictive models from kinase data [35, 36]. Therefore binary QSAR models were generated based on exactly the same input data as was used for the D-TOP run. 209 2D and 3Di MOE descriptors were calculated for the 400 input molecules and two binary QSAR models were analyzed: one using exactly the same number of independent parameters (seven) as the scoring function, and another one retaining the optimal 44 independent components as suggested by MOE. These models were applied to the external validation data set and compared to the results of the corresponding docking results.

Results

Learning the parameters

The training data set consists of CDK2, EGFR, p38α, and Src. The crystal structures of the kinases used as training and validation data sets are shown in supplementary figure S5. CDK2 is present with two structures (PDB 2R3M and 2R3O) showing two different Lys33 conformations. Erlotinib in complex with EGFR (PDB 1M17) extends to one of the back-pockets. P38α contributes one DFG-in (PDB 1A9U) and one DFG-out (PDB 1WBS) conformation. Src is represented by its ADP-PNP bound conformation (PDB 2SRC). The validation is performed on FGFR1 and VEGFR2: SU4984 in complex with FGFR1 (PDB 1AGW) does not address any of the back-pockets, in contrast to a pyridinyl-triazine inhibitor in complex with VEGFR2 (PDB 2P2H) which extends well behind the Lys–Asp salt bridge. Therefore, a sufficient structural diversity is present, which reduces the potential bias of the optimized parameters.

At the beginning, the initial scoring function was evaluated on the full training data set. The results are shown in Table 1. The areas under the ROC curves (AUR) indicate no or worse than random probabilities for correctly ranking ligands and decoys. The ANOVA F-ratio with respect to the factor activity is very low $F_A = 0.3$. The value of F_A measures the overall ability to see differences in the score distributions of active and inactive molecules. As expected, the initial set of scoring function parameters—which was optimized originally, like many other docking programs, towards binding mode accuracy [18]—is not suitable for virtual screening of kinase inhibitors, underlining the need for an optimized set of parameters for the special purpose of virtual screening.

Table 1 Initial scoring function: ANOVA $F_A = 0.3$

Protein[a]	Ligands[b]	Act.:inact.[c]	AUR[d]	SE[e]	Rem.[f]
PDB-CDK2-2R3M	DUD-CDK2[g]	71:2,057	0.58	0.03	0
PDB-CDK2-2R3O	DUD-CDK2[g]	71:2,059	0.60	0.03	0
PDB-EGFR-1M17	DUD-EGFR	468:15,915	0.54	0.01	0
PDB-p38a-1A9U	DUD-p38a[g]	441:8,944	0.36	0.01	–
PDB-p38a-1WBS	DUD-p38a[g]	432:8,685	0.32	0.01	–
PDB-Src-2SRC	DUD-Src	145:6,151	0.46	0.02	–

[a] Protein structure

[b] Ligand data set

[c] Number of docked active and inactive molecules

[d] Area under the ROC curve

[e] Standard error of AUR

[f] Remark: +, 0, – refer to AUR > 0.7, 0.7 > AUR > 0.5, AUR < 0.5, respectively

[g] Indicate identical ligand data sets used for docking

An optimization procedure with continuous re-docking would be prohibitive in terms of computing times if it was based on the full training data set of the size mentioned before. Therefore, random samples of 4 targets × 2 activity classes × 50 molecules = 400 molecules were generated from the training data set. Additionally, sets of molecules were docked into two different conformations of two kinases, resulting in 400 + 200 = 600 docking calculations per iteration of the D-TOP optimization procedure. The sampling of parameter space and the convergence of the objective function are shown in Fig. 1.

The "dividing rectangle" (DIRECT) algorithm [32] starts at the center of the hypercube given by the parameter boundaries and iteratively subdivides this space in order to focus on more promising regions. First, the algorithm discovers that higher values of the electrostatic parameter and lower values for the close contact interaction lead to higher values of the objective function (Fig. 1, events A and B). According to the relative improvement, the algorithm first reduces the close contact parameter, and then increases the electrostatic parameter. Similarly, lower values of the hydrogen bonding parameter prove to be better (Fig. 1, events C–E) causing the algorithm to reduce this parameter. This procedure is repeated for all parameters and the volume of the subspaces considered—i.e., the parameter step size—is reduced according to the tiling scheme shown in Fig. 1 (top). The algorithm stops when a preset minimal volume is reached. The DIRECT algorithm is designed to converge relatively fast and thereby limits the computational cost for evaluating the expensive objective function: for example, the optimization shown in Fig. 1 converged after 267 iterations with 267 × 600 = 160.200 docking calculations in total. This procedure has to be performed only once for a given target family and therefore will pay off rapidly.

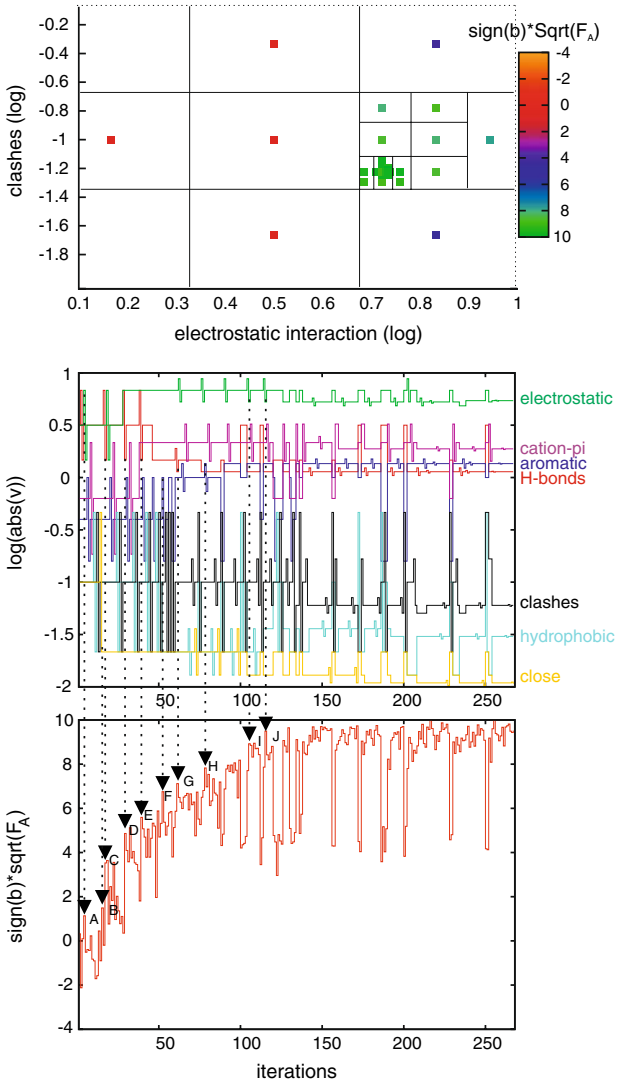

Fig. 1 Example of an optimization run: the division of parameter space into smaller rectangles by the DIRECT algorithm is shown (*top*). The development of the scoring function parameters during the optimization run shows the deterministic variation of the parameters by the DIRECT algorithm (*middle*) which is driven by the improvements found for the objective function (*bottom*). The convergence is accompanied by a series of events where more promising parameter combinations are discovered (*A–J*)

The robustness of these results was assessed by repeating the optimization with independent random subsets of ligands and decoys (set 2 and 3) entering the optimization procedure. The initial and final parameters of all three optimization runs using different random subsets are given in Table 2. The parameter values found after convergence are relatively stable. Additional experiments using more kinases, different crystal structures, and a different source for ligand-decoy data (bindingDB, 21) led to comparable effects, i.e., a down-scaling of hydrogen bonding interactions and an up-scaling of electrostatic interactions (data not shown, a summary of the experiments is given in the

supplementary information Table S1). CDK2 has been part of the single target optimization described in Ref. [16] allowing for a comparison of unifactorial and multifactorial optimization procedures. Both methods give rise to same trends, i.e., a reduction of the hydrogen bonding parameter, and an increase in electrostatic and cation-π parameter values, although the extent of the changes is different.

The pronounced reduction in absolute value of the hydrogen bonding parameter is notable. At first glance this seems counterintuitive due to the importance of hydrogen bonding of kinase inhibitors with the kinase hinge region. But it has to be taken into account that the optimization method focuses solely on ligand-decoy discrimination. Obviously, ligands and decoys cannot be distinguished by the hydrogen bonding interactions alone leading to a reduced relevance of this parameter. Indeed, the DUD data set was constructed such that compounds with similar numbers of hydrogen bond donors and acceptors were selected as decoy compounds for known active molecules [19]. The absolute value of the electrostatic parameter, in contrast, was increased significantly. The major partners for electrostatic interactions with kinase active sites are the residues involved in the salt bridge which links the two sub-domains of a protein kinase. Obviously, forming an interaction with these residues is important for successfully discriminating ligands and decoys. The ad-hoc value for cation-π interactions increased in absolute value, suggesting its importance for distinguishing ligands and decoys. In contrast, the parameter for aromatic interactions and the clash parameter increased in absolute value only in two out of three optimization runs.

Validation

Using the optimized scoring function parameters, the full data set was docked again, and the results are shown in Table 3 and depicted in Fig. 2. In contrast to the initial scoring function parameters, the optimized parameters provide much better discrimination of active and inactive molecules for all but one protein conformation: the area under the ROC curve (AUR) is now between 0.70 and 0.88 for five out six protein conformations. The AUR is a measure for the probability that the ranking of any pair of an active and an inactive molecule is correct [37]. The AUR for docking into p38α DFG-out conformation (PDB 1WBS) increased from 0.32 to 0.51 which means, unfortunately, that the negative enrichment was replaced by random selection. In contrast, the improvements found for docking into PDB 1A9U are undeniable. This preference for 1A9U is easily explained by the type of ligands within the p38α data set of the DUD: in the crystal structure PDB ?tul?> 1A9U, p38α is bound to SB203580—a pyridinyl imidozole compound—and the DUD contains a large fraction of SB203580-like ligands belonging to the same scaffold. In crystal structure PDB 1WBS, the DFG loop is flipped making this conformation unsuitable for binding SB203580-like compounds [38]. Notably, the inclusion of this unsuitable conformation did not stall the overall optimization procedure. Two protein conformations have been considered for CDK2 as well: in this case PDB 2R3M and 2R3O differ in the conformation of the Lys33 which is involved in the typical salt bridge with Asp145 located in the DFG loop [22]. This difference did not cause any pronounced effect on the ROC curves.

Table 2 Scoring function parameters

Parameters	Initial[a]	1[b]	2[b]	3[b]
Hydrogen bonds	−4.700	−1.136	−1.468	−1.285
Electrostatic	−2.000	−5.427	−5.325	−4.844
Aromatic	−0.700	−1.359	−0.736	−1.359
Cation-π	−0.700	−1.880	−1.967	−2.154
Hydrophobic	−0.100	−0.030	−0.026	−0.015
Clashes	0.250	0.060	0.063	0.035
Close contacts	−0.100	−0.011	−0.022	−0.013

[a] Initial parameter values

[b] Optimized parameters for different random subset used for optimization

The improvements due to the optimized parameters are visible in Fig. 2a. Originally, the score distributions overlap strongly and no discrimination between active and inactive molecules is found (data not shown). The optimized parameter set leads to a considerable shift of the active molecules versus the inactive molecules in the score spectra (Fig. 2a). Accordingly, the ROC curves indicate a much higher probability for correctly ranking active and inactive molecules.

An external validation was performed by docking into two new kinase targets, FGFR1 and VEGFR2. The significance of the improvements was assessed by ROCKIT analysis. Comparing DUD ligands not only with DUD decoys but additionally with random decoys showed a significant improvement of AUR in all cases (Fig. 2b): for FGFR1, the differences in AUR as well as in the true positives rate at 5% false positives rate ($TPR_{5\%}$) were highly significant ($p < 0.0001$). For VEGFR2, the difference in AUR was highly significant ($p = 0.0001$–0.0004), whereas the difference in $TPR_{5\%}$ was hardly significant. Obviously, the ROC curve shows an enrichment of active molecules even for the original scoring function, making the improvement in $TPR_{5\%}$ insignificant (see supporting information Fig. S7). Notably, the improvements are nearly insensitive to the choice of the reference database, no matter if it contains decoys similar to the ligands or completely unrelated random molecules. In summary, the

Table 3 Optimized scoring function: ANOVA $F_A = 543.2$

Proteins[a]	Ligands[b]	Act.:inact.[c]	AUR[d]	SE[e]	Rem.[f]
PDB-CDK2-2R3M	DUD-CDK2[g]	72:2,042	0.72	0.03	+
PDB-CDK2-2R3O	DUD-CDK2[g]	67:2,049	0.71	0.03	+
PDB-EGFR-1M17	DUD-EGFR	470:15,906	0.88	0.01	+
PDB-p38a-1A9U	DUD-p38a[g]	448:8,955	0.70	0.01	+
PDB-p38a-1WBS	DUD-p38a[g]	419:8,854	0.51	0.01	0
PDB-Src-2SRC	DUD-Src	118:6,135	0.79	0.02	+

[a] Protein structure

[b] Ligand data set

[c] Number of docked active and inactive molecules

[d] Area under the ROC curve

[e] Standard error of AUR

[f] Remark: +, 0, − refer to AUR > 0.7, 0.7 > AUR > 0.5, AUR < 0.5, respectively

[g] Indicate identical ligand data sets used for docking

twofold validation, including (1) generalization from random subset to the much larger full set and (2) generalization to external data, places confidence in the general applicability of the optimized parameter set. It has been argued that AUR is not a relevant performance measure for virtual screening since it does not take into account what is called "early recognition", i.e., finding actives very early in a ranking list [39]. However, the importance of "early recognition" depends on the costs attributed to true and false positives [40]. Therefore, a standard measure of ranking performance like AUR seems more appropriate unless consent is reached about the costs of virtual screening.

Next, a binary QSAR model was built using exactly the same data set as in the first optimization run, i.e., 400 molecules for which 209 2D and 3Di MOE descriptors were calculated. The binary QSAR method of MOE first transforms the descriptor vectors into orthogonal eigenvectors which serve as input for a naïve bayes classifier [35]. For sake of comparability, only seven independent eigenvectors were used for creating the model, similar to the seven free parameters of the docking scoring function. In contrast to docking scores, the distributions of the posterior probabilities for active molecules as computed by the classifier are clearly non-gaussian (see Fig. 3a). The ROC curves show that the model is highly predictive for the full data set of ligands and decoys. Subsequently, the same external validation was performed as for the docking method (see Fig. 3b). Obviously, DUD ligands can be discriminated from DUD decoys for both targets, FGFR and VEGFR. DUD ligands, however, cannot be discriminated from random decoys at all using this model. This indicates a strong dependence of the QSAR model's predictivity on data set structure.

Evidently, this stands in contrast to the results of the D-TOP optimized scoring functions, which do not suffer from this disadvantage (see Fig. 2b). A second QSAR model using 44 independent components—the optimal value suggested by MOE—improved the predictivity on the training data set but equally failed to discriminate DUD ligands from random compounds (see supporting information S8). The problem of the applicability domain of QSAR models is not restricted to naïve Bayes classifiers but is well described in the general QSAR literature [e.g., 41–43]. This study suggests that it might be advantageous to use docking instead of QSAR models for screening, especially when the content of the database to be screened is not structurally biased towards the applicability domain of the model.

Discussion

The results immediately give rise to question of what has been learned from the training data. In order to answer this question, one has to focus on the score, and the contributions to it, which directly determine screening performance. In the following, the contributions to the final score will be dissected and the most prominent factors influencing the scoring will be identified. In principle, there are two possibilities for achieving better scoring: either, the optimized scoring parameters lead to better pose prediction which subsequently causes improved affinity prediction and screening performance; or the scoring improves relatively independent of pose prediction. Therefore, the differences in binding mode geometries between original and optimized scoring parameters will be examined first. Due to a large number of binding modes to be analyzed, this will be done in a statistical manner and subdivided into three steps according to the three factors giving raise to the final score: after analyzing the number of interactions being formed, the score contribution per interaction type will be investigated, and finally the score contributions will be decomposed into their constituents.

A comparison of histograms, which show how many molecules form a specific number of interactions for various interaction types, gives an indication on the extent of the rearrangement in the number of molecular interactions for the various types. These histograms revealed only relatively small differences for the initial and optimized parameters (see supplementary information S6 for an example). In principle, this does not mean that the individual binding modes are unchanged but that, on average, the number of interactions per type does not change much. For example, the up-scaling of the electrostatic parameter leads to binding modes that are biased towards interaction with Lys–Asp salt bridge. Nevertheless the total number of interactions per type is relatively constant on average.

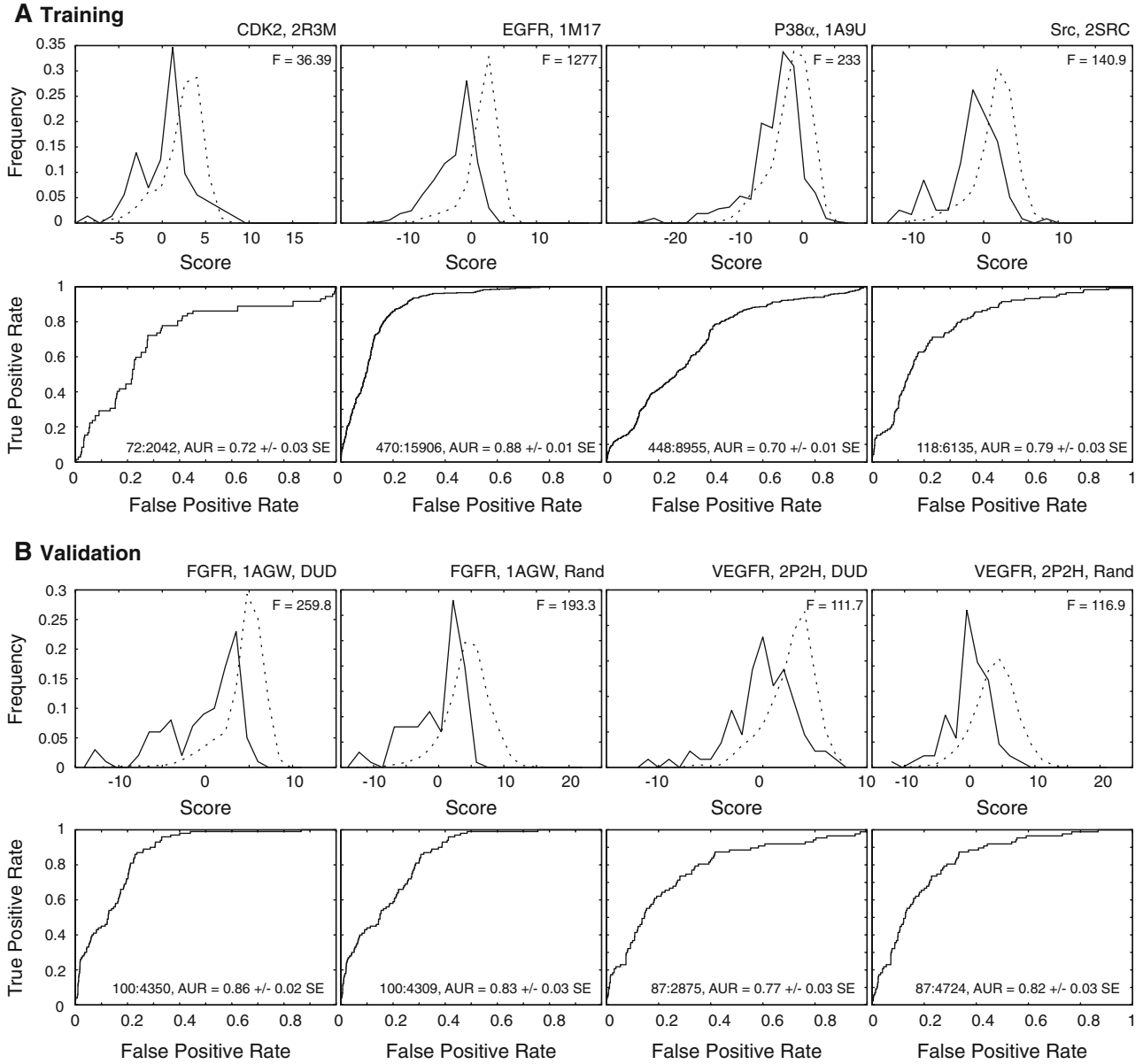

Fig. 2 Optimization of the scoring function for kinase docking. Training (**a**), and validation (**b**) histograms and ROC curves are shown. *Solid* and *dotted lines* depict ligand and decoy score histograms, respectively

In the next step, histograms were compared that evaluate the contribution of various interaction types to the total score. These histograms exhibited sizeable differences across all targets. Figure 4 depicts the results for p38α (PDB 1A9U): using the initial set of parameters, hydrogen bonding is the most important contribution to the interaction score, for both ligands and decoys. Electrostatic, aromatic, and cation-π interactions contribute only on a smaller order of magnitude. This picture changes dramatically when the optimized set of parameters is applied. Now the major part of the score is due to electrostatic, cation-π, and aromatic interactions, with hydrogen bonds having a strongly reduced impact. Additionally, notable differences can be detected now for ligands and decoys, for example with respect to shape and mean value of the distribution of the electrostatic scores. These histograms still contain information about two factors influencing the scoring: the geometrical changes and the overall weight attributed to the interaction types. Therefore these factors have to be dissected according to their contribution to the score.

This can be achieved by normalizing the raw score contributions, which is summarized in Table 4. Obviously the raw contributions to the score are clearly different between initial and optimized parameters, but the contributions after normalization by parameter value and average

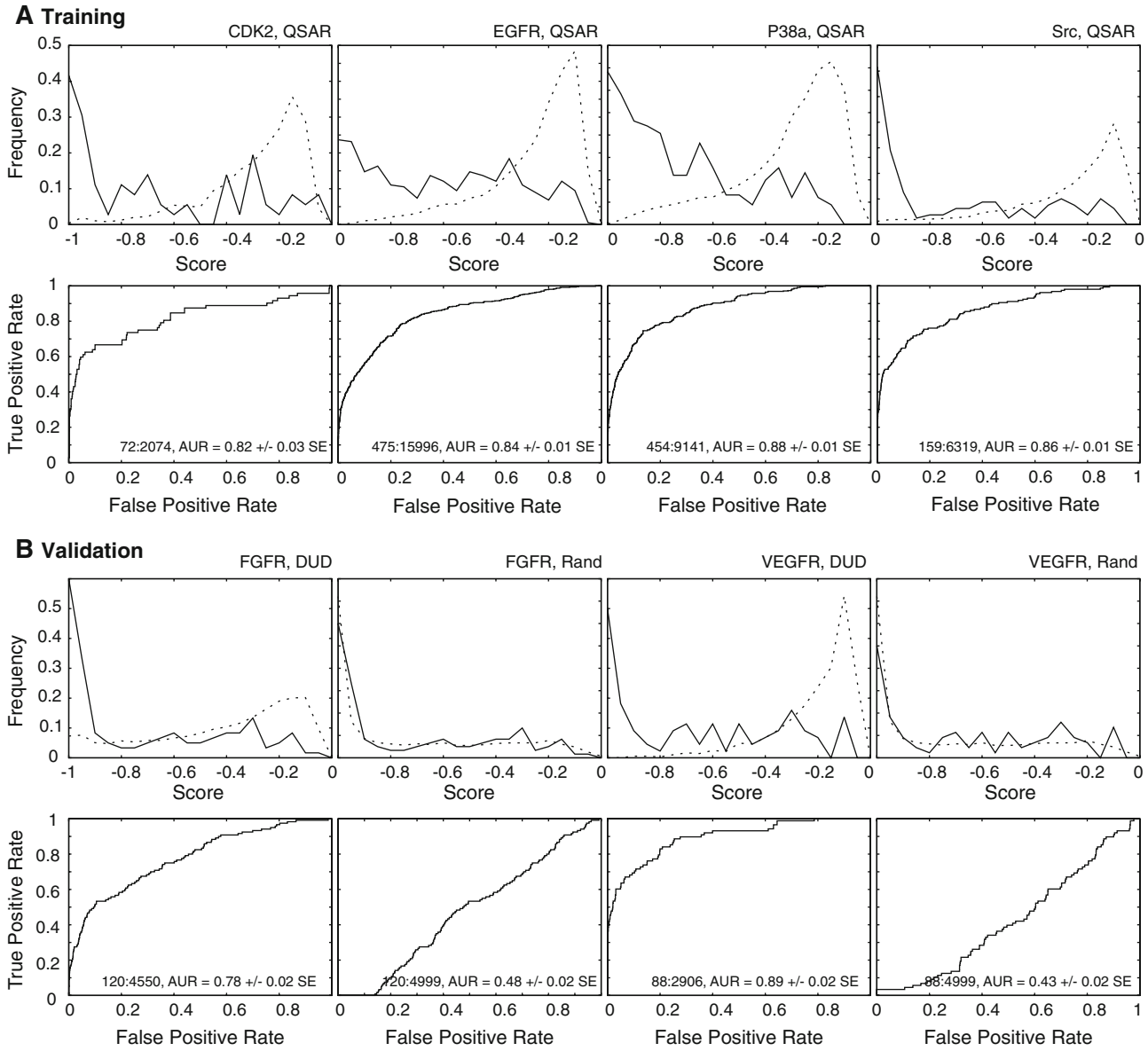

Fig. 3 Comparison to a binary QSAR model based on the same data set as used for scoring function optimization. The score given here corresponds to the negative of the posterior probability as calculated by the binary QSAR model. *Solid* and *dotted lines* depict ligand and decoy score histograms, respectively

number of interaction are relatively constant (see numbers in boldface). More precisely, after normalization by the parameter value, the contributions of clashes and close contacts dominate. Normalizing additionally by the median number of interactions per molecule shows that in all cases electrostatic and close contacts exhibit the highest geometry factors on average whereas aromatic interactions contribute only with low geometry factors. The low average geometry factor of hydrophobic interactions is downscaled even further by the optimization procedure. But overall, the average number of interactions per molecules and the average geometry factors are not affected much by the parameter optimization on average.

Therefore, we are left with one principal source of the improved scoring performance, namely the re-ordering of the overall weights attributed to the various interaction types. This result is not intuitive in the first place, but is in-line with other reports where no or little relationship between binding modes and screening performance was found [12, 44]. Additionally, it might explain why virtual screening and cross-docking were applied successfully [45], even when only rigid protein structures were considered: the binding mode might not be highly accurate, but a screening hit seems to have a definitely higher probability for being active. For example, when key interactions are of long range, like electrostatics, the impact of sub-optimal

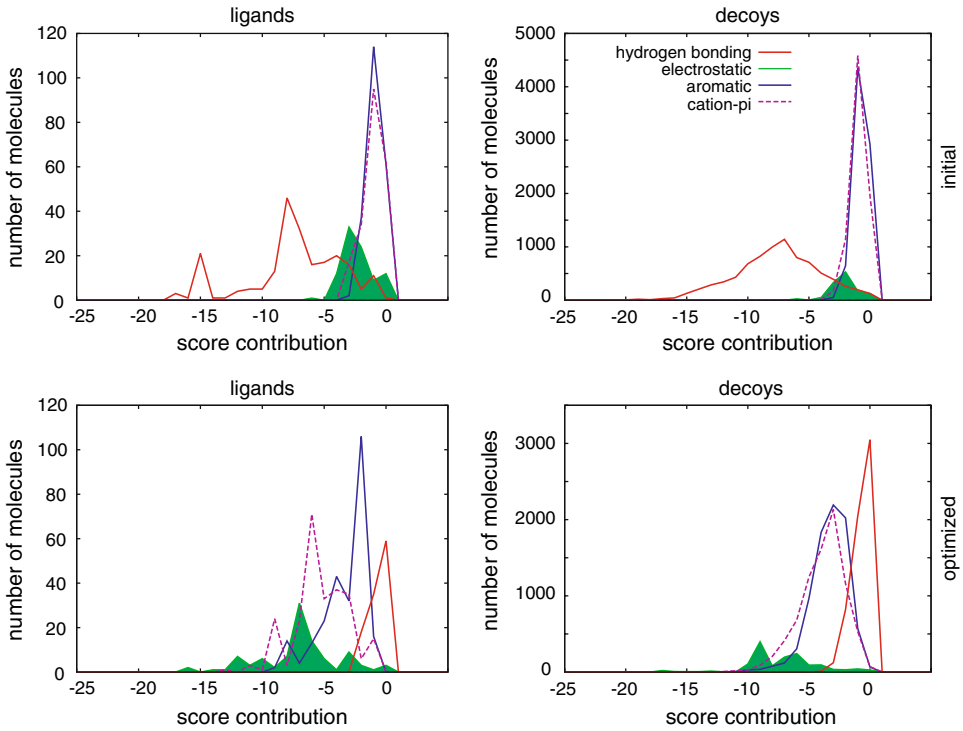

Fig. 4 The effect of the optimized parameters on the contribution of several interaction types to the overall score. As an example, the results for the full p38α data set (PDB 1A9U) are shown here. The *top* panel provides histograms of the score contributions for ligands and decoys using the initial parameter set. The *bottom* panel depicts the corresponding results when applying the optimized set of parameters. Hydrogen bonding (*red*), electrostatic (*green*, *filled*), aromatic (*blue*), and cation-π (*violet*, *dashed*) interaction parameters are shown

binding modes on the score might not be as severe as intuitively expected: depending on the spatial decay of the score with increasing separation between two charged moieties, even a displaced ligand may receive a significant electrostatic score. This issue, however, is still not fully resolved and certainly requires more research. In an earlier investigation it was observed that improved scoring functions were able to improve—as a secondary effect—binding mode accuracy in a ligand scaffold dependent way [16]. The current investigation did not show a significant trend for improved binding modes on average, most likely due to the strong up-scaling of electrostatic interactions. If desired, this issue can be addressed by multi-objective approaches (e.g., 46) which optimize both virtual screening performance and binding mode accuracy. Notably, the Dakota software [29] allows for a straight-forward implementation of such approaches.

Of course, one has to be cautious with interpreting these changes solely in terms of molecular interactions of the protein–ligand complex: first, protein–ligand docking only considers the structure of the complex, but not the properties of the protein and the ligands free in solution. This restriction certainly has an influence on any kind of affinity predictions. However, learning scoring function parameters empirically from experimental data promises to capture at least some of the neglected effects implicitly in the parameter values. Second, algorithm-specific effects cannot be excluded, but this issue has to be addressed by future research with other protein–ligand docking methodologies. In general, it does not seem to be advisable to transfer these parameter changes to other docking programs. There is, however, no obstacle for running this optimization procedure using any other docking program and receiving a specifically optimized parameter set.

Conclusions

In summary, the fully automated method described here allows for tuning protein–ligand scoring functions to specific target families, as exemplified here by the kinase family. The main conclusions from this investigation are: (1) It was shown that D-TOP is a valid procedure for improving the virtual screening performance on a target family, namely kinases. (2) Re-ordering of the interaction-type specific contributions to the final score is an important reason for improved virtual screening performance. (3) The robustness of this procedure with respect to reproducibility and reference database content is an advantage. Notably, this approach takes into account all details of software design and implementation, including all heuristics, by treating the protein–ligand docking software as a black-box. This principle, which is often used in engineering, e.g., for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis, allows for generalizing this method not only to other docking-specific parameters, but to ligand–ligand alignment and to any other virtual screening method as well where the complexity of the underlying algorithms hinders straightforward parameter optimization.

Table 4 Summary of score contribution statistics

Interaction type IA[a]	Raw score contribution S_c[b]		Parameter value P[c]	Score contribution S_c/P[d]		Number of interactions N[e]		Geometry factor $S_c/P/N$[f]	
	Median	IQR		Median′	IQR′	Median	IQR	Median″	IQR″
Initial: actives									
H-bonds	**−6.6**	*4.4*	−4.700	1.4	*0.9*	3	*2*	0.5	*0.3*
Electrostatic	−1.8	*1.8*	−2.000	0.9	*0.9*	1	*1*	**0.9**	*0.9*
Aromatic	−0.7	*0.8*	−0.700	1.0	*1.1*	11	*7*	0.1	*0.1*
Cation-π	−0.6	*0.8*	−0.700	0.9	*1.2*	4	*3*	0.2	*0.3*
Hydrophobic	−0.3	*0.5*	−0.100	2.7	*4.8*	48	*59*	0.1	*0.1*
Clashes	2.8	*2.1*	0.250	**11.3**	*8.5*	37	*26*	0.3	*0.2*
Close	**−18.0**	*7.0*	−0.100	**180.2**	*70.4*	271	*105*	**0.7**	*0.3*
Initial: inactives									
H-bonds	**−7.6**	*4.9*	−4.700	1.6	*1.0*	3	*2*	0.5	*0.3*
Electrostatic	−1.5	*2.0*	−2.000	0.8	*1.0*	1	*1*	**0.8**	*1.0*
Aromatic	−0.7	*0.8*	−0.700	1.0	*1.2*	9	*6*	0.1	*0.1*
Cation-π	−0.5	*0.6*	−0.700	0.7	*0.9*	3	*2*	0.2	*0.3*
Hydrophobic	−0.3	*0.5*	−0.100	3.5	*5.4*	56	*60*	0.1	*0.1*
Clashes	2.8	*2.0*	0.250	**11.1**	*7.8*	38	*22*	0.3	*0.2*
Close	**−18.8**	*6.3*	−0.100	**188.3**	*62.8*	284	*93*	**0.7**	*0.2*
Optim.: actives									
H-bonds	−0.5	*1.1*	−1.136	0.5	*1.0*	2	*2*	0.2	*0.5*
Electrostatic	**−6.5**	*2.8*	−5.427	1.2	*0.5*	1	*1*	**1.2**	*0.5*
Aromatic	**−3.9**	*4.0*	−1.359	2.9	*2.9*	15	*6*	0.2	*0.2*
Cation-π	−2.4	*3.3*	−1.880	1.3	*1.8*	4	*4*	0.3	*0.4*
Hydrophobic	−0.01	*0.1*	−0.030	0.3	*1.8*	11	*31*	0.03	*0.2*
Clashes	0.4	*0.4*	0.060	**7.4**	*7.4*	24	*21*	0.3	*0.3*
Close	−1.5	*0.7*	−0.011	**133.5**	*59.6*	203	*91*	**0.7**	*0.3*
Optim.: inactives									
H-bonds	−0.8	*1.3*	−1.136	0.7	*1.1*	2	*2*	0.4	*0.6*
Electrostatic	**−5.8**	*2.8*	−5.427	1.1	*0.5*	1	*1*	**1.1**	*0.5*
Aromatic	**−3.0**	*2.0*	−1.359	2.2	*1.5*	12	*6*	0.2	*0.1*
Cation-π	−2.1	*2.0*	−1.880	1.1	*1.1*	3	*2*	0.4	*0.4*
Hydrophobic	−0.03	*0.1*	−0.030	0.9	*3.1*	22	*48*	0.04	*0.1*
Clashes	0.5	*0.4*	0.060	**8.2**	*7.1*	28	*21*	0.3	*0.3*
Close	−1.6	*0.7*	−0.011	**144.3**	*64.9*	220	*99*	**0.7**	*0.3*

Italics indicate errors (IQR) similar to sdev

Bold and underlined values are discussed in more detail

[a] Interaction type

[b] Median and inter-quartile range (IQR) of the raw score contributions averaged over all targets and ligands

[c] Parameter value

[d] Raw score contribution normalized by parameter value

[e] Number of interactions per molecule

[f] Raw score contribution normalized by parameter value and average number of interactions per molecule

Supporting information available

A table summarizing the performed experiments (S1), a comparison of various kinase ligand data sets with respect to simple descriptors (S2) and substructures (S3), an example of effect encoding (S4), the active site definitions for the six kinases (S5), histograms of the number of interactions per molecule for various interaction types (S6), and plots of the statistical analysis of the external validation (S7) are presented. Additionally, the results of a 44-component binary QSAR model (S8) are shown.

Acknowledgments The author is indebted to Daniel Vitt and Bernd Kramer for unwavering support, and to all colleagues in the Chem- and Bioinformatics department for the excellent working climate.

References

1. Liao JJ (2007) J Med Chem 50:409. doi:10.1021/jm0608107
2. Kontoyianni M, Madhav P, Suchanek E et al (2008) Curr Med Chem 15:107. doi:10.2174/092986708783330566
3. Sperandio O, Miteva MA, Delfaud F et al (2006) Curr Protein Pept Sci 7:369. doi:10.2174/138920306778559377
4. Moitessier N, Englebienne P, Lee D et al (2008) Br J Pharmacol 153(Suppl 1):S7. doi:10.1038/sj.bjp.0707515
5. Yin S, Biedermannova L, Vondrasek J et al (2008) J Chem Inf Model 48:1656–1662. doi:10.1021/ci8001167
6. Kerzmann A, Fuhrmann J, Kohlbacher O et al (2008) J Chem Inf Model 48:1616. doi:10.1021/ci800103u
7. Raub S, Steffen A, Kämper A et al (2008) J Chem Inf Model 48:1492. doi:10.1021/ci7004669
8. Zhao X, Liu X, Wang Y et al (2008) J Chem Inf Model 48:1438. doi:10.1021/ci7004719
9. Sotriffer CA, Sanschagrin P, Matter H et al (2008) Proteins 73:395. doi:10.1002/prot.22058
10. Fukunishi H, Teramoto R, Takada T et al (2008) J Chem Inf Model 48:988. doi:10.1021/ci700204v
11. O'Boyle NM, Brewerton SC, Taylor R (2008) J Chem Inf Model 48:1269. doi:10.1021/ci8000452
12. Smith R, Hubbard RE, Gschwend DA et al (2003) J Mol Graph Model 22:41. doi:10.1016/S1093-3263(03)00125-6
13. Pham TA, Jain AN (2008) J Comput Aided Mol Des 22:269. doi:10.1007/s10822-008-9174-y
14. Martin EJ, Sullivan DC (2008) J Chem Inf Model 48:861. doi:10.1021/ci7004548
15. Martin EJ, Sullivan DC (2008) J Chem Inf Model 48:873. doi:10.1021/ci700455u
16. Seifert MH (2008) J Chem Inf Model 48:602. doi:10.1021/ci700345n
17. Seifert MH (2005) J Chem Inf Model 45:449. doi:10.1021/ci0496393
18. Seifert MH, Schmitt F, Herz T et al (2004) J Mol Model 10:342. doi:10.1007/s00894-004-0201-1
19. Huang N, Shoichet BK, Irwin JJ (2006) J Med Chem 49:6789. doi:10.1021/jm0608356
20. Manning G, Whyte DB, Martinez R et al (2002) Science 298:1912. doi:10.1126/science.1075762
21. Liu T, Lin Y, Wen X et al (2007) Nucleic Acids Res 35:198. doi:10.1093/nar/gkl999
22. Fischmann TO, Hruza A, Duca JS et al (2008) Biopolymers 89:372. doi:10.1002/bip.20868
23. Stamos J, Sliwkowski MX, Eigenbrot C (2002) J Biol Chem 277:46265. doi:10.1074/jbc.M207135200
24. Wang Z, Canagarajah BJ, Boehm JC et al (1998) Structure 6:1117. doi:10.1016/S0969-2126(98)00113-0
25. Gill AL, Frederickson M, Cleasby A et al (2005) J Med Chem 48:414. doi:10.1021/jm049575n
26. Xu W, Doshi A, Lei M et al (1999) Mol Cell 3:629. doi:10.1016/S1097-2765(00)80356-1
27. Mohammadi M, McMahon G, Sun L et al (1997) Science 276:955. doi:10.1126/science.276.5314.955
28. Hodous BL, Geuns-Meyer SD, Hughes PE (2007) J Med Chem 50:611. doi:10.1021/jm061107l
29. Eldred MS, Brown SL, Adams BM et al (2006) DAKOTA Version 4.0 developers manual, Sandia Technical Report SAND2006-4056. http://www.cs.sandia.gov/DAKOTA/index.html. Accessed 18 Sep 2008
30. Jones DR, Schonlau M, Welch WJ (2004) J Glob Optim 13:455. doi:10.1023/A:1008306431147
31. Jones DR, Perttunen C, Stuckman B (1993) J Optim Theory Appl 79:157. doi:10.1007/BF00941892
32. Gablonsky J, Kelley C (2001) J Glob Optim 21:27. doi:10.1023/A:1017930332101
33. Bortz J (2005) Statistik. Springer, Heidelberg, pp 247–288
34. ROCKIT version 1.1b (2001) Kurt Rossmann Laboratories for Radiological Image Research, University of Chicago: Chicago, IL. http://www-radiology.uchicago.edu/krl/roc_soft6.htm. Accessed 18 Sep 2008
35. Labute P (1999) Pac Symp Biocomput 444
36. Xia X, Maliski EG, Gallant P et al (2004) J Med Chem 47:4463. doi:10.1021/jm0303195
37. Reiser B, Guttman I (1986) Technometrics 28:253. doi:10.2307/1269081
38. Subramanian J, Sharma S, Rao C (2008) ChemMedChem 3:336. doi:10.1002/cmdc.200700255
39. Truchon JF, Bayly CI (2007) J Chem Inf Model 47:488. doi:10.1021/ci600426e
40. Nicholls A (2008) J Comput Aided Mol Des 22:239. doi:10.1007/s10822-008-9170-2
41. Weaver S, Gleeson MP (2008) J Mol Graph Model 26:1315. doi:10.1016/j.jmgm.2008.01.002
42. Benigni R, Bossa C (2008) J Chem Inf Model 48:971. doi:10.1021/ci8000088
43. Tetko IV, Sushko I, Pandey AK et al (2008) J Chem Inf Model 48:1733. doi:10.1021/ci800151m
44. Warren GL, Andrews CW, Capelli AM et al (2006) J Med Chem 49:5912. doi:10.1021/jm050362n
45. Seifert MH, Lang M (2008) Mini Rev Med Chem 8:63. doi:10.2174/138955708783331540
46. Grosdidier A, Zoete V, Michielin (2007) Proteins 67:1010 doi:10.1002/prot.21367