Detailed Analysis of Scoring Functions for Virtual Screening

Martin Stahl*,† and Matthias Rarey‡

Molecular Design, Pharmaceutical Division, F. Hoffmann−La Roche AG, CH-4070 Basel, Switzerland, and Institute for Algorithms and Scientific Computing (SCAI), GMD - German National Research Center for Information Technology, 53754 Sankt Augustin, Germany

Received September 11, 2000

We present a comprehensive study of the performance of fast scoring functions for library docking using the program FlexX as the docking engine. Four scoring functions, among them two recently developed knowledge-based potentials, are evaluated on seven target proteins whose binding sites represent a wide range of size, form, and polarity. The results of these calculations give valuable insight into strengths and weaknesses of current scoring functions. Furthermore, it is shown that a well-chosen combination of two of the tested scoring functions leads to a new, robust scoring scheme with superior performance in virtual screening.

Introduction

The need for fast and efficient discovery of new lead compounds in pharmaceutical research makes it imperative to intelligently use all information accessible for a target. Of special importance in drug discovery programs are 3D coordinates of the target protein obtained from X-ray structure analyses. This information is available for a constantly increasing number of targets, and it can be most efficiently exploited by automated procedures that quickly and objectively test the complementarity of many molecules with the target binding site.[1−4] Over the last years, small-molecule docking programs[5−10] have been established as enormously valuable tools to narrow down the size of compound libraries to the most promising candidates with high success rates.[11−15]

Flexible docking programs are today able to predict protein−ligand complexes with reasonable accuracy and are fast enough to be routinely applied to databases of some 10^4 compounds. The major weakness of docking programs currently lies not in the docking algorithms themselves but in the inaccuracy of the functions that are used to estimate the affinity between receptor and ligand, the so-called scoring functions.[16−18] Scoring functions are needed for two purposes: During the docking process, they serve as fitness functions in the optimization of ligand orientation and conformation, and for comparison with other molecules they are used as estimates of binding affinity for the completely docked molecule. Although in principle different functions can be used for these two purposes, in most applications the same function has been used. Thus there are various criteria for the quality of a scoring function: its ability to identify the correct binding mode of a ligand out of alternative docking solutions, its ability to rank related ligands with respect to their binding affinity, and its ability to select a number of (however weak) inhibitors out of a large database of inactive compounds. Here we will focus on the third criterion, which is the central issue in virtual screening.

Scoring functions used in library docking must be very fast and therefore invariably neglect many terms that are part of the full thermodynamic cycle defining a binding free energy in solution.[19] In addition, they must be error-tolerant, since fast flexible ligand docking approaches crystallographic accuracy only for relatively rigid ligands and in the absence of induced fit phenomena. Given these limitations, scoring functions cannot be expected to give accurate affinity predictions. Nevertheless, they should recognize solutions displaying good steric and electrostatic complementarity between receptor and ligand and give lower ranks other solutions with unlikely relative orientations of ligand and receptor groups. Of central importance is a balanced description of the two major driving forces of complex formation: hydrogen bonds and hydrophobic interactions.

For the present study we have employed the docking program FlexX[8,20,21] as the docking engine. We had observed earlier that the scoring function used so far in FlexX has clear preferences for hydrogen-bonded ligands and performs poorer with lipophilic binding sites and ligands in virtual screening experiments. Here, our goal is therefore to elucidate the strengths and weaknesses of alternative fast scoring functions. We have selected representatives of two classes of scoring functions: empirical and knowledge-based scoring functions. Empirical scoring functions[17] try to capture those elements of binding free energy that are intuitively deemed important by a sum of terms − mainly hydrogen bond, contact surface, and entropic contributions − whose relative weights are either derived by fitting to experimental data or by physical reasoning. Knowledge-based functions are derived from statistical analysis of protein−ligand atom pair distances in X-ray structures of protein−ligand complexes. Generally applicable functions of this type have only been developed recently[22−24] and have only partially been tested for virtual screening.

An unbiased assessment of scoring functions is only possible if results for a variety of different targets are compared that cover a wide range in size, form, and polarity of their binding sites and ligands. Therefore,

* To whom correspondence should be addressed. Tel: +41 61 68 88421. Fax: +41 61 68 86459. E-mail: martin.stahl@roche.com.
† F. Hoffmann−La Roche AG.
‡ GMD - German National Research Center for Information Technology.

10.1021/jm0003992 CCC: $20.00
Published on Web 03/06/2001

Table 1. Number and Origin of Active Compounds Used in This Docking Study

no. of compds	target	origin
128	cyclooxygenase-2	refs 26−28
55	estrogen receptor	refs 29−31
72	p38 MAP kinase	Roche, ref 32
36	gyrase B	Roche
67	thrombin	refs 33, 34
43	gelatinase A and general MMP	WDI, PDB, ref 35
51	neuraminidase	PDB, Roche

we have selected seven targets of high pharmaceutical relevance that represent different classes of enzymes and that differ substantially in their active sites. The ability of the scoring functions to select known inhibitors out of a random library of "drug-like" compounds is compared and analyzed. Potential benefits from consensus scoring with pairs of these scoring functions are discussed. Finally, the knowledge gained from this analysis is used to optimize the combination of two individual scoring functions to a new and general scoring scheme that is of superior performance in virtual screening.

Materials and Methods

In this section, we give details of ligand and target data preparation, briefly outline the scoring functions used in this study, and describe modifications to the FlexX docking software and the setup of the docking calculations.

Preparation of Docking Libraries. Sets of inhibitors of seven well-established pharmaceutical targets listed in Table 1 were compiled manually from Roche therapeutic project databases and public sources such as the PDB database[25] and review articles. Our main selection criterion was to cover as many different compound classes as possible for each target to establish structural diversity among the inhibitors. For all targets, inhibitors activities ranged from low-micromolar to nanomolar affinity. All compounds were stored as SMILES[36,37] and converted to single 3D conformations in Sybyl mol2 format[38,39] by means of CORINA.[40,41] Protonation states were adjusted to generate the structure most likely to be dominant at neutral pH. This was done by a C routine that adds or deletes hydrogen atoms on the CORINA-generated 3D structure according to approximate pK_a values calculated by the program pKalc.[42]

A subset of the WDI database[43] was prepared by removal of compounds with molecular weights greater than 800 or less than 200. Furthermore, compounds with saturated carbon chains longer than 7 carbon atoms, without at least 1 oxygen and 1 hydrogen atom, or containing elements other than C, N, O, P, S, and halogens were removed. The remaining WDI compounds were clustered according to Daylight fingerprint similarity by means of the Jarvis−Patrick algorithm[44] as implemented in the Daylight toolkit[45] considering the 14 nearest neighbors and requiring at least 8 cluster members per cluster. One compound per cluster was selected, resulting in a database of approximately 10 000 compounds. Further selection steps included the removal of macrocycles with larger than 12-membered rings (they are not handled flexibly in FlexX), approximately 70% of molecules with a steroid-type scaffold (because they were over-represented in the 10 000-compound subset), and molecules with more than 13 pharmacophore centers as determined by a Roche in-house program.[46] This step essentially served to remove large molecules with many polar functional groups. The groups of active molecules in Table 1 all passed this filter. The final size of the database was 7528 compounds, which were stored as a SMILES list and converted to Sybyl mol2 format by means of CORINA. Protonation states were corrected as described above.

Preparation of Target Structures. For each of the seven targets, coordinates of protein crystal structures were retrieved from the PDB or the Roche in-house structure collection. Binding pockets were defined manually by means of the interactive modeling program MOLOC developed at Roche. For thrombin, the PDB complex 1dwd was selected and the water molecule adjacent to Tyr 228 in the S1 pocket included as part of the active site. For gelatinase A, the only available X-ray structure was that of a proenzyme mutant (1ck7). Since only minor structural changes can be expected upon complexation and the active site region can be well-superimposed onto other MMP structures, the structure was used in an "in silico activated" form: The N-terminal propeptide was removed and residue 404 mutated to glutamic acid as in the wild-type enzyme. For the estrogen receptor, the PDB structure 1err was selected, which displays the open conformation of the enzyme that can accommodate both agonist and antagonist ligands. For the remaining four targets, unpublished structures solved at Roche were chosen (complexed with Roche inhibitors: neuraminidase, RO-33-0721; gyrase B, RO-60-1034; p38 MAP kinase, RO-115-3528; cyclooxygenase 2, RO-110-3472).

Scoring Functions. The standard scoring function used in FlexX[8] is a modified version of the empirical scoring function by Boehm.[47] It can be written as a sum of five contributions:

$$\Delta G_{\mathrm{bind}} = \Delta G_{\mathrm{match}} F_{\mathrm{match}} + \Delta G_{\mathrm{lipo}} F_{\mathrm{lipo}} + \Delta G_{\mathrm{ambig}} F_{\mathrm{ambig}} + \Delta G_{\mathrm{clash}} F_{\mathrm{clash}} + \Delta G_{\mathrm{rot}} n_{\mathrm{rot}}$$

where the ΔG_i are coefficients of functions F_i operating on the protein and ligand coordinates. $\Delta G_{\mathrm{match}}$ is a sum of scores for directed interactions between receptor and ligand. It consists of individual energy contributions for each hydrogen bond, metal contact, and specific aromatic interaction multiplied by two linear penalty functions for angle and distance deviations from predefined ideal values.[47] The original Boehm and the default FlexX match energy put additional weight on charged hydrogen bonds by means of a scaling factor of 1.667 per charged hydrogen bond partner, provided that the partial charge of the participating ligand atom is above a given threshold. In the present study, this factor was omitted, since according to our experience it has negligible effect on structure prediction and can lead to many false positives in virtual screening experiments because of the large individual score contributions for an individual charged interactions. Neglecting the charge factor was detrimental in the case of specific metal−ligand interactions only. For these, the interaction energy was therefore increased from −2.35 to −3.5 to compensate for the loss of the charge factor. The terms F_{lipo} and F_{ambig} provide a measure of hydrophobic contact surface as functions of receptor−ligand atom pairs, F_{lipo} involving only pairs of unpolar atoms and F_{ambig} involving pairs of one polar and one unpolar atom. Finally, F_{clash} is a penalty function for protein−ligand overlap, and n_{rot} is equal to the number of rotatable bonds in the ligand times a weighting factor. The term $\Delta G_{\mathrm{rot}} n_{\mathrm{rot}}$ was originally intended as a measure of the entropic cost of freezing intramolecular degrees of freedom in the ligand upon complexation,but, in virtual screening, mainly serves to suppress the dependence of the score on the molecular weight. The third empirical scoring function employed here is PLP, a simple four-parameter potential that is a piecewise linear approximation of a potential well for hydrogen bonds and lipophilic interactions without angular terms.[48] The latest version of the PLP potential[49] including a crude hydrogen bond directionality term was not available to us at the time this study was performed.

Two recently published knowledge-based scoring functions were included in this study: DrugScore by Gohlke et al.[24] and PMF by Muegge and Martin.[22] DrugScore was used as a standalone executable supplied by the authors. PMF was implemented according to the original publication and discussions with the author. The original PMF had performance problems in docking applications due to the fact that several pair potentials were repulsive already at relatively long interatom distances. In this study we used a modified parameter file (I. Muegge, unpublished results), in which minima have been extended toward shorter distances and a van der Waals repulsive potential is used at short distances.

In their original form, PLP, DrugScore, and PMF are sums of protein−ligand atom pair contributions. This naturally makes the score dependent on the size of the molecule. In FlexX, size dependence is reduced through the rotatable-bonds penalty term described above. The number of rotatable bonds is roughly proportional to molecular size for many "drug-like" compounds. The rotatable-bonds penalty term (definition of "rotatable" according to Boehm)[47] was used for all scoring functions, albeit with different coefficients of +3.0 for PLP and PMF and +14000 for DrugScore instead of +1.4 in the FlexX function. These coefficients reflect the different scales of the scoring functions relative to the FlexX scores and were derived from regression analyses of docking results on several targets. They are chosen such that they remove any statistical dependence of docking results on molecular size measured in terms of numbers of atoms or rotatable bonds. Inclusion of the rotatable-bonds term effectively improves the performance of PLP, DrugScore, and PMF in virtual screening experiments and facilitates comparison with FlexX results.

The combination scoring function ScreenScore was derived from PLP and FlexX score contributions in the following way: The FlexX implementation of the PLP scoring function (see next paragraph) was used such that all protein−ligand atom pairs already covered by F_{match} term of the FlexX function were excluded from the PLP pair contributions. Since docking more than 7500 compounds in seven binding sites for many different combinations of score terms could not be afforded, the docking run was performed once with the original FlexX scoring function as a fitness function. For each docked compound, a table with FlexX and PLP score contributions was generated for all (up to 800) solutions. Various combinations of PLP and FlexX scores could then quickly be tested by means of these tables. Each new combined scoring function was first used to select the best solution per compound and then for database ranking. The optimization process involved more than 26 million docking solutions for about 50 000 docked compounds. We used an interactively guided systematic optimization scheme with the goal of achieving good enrichment for all targets instead of peak performance for few targets. First, only combinations of F_{match} and the match-excluded PLP term (F_{PLP}) were tested. In a second step, the best fixed-ratio combinations of F_{match} and F_{PLP} were combined with F_{lipo} and/or F_{ambig} contributions. Finally, the optimum value of $\Delta G_{n\text{rot}}$ was determined for the best overall combinations. The final function has the form:

$$\Delta G_{\text{bind}} = F_{\text{match}} + 0.07(F_{\text{lipo}} + F_{\text{ambig}}) + 0.3F_{\text{PLP}} + 1.6n_{\text{rot}}$$

Docking Software. All docking calculations were performed with FlexX version 1.9.2, which contains a number of enhancements and new features with respect to virtual screening which should be briefly described here.[50] For efficient handling of large compound databases, a parallel version named FlexX-PVM was developed. FlexX-PVM, based on the PVM (parallel virtual machine) library,[51,52] is able to run in parallel on heterogeneous hardware environments. To accomplish this, an automatic scheduling system was developed distributing individual protein−ligand docking calculations on the available processors. The scheduling system is robust and allows for stopping and restarting of screening calculations and reconfiguration of the hardware setup during the screening calculation. The scheduler automatically handles defective ligand data and system failures of individual processors. Due to an automatic output file merging, the same output files are created in a parallel run as in a sequential run.

Frequently, the compound database used for virtual screening is created from 2D information with a 3D structure generator. The question arises which enantiomer should be generated for the docking calculation. To avoid time-consuming generation and individual docking of all enantiomers, we extended the definition of degrees of freedom within FlexX to stereocenters. FlexX differentiates between three kinds of stereocenters: pseudo-*R*/*S* (3-bonded, pyramidal nitrogens), *R*/*S*, and *Z*/*E*, which can be handled as additional degrees of freedom during the docking calculation. FlexX then automatically generates the stereoisomer with the best fit to the active site. Since in this study all enantiomers of active compounds were known, this feature was not used here.

Concerning scoring, FlexX was extended in various ways. First, all parameters of the FlexX scoring function including the functional form for the lipophilic and ambiguous contact terms as well as the hydrophobicity definition can be modified externally. Second, we integrated the PLP[48] function into FlexX. Here, the only difference to the original implementation is that halogen atoms are not disregarded but treated as lipophilic atoms. Terms of the three scoring functions can be arbitrarily weighted and combined in FlexX. Furthermore, the user can specify for each term whether it should be used in the complex construction phase or for scoring the final solutions.

An additional feature used throughout this study is a conformation filter developed at Roche. During the incremental construction of docking solutions, the filter removes poses displaying strongly repulsive 1,5-interactions along flexible chains of the ligand. In FlexX, during each incremental construction step, the conformation space of the single bond formed in the previous step is explored. To generate low-energy conformations, the MIMUMBA torsion angle library[53,54] is used. Since in this procedure only one rotatable bond is regarded at a time, high-energy conformations can still result from unfavorable combinations of two dihedral angles along a chain.[55] The intramolecular clash terms are too crude to prevent the generation of such conformations. The new filter detects these conformations − either planar (sp,sp) or *syn*-pentane-type (sc+,sc−) arrangements − and eliminates those which are associated with at least 2 kcal/mol of strain energy. Thus the filter increases the quality of FlexX conformations along flexible chains. It has a positive effect on structure prediction performance in that it helps to focus the conformational search on low-energy structures.

Docking Procedure. For each target, the corresponding set of inhibitors was combined with the WDI subset. The combined library was docked into the target active site using the default FlexX parameter settings contained in the FlexX distribution. The main settings are 800 solutions per iteration during the incremental construction algorithm and a maximum protein−ligand atom−atom overlap of 2.5 Å^3. Calculations were run in parallel on 16 SGI R12k 400-MHz processors with an average wall clock run time of 52 s/molecule. Final scores were calculated for all FlexX solutions (up to 800) per compound. In this way each of the alternative scoring functions was given the freedom to pick a different best solution per compound, whose score value was used for database ranking. Compounds for which no docking solutions could be obtained (on average 6% of the WDI set) were appended to the sorted rank list in arbitrary order.

Results and Discussion

Individual Performance of Scoring Functions. Figure 1 shows the enrichment of inhibitors obtained with four scoring functions for each of the seven targets. The accumulated percentage of inhibitors contained in the top *X%* of the ranked database is plotted. Note the logarithmic scale on the *x*-axis expanding the most important region of the plots from 1−10% of the database.[56] Targets in Figure 1 are roughly ordered in increasing order of the polarity of the binding sites.

The FlexX scoring function performs best for those target−ligand combinations that form a significant number of hydrogen bonds, i.e., p38 MAP kinase, thrombin, gelatinase A, and neuraminidase. In thrombin, most known inhibitors form salt bridges to Asp 189 at the bottom of the S1 pocket and also to Gly 216. These receive high ranks with the FlexX scoring function, whereas inhibitors placing lipophilic groups into the S1

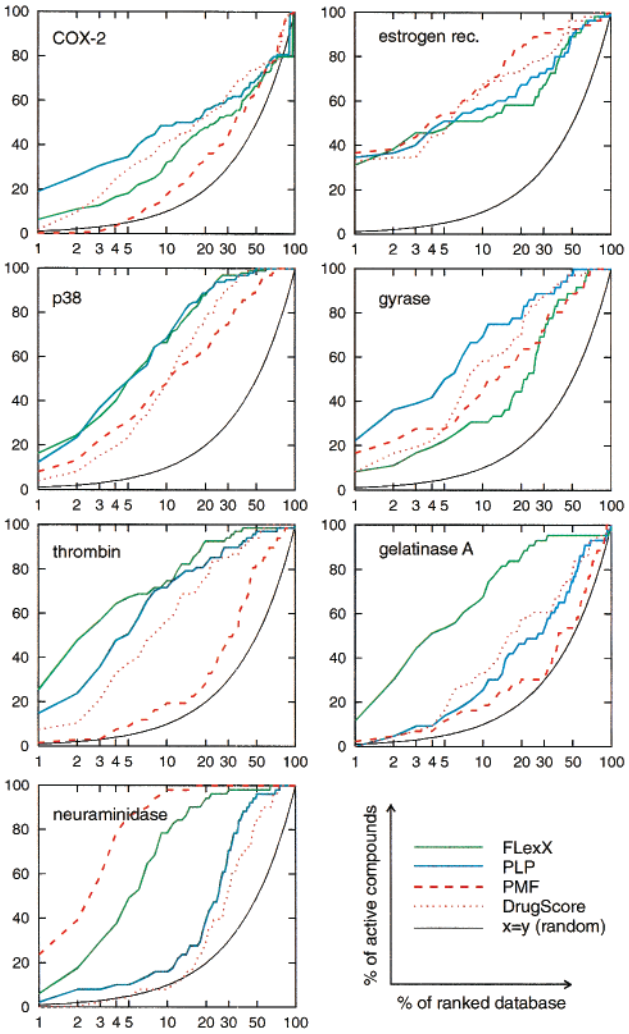

Figure 1. Enrichment of inhibitors for seven targets calculated with four scoring functions. The legend in the lower right-hand corner pertains to all seven panels.

pocket, such as **1** (Figure 2), receive very low ranks although the complex structure is reasonably well-predicted. Gelatinase A has a valley-shaped binding site with a deep S1′ pocket, which can accommodate lipophilic groups as large as biphenyl moieties. The active site is flanked on either side by β-strands allowing for hydrogen bonds to the ligand. These interactions and contacts of inhibitor carboxylate or hydroxamate groups with the catalytic zinc atom result in high ranks once the inhibitors are correctly placed. It is not surprising that only the FlexX function is able to significantly enrich gelatinase A inhibitors, since it is the only one taking metal−ligand interactions into account in an explicit manner. The binding site of p38 MAP kinase is a narrow lipophilic cleft that accommodates planar conjugated systems in the adenine binding region. One rim of the adenine binding pocket is formed by the "hinge" strand. Inhibitors invariably form a hydrogen bond to the NH group of Met 109 in the hinge and frequently another hydrogen bond to a flanking carbonyl group. FlexX again assigns the highest ranks to those inhibitors that form more than one hydrogen bond to the protein. As a point in case, consider the inhibitors **2** and **3**, which are of similar size and shape and adopt the same type of binding mode. Compound **2** is on rank 196, while compound **3** is on rank 69 in the FlexX rank list. The estrogen receptor is a large steroid-size lipophilic cavity with acceptor groups at either end that can form hydrogen bonds with ligand hydroxyl groups, for example, as present in the agonist **4** and antagonist **5**. For both agonists and antagonists, lipophilic interaction energies largely determine the binding energy, but most antagonists form an additional salt bridge to Glu 351. Antagonists are ranked highly by FlexX, while agonists are not. The COX-2 binding site is a narrow, completely buried lipophilic cavity. Hydrogen bond formation is not a predominant feature of inhibitor binding, although most known inhibitors, exemplified by compound **6**, have a sulfonamide group forming hydrogen bonds to various residues (Gln 192, His 90, Phe 518). The FlexX scoring function is not very effective at enriching inhibitors for COX-2, because many docking solutions of WDI compounds display spurious hydrogen bonds, which nevertheless contribute much to the score and lead to high ranks.

Results for the PLP function differ considerably from the FlexX scoring results. Its different functional form

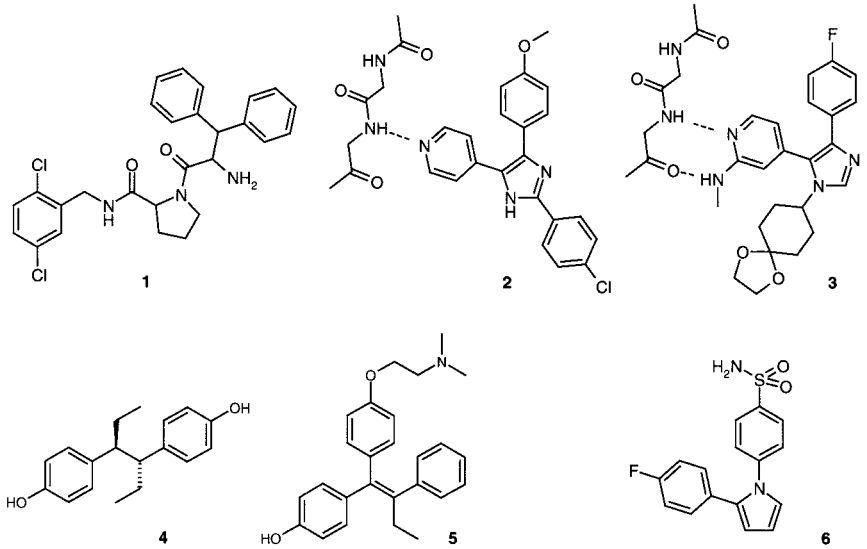

Figure 2. Individual inhibitors discussed in the text: thrombin inhibitor with a lipophilic S1 binding moiety (**1**), p38 MAP kinase inhibitors (**2**, **3**), agonist (**4**) and antagonist (**5**) of the estrogen receptor, and COX-2 inhibitor (**6**).

− no directionality, broad minima − makes the score less dependent on hydrogen bonds and pronounces general steric fit. Thus it is understandable that the PLP potential performs significantly better for COX-2 but significantly worse for neuraminidase. Also, it is striking that PLP is the only scoring function that performs well for gyrase B. In gyrase B an ATP binding site is targeted like in the p38 MAP kinase, but the cavity is more shallow and inhibitors occupy only a fraction of it.

On average, the two knowledge-based scoring functions perform worse than PLP and FlexX. In general it seems that DrugScore can model lipophilic interactions well (COX-2) but fails completely when mainly hydrogen bond interactions count (neuraminidase). A positive feature of DrugScore is the observation that it ranks a considerable number of agonists among the top 10% of the database for the estrogen receptor. Furthermore, DrugScore is the only function to give a relatively high rank (247) to thrombin inhibitor **1** for a docking solution that indeed places the chlorinated phenyl ring in the S1 pocket. On the contrary, PMF outperforms all other functions for neuraminidase but performs poorly in those cases where inhibitors must be fit into relatively narrow cavities. It is difficult to point out the reasons for these observations, since knowledge-based scores are not easily interpretable. However, an obvious weakness of PMF is the fact that many of its carbon−nitrogen and carbon−oxygen pair potentials are strongly repulsive at distances between 3 and 4.5 Å, a situation that results from combined sampling of directed and undirected interactions in crystal structures and combining these statistics in an undirected pair potential. This leads to repulsive interactions, e.g., for phenyl rings located close to amide bonds, e.g., as is the case in benzamidines binding to the thrombin S1 pocket. It should be noted that the observed weak performance of PMF in library ranking does not contradict its demonstrated usefulness in structure prediction or for K_i prediction.[57−59] Finding a correlation among a set of related compounds is a completely different task than ranking a set of completely unrelated compounds.

One might argue that the above results are somewhat biased by the use of the FlexX scoring function as the fitness function for the generation of ligand placements. To address this issue, we have minimized docking solutions for all targets with the PLP and PMF scoring functions using a simplex minimizer.[60] The enrichment curves obtained in this way for PLP and PMF are very similar to the ones in Figure 1 and lead to the same conclusions (results not shown).

Consensus Scoring. Each of the four scoring functions has strengths and weaknesses. To increase the success rate in virtual screening, it may therefore be advisable to use several scoring functions. For the combination of results from different scoring functions, a method called "consensus scoring" has recently been proposed,[61] in which only those compounds are regarded that receive high ranks with two or more scoring functions. A considerable reduction of false positives has been reported with this method. Table 2 shows results for pairwise combination of the four scoring functions discussed individually above. Each table entry consists of three numbers: C_{tot} is the total number of compounds common to the top 5% of two rank lists; C_{act} is the number of active compounds contained therein; I_{act} is the number of active compounds one would have obtained, if one had selected the top ranking C_{tot} compounds of the better individual scoring function. Consensus scoring is successful when the number of false positives is reduced more significantly than the number of active compounds and when C_{act} is greater than I_{act}.

Table 2. Results from Consensus Scoring with Pairs of Four Published Scoring Functions[a]

	FlexX			PLP			DrugScore			PMF		
	C_{tot}	C_{act}	I_{act}	C_{tot}	C_{act}	I_{act}	C_{tot}	C_{act}	I_{act}	C_{tot}	C_{act}	I_{act}
COX-2												
FlexX	382	23										
PLP	149	22	33	382	44							
DrugScore	93	16	13	169	30	34	382	37				
PMF	54	2	7	116	7	29	120	6	10	382	8	
Estrogen Receptor												
FlexX	379	26										
PLP	184	25	22	379	28							
DrugScore	129	20	21	181	22	21	379	25				
PMF	112	22	21	180	25	22	156	21	21	379	29	
p38 MAP Kinase												
FlexX	381	48										
PLP	121	30	20	381	48							
DrugScore	100	16	19	195	18	30	381	23				
PMF	84	16	19	159	19	24	141	11	12	381	30	
Gyrase												
FlexX	378	8										
PLP	111	7	12	378	18							
DrugScore	66	4	3	129	7	12	378	10				
PMF	52	5	5	71	7	8	88	4	7	378	10	
Thrombin												
FlexX	379	45										
PLP	184	31	36	379	34							
DrugScore	117	24	26	174	18	19	379	25				
PMF	52	5	14	95	4	12	114	5	6	379	6	
Gelatinase A												
FlexX	378	23										
PLP	116	6	9	378	6							
DrugScore	89	6	6	212	4	3	378	7				
PMF	50	4	3	89	2	1	88	3	1	378	5	
Neuraminidase												
FlexX	379	25										
PLP	73	4	3	379	5							
DrugScore	28	2	1	148	1	4	379	2				
PMF	93	23	12	94	5	12	84	2	12	379	44	

[a] C_{tot} is the total number of molecules common to the top 5% of both rank lists; C_{act} is the number of active compounds contained therein; I_{act} is the number of active compounds among the C_{tot} top ranking molecules of the better of the compared scoring functions. Diagonal elements are combinations of each scoring function with itself, where C_{tot} is the number of molecules in the top 5% of the database. This varies because the number of inhibitors was different for each target (Table 1).

From Table 2 it can be seen that consensus scoring is generally successful when two scoring functions are combined that perform well individually. For example, the combination of PLP and FlexX for p38 MAP kinase results in 121 common molecules or less than one-third of the top 5% of (381) compounds of either rank list, whereas the number of active compounds decreases to about two-thirds, from 48 to 30. However, for many combinations, I_{act} is higher than C_{act}, because most active compounds are concentrated at the very top of the rank files. In other words, if one knew in advance which scoring function works better for a given target, better performance can be achieved by using this function alone and concentrating on the highest ranking compounds only. In practice, however, consensus scoring

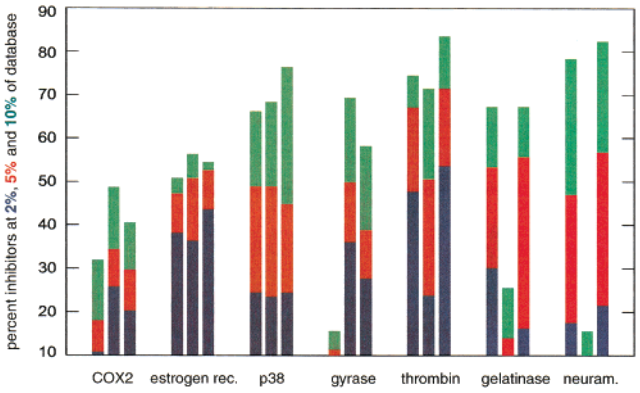

Figure 3. Comparison of the FlexX and PLP scoring functions with the FlexX-PLP combination ScreenScore. For each target, the left column of the triplet shows FlexX results, the middle column PLP results, and the right column results calculated with ScreenScore.

offers more robust results than an individual scoring function. The best general combination is FlexX score and PLP score. We have also investigated the use of three scoring functions instead of two for consensus scoring, but we observed a significant loss of inhibitors in most cases.

Increased Performance through a Combination of Scoring Terms. Even though consensus scoring is a viable approach to increase the efficiency in virtual screening, it is desirable to cover many targets with a single robust scoring function. None of the four scoring functions discussed here performs satisfactorily for all seven targets. Rather, it seems that pairs of scoring functions are complementary to each other. Therefore we sought to find a combination of scoring function terms that would lead to better overall performance rather than peak performance in a few individual cases. It seemed reasonable to combine the localized and directed FlexX hydrogen bond contributions with the ability of PLP to model lipophilic interactions with a simple pair potential approach. The final combined PLP-FlexX combination, which we will call ScreenScore in the following, has the form:

$$\Delta G_{\mathrm{bind}} = F_{\mathrm{match}} + 0.07(F_{\mathrm{lipo}} + F_{\mathrm{ambig}}) + 0.3F_{\mathrm{PLP}} + 1.6n_{\mathrm{rot}}$$

The lipophilic term is thus dominated by the PLP contribution, while the FlexX lipophilic contribution (F_{lipo} and F_{ambig}), being a short-range term, merely stresses surface complementarity. The optimal ΔG_{rot} coefficient is slightly larger than the original FlexX value of 1.4, which reflects the larger absolute score values calculated with ScreenScore. ScreenScore contains no contribution from the FlexX clash penalty function F_{clash}. In Figure 3, ScreenScore is compared with its two ancestors. Its performance is good for all seven targets, whether polar or nonpolar, and in some cases even higher than that of either FlexX or PLP alone. It does not reach the high enrichment for COX-2 or gyrase B that is obtained with PLP or the high enrichment of gelatinase A inhibitors obtained with FlexX at the top 2% of the database, but it successfully balances the properties of FlexX score and PLP score.

ScreenScore certainly does not solve all the problems of fast scoring functions. For example, a compound

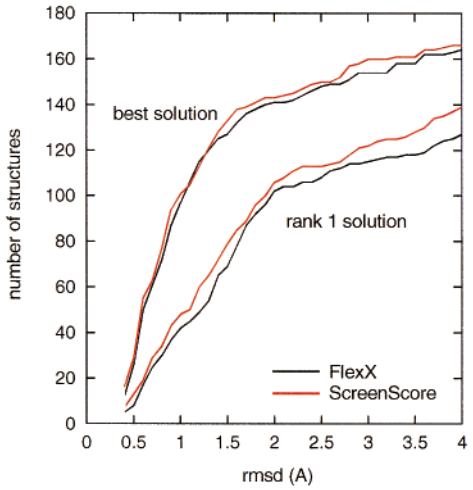

Figure 4. Analysis of docking accuracy for 200 structures from the PDB.[62] The plot shows the numbers of compounds docked better than a given root-mean-square deviation (rmsd) threshold. "Best solution" means that the structure closest to the X-ray structure was chosen regardless of its rank.

forming two hydrogen bonds to the receptor still tends to get a better score than a similar compound forming only one hydrogen bond, even though this may not reflect reality in all situations. Such spurious behavior can only be removed by a more detailed − but also more CPU-intensive − treatment of hydrogen bonds and electrostatic effects.

Database rankings discussed up to this point were calculated by rescoring docking solutions generated with the FlexX scoring function. We have also tested ScreenScore as a combined fitness and scoring function. Overall results remain comparable: Enrichment of inhibitors in the top percentiles of the database drops by about 5−10% for the polar targets neuraminidase, gelatinase A, and gyrase, while for thrombin and p38 MAP kinase it rises by the same amount. For COX-2 and the estrogen receptor, changes are minimal. These results may be explained by the fact that the PLP potential itself is a softer, more error-tolerant function and introduction of any component of PLP into the fitness function will lead to a solution set that satisfies directed interactions less than the FlexX score itself. For polar targets this will lead to some more lipophilic compounds receiving better scores and thus higher ranks than before. Interestingly, when applied as a fitness function on a previously published set of 200 protein−ligand complexes from the PDB,[62] ScreenScore is slightly superior to the original FlexX score (Figure 4) but still inferior to DrugScore, which was tested on a subset of these 200 complexes.[24]

Although the performance of ScreenScore is an improvement in both structure prediction and virtual screening, the above findings show that it generally makes sense to separate fitness and scoring functions from each other. This allows for exaggerating specific binding phenomena such as hydrogen bonds and metal interactions in the docking phase and reducing their weight in the final scoring phase. Such a strategy is in accord with the docking approach realized in FlexX, which focuses on specific interactions throughout the docking phase.

Conclusions

We have tested the performance of four fast scoring functions in seven virtual screening experiments and showed that all four functions have specific shortcomings that reduce their usefulness as general scoring functions for database ranking. Particularly, some functions are more suitable for lipophilic targets and others for polar ones. We then showed that a combination of elements of two scoring functions leads to significantly higher performance than any of the individual functions. An optimized combination of FlexX and PLP scoring functions, called ScreenScore, successfully balances the weight of undirected lipophilic interactions and directed hydrogen bonds. In this study, we have used the original FlexX scoring function to generate docking solutions for each compound. This has proven to be a good choice, since it was shown that emphasizing specific directed interactions in the docking phase, as is done in FlexX, can increase performance in virtual screening. Although this study was performed with FlexX only, we anticipate that the ScreenScore will prove to be a robust and valuable scoring function in combination with other docking engines as well.

Acknowledgment. The authors thank Ingo Muegge for help with the PMF scoring function and Gerhard Klebe and Holger Gohlke for making the DrugScore executable available to us. M.S. thanks his colleagues at Roche Bioscience and Roche Basel for a stimulating and supportive research atmosphere. M.R. thanks Roche Bioscience for funding his research stay in Palo Alto, CA, and for providing a friendly and efficient working environment. The majority of the FlexX extensions were developed in the RELIMO Project funded by the BMBF under Grant 0311 620 and the TTN Dechema Project funded by the EU under Grant IV38463 (parallel version of FlexX).

Supporting Information Available: Ligand input geometries (SMILES format) of all non-Roche compounds listed in Table 1, FlexX parameters for PLP, FlexX and ScreenScore, and an additional Figure highlighting the statistical significance of the docking results in Figure 1. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Blaney, J. M.; Dixon, J. S. A good ligand is hard to find: Automated docking methods. In *Perspectives in Drug Discovery and Design*; Kluwer/Escom: Dordrecht, 1993; Vol. 1, pp 301−319.
(2) Kuntz, I. D.; Meng, E. C.; Shoichet, B. K. Structure-based molecular design. *Acc. Chem. Res.* **1994**, *27*, 117−123.
(3) Stahl, M. Structure-based library design. In *Virtual Screening for Bioactive Molecules*; Schneider, G., Boehm, H.-J., Eds.; VCH: Weinheim, 2000; pp 229−264.
(4) Gane, P. J.; Dean, P. M. Recent advances in structure-based rational drug design. *Curr. Opin. Struct. Biol.* **2000**, *10*, 401−404.
(5) Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. FLOG: A system to select "quasi-flexible" ligands complementary to a receptor of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 153−174.
(6) Abagyan, R.; Trotov, M.; Kuznetsov, D. ICM − A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488−506.
(7) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.
(8) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.
(9) Ewing, T. J. A.; Kuntz, I. D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.* **1997**, *18*, 1175−1189.
(10) McMartin, C.; Bohacek, R. S. QXP: Powerful, rapid computer algorithms for structure-based drug design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333−344.
(11) Burkhard, P.; Hommel, U.; Sanner, M.; Walkinshaw, M. D. The discovery of steroids and other novel FKBP inhibitors using a molecular docking program. *J. Mol. Biol.* **1999**, *287*, 853−858.
(12) Godden, J. W.; Stahura, F.; Bajorath, J. Evaluation of docking strategies for virtual screening of compound databases: cAMP-dependent serine/threonine kinase as an example. *J. Mol. Graphics Mod.* **1999**, *16*, 139−143.
(13) Baxter, C. A.; Murray, C. W.; Waszkowycz, B.; Li, J.; Sykes, R. A.; Bone, R. G. A.; Perkins, T. D. J.; Wylie, W. New approach to molecular docking and its application to virtual screening of chemical databases. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 254−262.
(14) Filikov, A. V.; Monan, V.; Vickers, T. A.; Griffey, R. H.; Cook, P. D.; Abagyan, R. A.; James, T. L. Identification of ligands for RNA targets via structure-based virtual screening: HIV-1 TAR. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 593−610.
(15) Perola, E.; Xu, K.; Kollmeyer, T. M.; Kaufmann, S. H.; Prendergast, F. G.; Pang, Y.-P. Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *J. Med. Chem.* **2000**, *43*, 401−408.
(16) Ajay; Murcko, M. A. Computational methods to predict binding free energy in ligand−receptor complexes. *J. Med. Chem.* **1995**, *38*, 4953−4967.
(17) Böhm, H.-J.; Stahl, M. Rapid empirical scoring functions in virtual screening applications. *Med. Chem. Res.* **1999**, *9*, 445−462.
(18) Tame, J. R. H. Scoring functions: A view from the bench. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 99−108.
(19) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys. J.* **1997**, *72*, 1047−1069.
(20) Rarey, M.; Kramer, B.; Lengauer, T. Multiple automatic base selection: Protein−ligand docking based on incremental construction without manual intervention. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 369−384.
(21) Rarey, M.; Kramer, B.; Lengauer, T. Docking of hydrophobic ligands with interaction-based matching algorithms. *Bioinformatics* **1999**, *15*, 243−250.
(22) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.
(23) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. BLEEP − A potential of mean force describing protein−ligand interactions: I. Generating the potential. *J. Comput. Chem.* **1999**, *20*, 1165−1177.
(24) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(25) Bernstein, F. C.; Koetzle, T. E.; Williams, G. J. B.; Meyer, J., E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; M., T. The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **1977**, *112*, 535−542.
(26) Carter, J. S. Recently reported inhibitors of cyclooxygenase-2. *Exp. Opin. Ther. Patents* **1997**, *8*, 21−29.
(27) Friesen, R. W.; Brideau, C.; Chan, C. C.; Charleson, S.; Deschenes, D.; Dubé, D.; Ethier, D.; Fortin, R.; Gauthier, J. Y.; Girard, Y.; Gordon, R.; Greig, G. M.; Riendau, D.; Savoie, C.; Wang, Z.; Wong, E.; Visco, D.; Xu, L. J.; Young, R. N. 2-Pyridinyl-3-(4-methylsulfonyl)phenylpyridines: Selective and orally active cyclooxygenase-2 inhibitors. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 2777−2782.
(28) Kalgutkar, A. S. Selective cyclooxygenase-2 inhibitors as nonulcerogenic antiinflammatory agents. *Exp. Opin. Ther. Patents* **1999**, *9*, 831−849.
(29) Protherics web site: http://www.protherics.com.
(30) Magarian, R. A.; Overacre, L. B.; Singh, S.; Meyer, K. L. The medicinal chemistry of nonsteroidal antiestrogens: A review. *Curr. Med. Chem.* **1994**, *1*, 61−104.
(31) Fink, B. E.; Mortensen, D. S.; Stauffer, S. R.; Zachary, D. A.; Katzenellenbogen, J. A. Novel structural templates for estrogen-receptor ligands and prospects for combinatorial synthesis of estrogens. *Chem. Biol.* **1999**, *6*, 205−219.
(32) Hanson, G. J. Inhibitors of p38 kinase. *Exp. Opin. Ther. Patents* **1997**, *7*, 729−733.
(33) Wiley, M. R.; Fisher, M. J. Small-molecule direct thrombin inhibitors. *Exp. Opin. Ther. Patents* **1997**, *7*, 1265−1282.
(34) Sanderson, P. E. J.; Nayor-Olsen, A. M. Thrombin inhibitor design. *Curr. Med. Chem.* **1998**, *5*, 289−304.
(35) Beckett, R. P.; Whittaker, M. Matrix metalloproteinase inhibitors 1998. *Exp. Opin. Ther. Patents* **1998**, *8*, 259−282.

(36) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(37) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.
(38) Clark, M.; Cramer III, R. D.; Van Opdenbosch, N. Validation of the general purpose Tripos 5.2 force field. *J. Comput. Chem.* **1989**, *10*, 982−1012.
(39) *Sybyl* molecular modeling software, version 6.2, Tripos Associates, St. Louis, MO, 1994.
(40) Sadowski, J.; Rudolph, C.; Gasteiger, J. Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537−547.
(41) Sadowski, J.; Schwab, C. H.; Gasteiger, J. *Corina*, version 2.1, Molecular Networks GmbH Computerchemie, Erlangen, 1998.
(42) pKalc, CompuDrug Inc., South San Francisco, CA.
(43) WDI: World Drug Index 2/96, Derwent Information, 1996.
(44) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared near neighbors. *IEEE Trans. Computers* **1973**, *C22*, 1025−1034.
(45) Daylight Chemical Information Systems, Daylight Inc., Mission Viejo, CA.
(46) Gerber, P. R. Topological pharmacophore description of chemical structures using MAB-force-field derived data and corresponding similarity measures. Proceedings of IV Girona Seminar on Molecular Similarity, Girona, Spain, July 5−7, 1999.
(47) 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.
(48) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: Conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *2*, 317−324.
(49) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Reijto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering common failures in molecular docking of ligand-protein complexes. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 731−751.
(50) The FlexX software is available for various UNIX platforms. More information about FlexX can be obtained from: http://cartan.gmd.de/FlexX or from M.R. (rarey@gmd.de).
(51) Sunderarm, V.; Dongarra, J.; Geist, A.; Manchek, R. The PVM concurrent computing system: Evolution, experiences and trends. *Parallel Computing* **1994**, *20*, 531−547.
(52) PVM home page: http://www.epm.ornl.gov/pvm/.
(53) Klebe, G.; Mietzner, T. A fast and efficient method to generate biologically relevant conformations. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 583−606.
(54) Klebe, G. Toward a more efficient handling of conformational flexibility in computer-assisted modelling of drug molecules. In *De Novo Design*; Müller, K., Ed.; Escom: Leiden, 1995; Vol. 3, pp 99−114.
(55) Hoffmann, R. W.; Stahl, M.; Schopfer, U.; Frenking, G. Conformation design of hydrocarbon backbones. *Chem. Eur. J.* **1998**, *4*, 559−566.
(56) The question often arises what percentage of the ranked database is of interest in a virtual screening run. This obviously depends on the database size and ratio between active and inactive compounds. Consider a test database of 100 active and 1000 inactive compounds. In the ideal case of database ranking, the active compounds would occupy the top 100 rank positions or the top 10% of the database. Repeating the same run with a larger database of 100 000 inactive compounds would ideally place the active compounds on the top 0.1% of the database. Even though the ideal case is never reached, this kind of analysis pertains to real-world applications. In the present case about 75 compounds constitute the top 1% and the number of active compounds is on the same order of magnitude. Ranking efficiency of the different scoring functions is in this case best compared at the 5% level. The effect of database size can also be studied by comparing Figure 1 to the figure in the Supporting Information, for which the WDI subset has been partitioned in 7 sets of 1000 compounds.
(57) Muegge, I. The effect of small changes in protein structure on predicted binding modes of known inhibitors of influenza virus neuraminidase: PMF-scoring in DOCK4. *Med. Chem. Res.* **1999**, *9*, 490−500.
(58) Ha, S.; Andreani, R.; Robbins, A.; Muegge, I. Evaluation of docking/scoring approaches: A comparative study based on MMP3 inhibitors. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 435−448.
(59) Stahl, M. Modifications of the scoring function in FlexX for virtual screening purposes. *Persp. Drug Discovery Des.* **2000**, *20*, 83−98.
(60) Press, W. H.; Teukolsky, S. A.; Vettering, W. T.; Flannery, B. P. *Numerical Recipes in C*; Cambridge University Press: Cambridge, 1992; p 408.
(61) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.
(62) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FlexX incremental construction algorithm for protein−ligand docking. *Proteins* **1999**, *37*, 228−241.

JM0003992