Docking Ligands into Flexible and Solvated Macromolecules. 4. Are Popular Scoring Functions Accurate for this Class of Proteins?

Pablo Englebienne and Nicolas Moitessier*

Department of Chemistry, McGill University, 801 Sherbrooke St. W, Montreal, Quebec, Canada H3A 2K6

Received November 25, 2008

In our previous report, we investigated the impact of protein flexibility and the presence of water molecules on the pose-prediction accuracy of major docking programs. To complete these investigations, we report herein a study of the impact of these two aspects on the accuracy of scoring functions. To this effect, we developed two sets of protein/ligand complexes made up of ligands cross-docked or cocrystallized with a large variety of proteins, featuring bridging water molecules and demonstrating protein flexibility. Efforts were made to reduce the correlation between the molecular weights of the selected ligands and their binding affinities, a major bias in some previously reported benchmark sets. Using these sets, 18 available scoring functions have been assessed for their accuracy to predict binding affinities and to rank-order compounds by their affinity to cocrystallized proteins. This study confirmed the good and similar accuracy of Xscore, GlideScore, DrugScore^{CSD}, GoldScore, PLP1, ChemScore, RankScore, and the eHiTS scoring function. Our next investigations demonstrated that most of the assessed scoring functions were much less accurate when the correct protein conformation was not provided. This study also revealed that considering the water molecules for scoring does not greatly affect the accuracy. Finally, this work sheds light on the high correlation between scoring functions and the poor increase in accuracy one can expect from consensus scoring.

INTRODUCTION

As the time frame and costs of the traditional drug discovery approach are ever-increasing, computational/virtual approaches arose as promising techniques in the medicinal chemist toolkit. In particular, docking-based virtual screening (VS) methods are becoming increasingly popular as a fast, cost-effective alternative or complement to classical high throughput screening (HTS).[1] The purpose of docking methods in this context is 2-fold: i. prediction of the binding mode of a ligand to a given biologically relevant target receptor, enzyme or nucleic acid and ii. prediction of the binding affinity of the complex formed.[2] To carry out these two tasks, docking programs rely on two major components: a conformational search algorithm and a scoring function. The former samples the translational, rotational, and conformational space of the complex, and several approaches that accurately dock flexible ligands to proteins have been disclosed.[3,4] As soon as a putative binding mode of the ligand—referred to as a pose—is proposed, its binding affinity must be predicted. Scoring functions often provide a fast evaluation of the free energy of binding during and after the conformational sampling stage.[5] Despite the intense effort in the area of docking methods in recent years, the moderate performance of scoring functions revealed by independent comparative studies remains the chief issue to address in the improvement of docking methods,[2] and strategies have been developed to increase the identification of actives compounds.[6]

Commonly used scoring functions can be classified into force-field based, knowledge-based, and regression-based, even though some functions may combine two approaches (e.g., regression and FF-based AutoDock scoring function). Several issues need to be considered when assessing or selecting scoring functions for docking and scoring of potential drugs. First, force fields (e.g., AMBER in DOCK scoring function) are known to overestimate the binding affinities, and the calculated intermolecular energy values need to be scaled down for more accurate predictions. An example is the Linear Interaction Energy (LIE) method.[7] In addition, force fields only approximate the enthalpic interaction energy, disregarding some contributions to the free energy of binding such as entropy and solvation. Second, knowledge-based potentials are developed from statistical analysis of protein/ligand complexes regardless of their affinities. Third, regression-based scoring functions, also called empirical scoring functions, are also trained against a set of protein/ligand complexes which are related to known 3D structures and binding affinities. Clearly, the scoring functions from the last two categories are strongly dependent on the training set used to derive them and rely on the accuracy of the binding affinity data, crystallographic experiments, model fitting, and transferability of the parameters to other complexes not present in the training sets.[8] To partially address this issue, large data sets (i.e., the whole PDB database) can be used. However, as these training sets do not contain compounds that are too large to fit into the binding site or otherwise inactive, the potentials derived from these sets may fail to discriminate inactive compounds from actives thus leading to the occurrence of false positives.

We have recently reported the development of FITTED (versions 1.0, 1.5, and 2.6), a docking program accounting for protein flexibility and essential water molecules.[9−11] In parallel, we have reported the evaluation of the impact of

* Corresponding author e-mail: nicolas.moitessier@mcgill.ca.

10.1021/ci8004308 CCC: $40.75
Published on Web 05/15/2009

ligand and protein structures and presence of water molecules on the pose prediction accuracy of major docking programs.[11] In order to complement this previous study and later develop a scoring function that accounts for these two aspects, we have investigated the impact of protein flexibility and water molecules on the accuracy of several commonly used scoring functions including our current version of RankScore.[11] Thus we report herein the development of two sets of protein/ligand complexes, their selection criteria, and their preparation as well as their use as testing sets for the evaluation of 18 commonly used scoring functions.

METHODS

Training Set Selection Criteria. The accuracy of a scoring function is largely dependent on the training set used to calibrate it. It has been reported that several commonly used scoring functions are less accurate than the ligand molecular weight (MW) used as a descriptor.[12,13] Indeed, a close look at training sets used in the development and evaluation of scoring functions shows that there is sometimes a strong correlation between binding affinities and MW, an artifact that should be considered. It is well-known that truncating a large ligand often results in a significant loss of binding affinity, a property clearly captured by several scoring functions. However, as VS is often carried out with libraries of druglike, leadlike, and even fragmentlike molecules with similar molecular weights, these scoring functions do not perform as well. The current challenge is therefore to develop a scoring function able to discriminate between actives and inactives of similar sizes. Some training sets also lack diversity in protein and ligand structures. For instance, scoring functions may be developed from training sets excluding metalloenzymes and/or highly polar enzymes (e.g., neuraminidase), therefore being poorly transferable to these classes of targets. In addition, most of these sets are developed from crystal structures and therefore trained to perform well in self-docking experiments. This is a significant limitation as cross-docking is a more realistic experiment simulating a VS situation. More recently, Verdonk and coworkers have reported a training set carefully prepared using strict criteria.[14,15] In order to evaluate the state-of-the-art in the development of scoring functions, we propose herein to report our training/testing sets that follow a number of restrictions and conditions:

i. The training set should be large enough to be statistically relevant. We targeted a minimum of 200 complexes in order to guarantee an accurate evaluation of existing scoring functions and a good predictivity of the scoring function that will be eventually developed.[16]

ii. The ligands should be as diverse as possible (both in shape, bioactivity, and functional groups) to assess the transferability of scoring functions and eventually ensure transferability of the scoring function to be developed.

iii. The ligand molecular weight should be higher than 250 and not exceed 700.

iv. The affinity of the ligand toward the cocrystallized receptor should be known.

v. Crystal structures should be available at a good resolution (≤2.5 Å). Although this criterion is not strict enough to evaluate the "quality" of the complexes, it is easily accessible.[17]

vi. Some proteins should appear more than 5 times in this set so that cross-docked structures can be considered.

vii. Proteins with both hydrophobic and hydrophilic binding sites should be included.

viii. Different aspects of protein−ligand binding, such as water-mediated binding and metal binding should be represented.

ix. Correlation between ligand molecular weight and binding affinities should be as small as possible.

x. Metal-containing and covalently bound ligands should not be included as they may be poorly defined in scoring functions and would require specific terms.

xi. Binding affinities should cover a range as wide as possible without overweighting one range of affinities and should include as many poor binders as possible.

Due to these many criteria, the set developed herein is very different from the set we have previously used to assess docking programs.[11]

Correlation Metrics and Statistical Significance. Most commonly, the square of Pearson's correlation coefficient (r^2) and Spearman's ρ are used to assess the correlation between observed and predicted affinities. While r^2 is the traditional correlation metrics, measuring the correlation between experimental binding affinities and *scores*, ρ is a nonparametric measure of the correlation between the *ranked lists* of experimental binding affinities and predicted scores. A ρ of ±1.0 corresponds to a perfect match between the two ranked lists (but a negative ρ indicates an inverse order for one of the lists), while a value of 0.0 is consistent with random ordering. As suggested by Nicholls and Jain and a reviewer of this manuscript, we also considered Kendall's tau (τ) as an alternative to Spearman's for the assessment of the rank-ordered correlation.[17] Kendall's has the advantage of being more robust than Spearman's, while also being easier to interpret, as it is an estimate of the probability of having the same trend between two sets of ranks. We used the bootstrap technique to assess the statistical significance of the correlation coefficients: random subsets of the data set were drawn, allowing for duplicates, and the correlation of each subset was calculated. The range containing 95% of the values is taken as the confidence interval for the given descriptor (either ρ or τ).

Training Set Preparation. A meticulous preparation of the complexes was believed to be essential to provide objective results. As described above, the ligands were chosen to be varied in shape, size, bioactivity, and functional groups, with known activity toward a given receptor. To follow the criterion vi, a careful selection of complexes from the PDB led to the selection of 58 complexes of five highly studied proteins: HIV-1 protease, thrombin, trypsin, and matrix metalloproteases (MMP-3 - stromelysin 1 and MMP-8 - collagenase 2). This set was next expanded through the addition of complexes from the PDBBind database,[18,19] taking into account the chemical diversity of the ligands (criteria ii and x), an even distribution of the binding affinity of the complexes (criterion xi) and a proper selection of proteins (criteria vii and viii). After filtering out the complexes that were not well characterized (i.e., missing residues, multiple ligands in binding site), a set of 223 complexes was obtained. In order to meet criterion ix, another 14 complexes were removed and led to a final set of 209 complexes with 82 proteins represented (Table 1, see also

Table 1. Proteins Represented More than Once[b]

proteins	number of ligands	sets
thrombin	22	1, 2
trypsin	13	1, 2
HIV-1 protease	10	1, 2
MMP-8	8	1, 2
factor Xa	7	1, 2
purine nucleoside phosphorylase	6	1[a]
scytalone dehydratase	6	1, 2
urokinase-type plasminogen activator	6	1, 2
carbonic anhydrase	5	1, 2
MMP-3	5	1, 2
PTP-1b	5	1, 2
acetylcholinesterase	4	1
neuraminidase	4	1
retinoic acid receptor gamma-1	4	1
xylanase beta-1	4	1
2,2-dialkylglycine decarboxylase	3	1
cyclin dependent kinase 2	3	1
glutathione s-transferase	3	1
ribonuclease a	3	1
thymidylate synthase	3	1
carboxypeptidase	2	1
orotidine 5′-monophosphate decarboxylase	2	1
serine/threonine-protein kinase chk1	2	1
sex hormone-binding globulin	2	1

[a] The PNP proteins were from different species and thus were not included in Set 2. [b] Set 1 is the self-docking set, while Set 2 includes the cross-docked structures. The complete sets are given as Supporting Information (Tables S1 and S2).

Supporting Information Table S1). This first set—referred to as Set 1—was used to evaluate SFs in a self-docking situation as described below. Proteins for which 5 or more structures were found were further processed. For these systems, protein and ligand structures were swapped to generate cross-docked structures. This second set of nearly 1000 complexes was used to evaluate SFs in a cross-docking context (Set 2).

As shown in Figure 1 (and Table 3), only a little correlation between MWs and affinities was obtained ($r^2 = 0.109$; $\rho = 0.330$; $\tau = 0.230$). In addition, the values of affinity constants span 9 orders of magnitude, with an even distribution in the 0.1 μM−0.1 nM range. Unfortunately a very low number of millimolar ligands fulfilling the constraints defined above were found in the PDB, and this extreme range of activities is underrepresented (criterion ix). In fact, in order for a crystal structure of a protein/ligand complex to be solvable, a ligand needs to present a measurable binding affinity. In addition, crystal structures with highly active compounds provide more information to medicinal chemists and are prioritized by crystallographers.

Careful manipulations were necessary to set up the complexes for calculations. For most of the steps, the preparation was not fully automated in order to reduce potential errors. First, critical water molecules were carefully selected. To do so, crystallographic water molecules were removed unless they were involved in at least 3 hydrogen bonds with the ligand and the protein simultaneously. Next, ligand bond orders (not present in the source PDB files) were properly set, hydrogens were added, and atom types and partial charges were assigned. Special attention was given to the protonation state of ionizable groups in the complexes. For instance, the catalytic dyads in most aspartyl proteases (such as HIV-1 protease) are required to be monoprotonated for the catalytic mechanism to proceed.[20] However, X-ray crystallography and modeling studies indicate that fully protonated states are also observed with some diol ligands.[21] Klebe and co-workers suggested that ligands with an ammonium group facing the catalytic dyad might stabilize the dideprotonated state.[22] Histidine protonation was also carefully assigned by optimizing the hydrogen bond network with neighboring residues. On the ligand side, reasonable protonation states of ionizable groups were assumed. The next step was the full optimization of the hydrogen positions and ligand position (see the Experimental Section) through energy minimization. Initial attempts to fully relax the complexes led to unreliable structures often far from the crystal structure. Freezing the protein and water heavy atoms and constraining ($K = 5$ kcal/mol•Å) ligand heavy atoms was necessary to restrict large motion of some of the ligands, thus keeping the poses close to the experimentally observed ones. As pointed out by a reviewer, the accuracy of scoring functions may be dependent on the method of preparation of the systems. We assessed the scoring functions on unoptimized (i.e., raw PDB files) structures and observed poorer predictions. In the following sections, we will consider only the optimized poses.

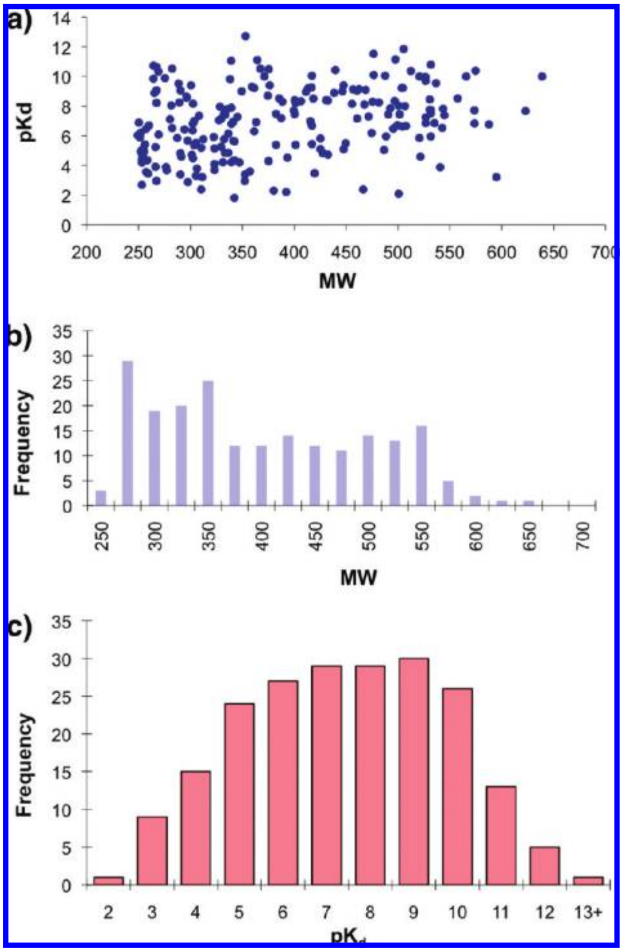


Figure 1. Properties of the training set: a) profile of molecular weight and binding affinity dependence; b) distribution of ligand MW; and c) distribution of ligand binding affinities.

Cross-Scoring Training Set. To explore the sensitivity of scoring functions to the protein conformation, we decided to score every ligand/protein combinations for proteins in set 2 (see Table S2, Supporting Information). First, all complexes within a family were superimposed by aligning

Table 2. Selected Scoring Functions Used in This Study

scoring function	implementation	class	training sets used for development[a]
ChemScore	Sybyl	empirical	82 complexes in 5 classes
DockScore	Sybyl	FF-based	no training set
DrugScore^{CSD}	standalone	knowledge-based	28642 small molecules
DrugScore^{PDB}	standalone	knowledge-based	6026 complexes
eHiTS SF	eHiTS	knowledge-based	133 complexes + extended training set, protein class-specific
FlexXScore	Sybyl	empirical	45 complexes (LUDI SCORE1)
GlideScore	Glide	empirical	82 complexes in 5 classes (ChemScore)
GoldScore	Sybyl	FF-based	no training set
Hammerhead	Cerius2	empirical	34 complexes
LigScore1	Cerius2	empirical	50 complexes
LigScore2	Cerius2	empirical	112 complexes
PLP1	Cerius2	empirical	3 complexes (DHFR, FKBP, HIV-1P)
PLP2	Cerius2	empirical	3 complexes (DHFR, FKBP, HIV-1P)
PMF	Sybyl	knowledge-based	697 complexes
PMF	Cerius2	knowledge-based	697 complexes
RankScore	Fitted	FF-based	50 BACE-1 inhibitors and 4 complexes
Surflex SF	Surflex	empirical	34 complexes
XScore	standalone	empirical/consensus	200 complexes

[a] These training sets are those reported. In some cases, improved versions have been released but the training sets used to derive them have not been reported.

Table 3. Accuracy of the Scoring Functions on the Complete Set 1 and on Two Reduced Sets Compared to Previously Reported Data[a]

	entire set (209 complexes)			5 outliers removed (204 complexes)			metalloenzymes removed (188 complexes)			from refs 12 and 13
	r^2	ρ	τ	r^2	ρ	τ	r^2	ρ	τ	ρ
ChemScore (Sybyl)	**0.197**	**0.469**	**0.324**	**0.276**	**0.547**	**0.374**	**0.224**	**0.510**	0.351	0.431/0.432
DockScore (Sybyl)	0.098	0.342	**0.231**	**0.166**	0.421	**0.281**	**0.123**	**0.389**	0.263	0.475/0.476
DrugScore^{CSD}	**0.176**	**0.427**	**0.292**	**0.241**	**0.498**	**0.337**	**0.251**	**0.472**	**0.320**	**0.624**
DrugScore^{PDB}	**0.121**	**0.383**	**0.264**	**0.169**	**0.455**	**0.309**	**0.148**	**0.431**	**0.296**	**0.587/0.589**
eHiTS SF	0.208	**0.421**	**0.296**	**0.228**	**0.513**	**0.353**	**0.228**	**0.446**	0.315	-
FlexXScore	0.038	0.195	0.133	0.094	0.272	0.183	0.046	0.216	0.149	0.283/0.287
GlideScore	**0.114**	**0.378**	**0.26**	**0.176**	**0.454**	**0.309**	**0.365**	**0.423**	0.292	-
GoldScore (Sybyl)	**0.169**	**0.434**	**0.295**	**0.235**	**0.507**	**0.342**	**0.205**	**0.487**	**0.330**	**0.569/0.570**
Hammerhead (Cerius2)	**0.115**	0.345	**0.237**	**0.166**	0.415	**0.282**	**0.144**	**0.383**	0.262	-
LigScore1 (Cerius2)	0.011	0.112	0.074	0.030	0.176	0.116	0.017	0.141	0.094	-
LigScore2 (Cerius2)	0.096	0.332	**0.225**	**0.141**	0.401	**0.269**	**0.122**	**0.367**	0.248	0.363/0.368
PLP1 (Cerius2)	**0.139**	**0.387**	**0.262**	**0.190**	**0.453**	**0.306**	**0.173**	**0.434**	**0.292**	**0.592/0.593**
PLP2 (Cerius2)	**0.116**	**0.364**	**0.248**	**0.185**	**0.443**	**0.299**	**0.122**	**0.400**	**0.273**	**-**
PMF (Sybyl)	0.000	0.012	0.011	0.011	0.073	0.051	0.000	0.023	0.021	-
PMF (Cerius2)	0.050	0.212	0.141	0.093	0.284	0.186	0.054	0.236	0.156	0.369/0.370
RankScore	**0.148**	**0.418**	**0.298**	**0.216**	**0.503**	**0.353**	**0.177**	**0.458**	0.328	-
Surflex SF	**0.143**	**0.409**	**0.274**	**0.161**	**0.476**	**0.317**	**0.167**	**0.448**	0.301	-
XScore	**0.239**	**0.526**	**0.365**	**0.320**	**0.605**	**0.416**	**0.259**	**0.566**	**0.392**	**0.660/0.660**
MW	0.109	0.349	0.230	0.117	0.422	0.277	0.110	0.350	0.229	0.560

[a] The AutoDock scoring function has not been included in the present study. In bold if better than MW.

the alpha carbons on the proteins. Then new complexes were constructed by swapping ligand and protein structures. To relieve any undesired clashes between ligand, protein, and waters in these manually docked structures, a local conformational search of the ligand was performed on each complex. For this purpose, we have implemented a local search mode in our docking program FITTED. This conformational search mode was designed to optimize the intermolecular interactions among ligand, protein, and waters, without greatly disturbing the initial conformation of the ligand. The rmsd (vs crystal structure) of the resulting ligand conformations was below 1.5 Å in 84% of the cases, while in the case of ligands in their cognate receptor, all poses were below 1.5 Å rmsd from the experimentally observed ones. In order to evaluate the impact of water molecules in the final scores, the cross-docked structures were optimized using the local search algorithm keeping the water molecules and then scored with or without the water molecules present in the binding site. Alternatively, the ligand poses were also optimized with no waters included and then scored.

RESULTS AND DISCUSSION

Accuracy of the Selected Scoring Functions on the Entire Set. With the MW-unbiased set (Set 1) in hand, we evaluated the accuracy of well-established scoring functions. Thus, two implementations of PMF (Cerius2 and CScore),[23] PLP1/2,[24] LigScore1/2,[25] ChemScore,[26] GoldScore,[27] XScore,[16] six different versions of DrugScore,[13] GlideScore,[28] the eHiTS scoring function,[29] the Surflex scoring function[30] and its predecessor, the Hammerhead scoring function, and our first version of RankScore[31] were used to predict the binding affinities of the ligand set (Table 2). These scoring functions have been derived using different training sets, as shown in the rightmost column of Table 2. At this stage, we can only discuss the reported training sets used to derive these scoring functions, although we are aware that some changes may have been made to the latest releases. Knowledge-based scoring functions (PMF, DrugScore) feature the largest training sets, although in these cases the training sets are not used to adjust coefficients by regression. Of the disclosed training sets used by empirical scoring functions, XScore features the largest training set, followed by the eHiTS scoring function. The latter includes an additional tunable property: the set of regression coefficients has been calibrated against multiple subsets of the whole Protein Data Bank, and a specific scoring function is selected for each complex to be scored.

We next looked at the overlap between these sets and our sets. In fact, our Set 1 has little in common with the previous training sets used in the development of these regression-based scoring functions: either one (FlexXScore, Hammerhead/Surflex), three (LigScore), five (ChemScore), or seven (XScore) complexes were shared. In addition, although the sets used to train DrugScore and the eHiTS scoring functions are larger and may share a greater number of structures with our Set 1, the use of most of the available PDB structures in their training dilutes the effect of a single one.

The relative ranking of the ligands for their binding affinity were also computed. The correlation between experimental and calculated data was next evaluated and compared to the reported accuracies computed with a more MW-biased testing set (Wang's set).[12] In the upcoming sections, we may suggest that a scoring function is better than another one although in many cases the differences are within the error bars.

As previously noted, some reported biological activities can be highly dependent on the experimental conditions of the assay.[2] In order to remove part of the noise due to experimental errors, we removed the worse 5 predictions for each scoring function. A close look at these reduced sets showed that 1bn4, 1bnn, 1m0n, 1m0o, 1m0q, and 1osv are the most frequently found within the selected 5 outliers, and their activities were poorly predicted with all the scoring functions. We also looked at a reduced set with no metalloenzymes as it is known that metal coordination is often poorly scored and/or scoring functions not trained on metalloenzymes.[32]

The correlation between the scores computed with the selected scoring functions and the experimental binding affinities is low (see Table 3), with XScore (r^2=0.320) being the most predictive followed by ChemScore (r^2=0.276), DrugScore^{CSD} (r^2=0.241), GoldScore (r^2=0.235), the eHiTS scoring function (r^2=0.228), and RankScore (r^2=0.216). However, although this correlation is a good indicator of the potential of scoring functions, we—as others—[33] believe that the correlation between the ranked lists using either Spearman ρ or Kendall τ provides a more useful indication of the predictive power of scoring functions in the context of virtual screening. ρ and τ also identify XScore (ρ=0.606, τ=0.416), as the most accurate scoring functions followed by ChemScore (ρ=0.547, τ=0.374), the eHiTS scoring function (ρ=0.487, τ=0.353), RankScore (ρ=0.482, τ=0.353), GoldScore (ρ=0.543, τ=0.342), DrugScore^{CSD} (ρ=0.559, τ=0.337), and PLP1 (ρ=0.516, τ=0.306). Clearly, r^2, ρ, and τ identified the same scoring functions as the most accurate with insignificant changes in the ranking. Among these top-scoring functions, the eHiTS scoring function is the least sensitive to the presence of outliers, while GlideScore and LigScore1 are very sensitive to the presence of metalloenzymes and outliers, respectively. When correlations were computed on nonmetal containing enzymes, XScore remains the most accurate scoring function.

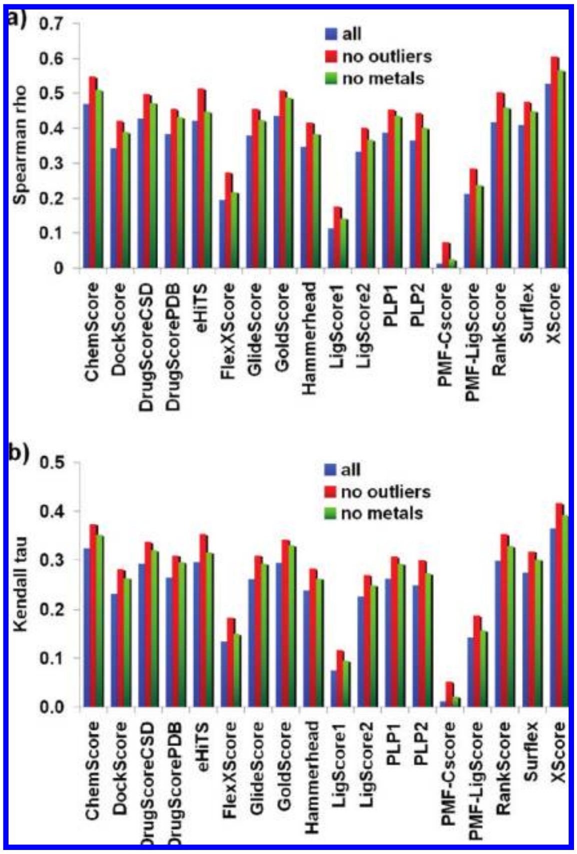


Figure 2. Spearman (a) and Kendall (b) coefficients for three subsets of the training set: in blue, 209 complexes (whole set); in red, 5 outliers removed (204 complexes); and in green, all transition metal-containing proteins removed (188 complexes).

Interestingly, the collected data summarized in Table 3 and Figure 3 reveal the change in accuracy when going from Wang's set to ours and also confirm the variety of accuracies measured with this set of scoring functions. Comparing Wang's data to our data indicates that the accuracy of all the scoring functions but ChemScore, LigScore, and FlexXScore were affected by the reduced dependence on MW. In fact, the major difference between the scoring functions accuracy on our set and on Wang's set is the data collected with ChemScore and LigScore2. These scoring functions were previously found to be poorer than MW used as a descriptor, while in the present work they were found to perform better, clearly capturing some of the binding aspects other than ligand size. It is not clear why these two scoring functions perform better with our more challenging set.

When considering the error bars arising from bootstrapping the data set with either correlation coefficient (Figure 4), XScore appears to be better than most of the other scoring functions by a marginal value, while LigScore1, FlexXScore, and both implementations of PMF poorly rank-ordered the set by scores. As a matter of fact, with the CScore implementation of PMF and LigScore1, one cannot rule out the possibility of a chance correlation, as a null correlation coefficient is included in the respective interval of confidence for both scoring functions.

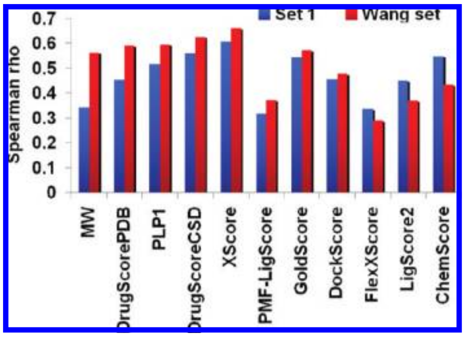


Figure 3. Comparison of the Spearman correlation for ten scoring functions. Blue columns indicate the correlation of the “no outliers” set (204 complexes); red bars denote the Spearman coefficients as reported by Wang et al.[12]

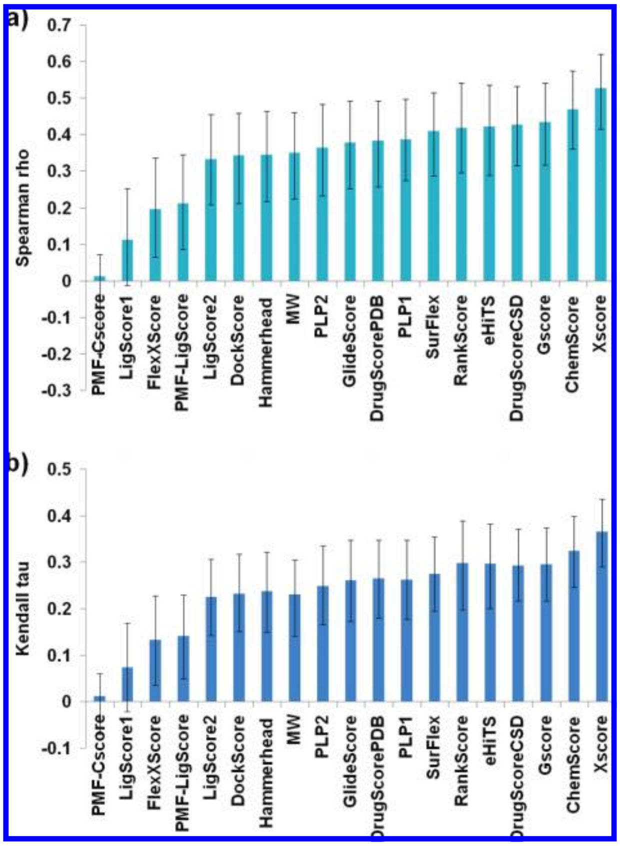


Figure 4. Comparison of Spearman (a) and Kendall (b) coefficients for the scoring functions considered. Error bars calculated through the bootstrap method (see the Methods section).

Although the eHiTS scoring function and DrugScore have been trained on large sets, their accuracy does not exceed the ones observed with some regression-based scoring functions. We believe that the uneven number of protein complexes represented in the Protein Data Bank (e.g., several hundreds of thrombin/ligand complexes, 1% of the PDB) may lead to the overtraining of these scoring functions for these specific proteins. As a result, increasing the number of structures in the training set if not accompanied by an increase in the diversity is not expected to improve the training and transferability of the developed scoring functions.

Accuracy of the Selected Scoring Functions within Protein Classes. A scoring function is of interest for VS applications (e.g., hit compound discovery) only if it can discriminate between active and inactive compounds in a given library against a particular target. In structure−activity relationship (lead optimization), a scoring function should rank compounds with subtle structural changes. To evaluate this ability, one has to investigate families of protein/ligand complexes. We used the collected data and extracted the scores obtained with serine proteases (thrombin, trypsin, and factor Xa, 42 complexes), HIV-1 protease (11 complexes), and thrombin alone (22 complexes) (Figure 5). Although XScore was found to be the most accurate on the entire set, RankScore, GoldScore and DockScore were found to be the most predictive with the selected serine proteases inhibitors and the only 3 scoring functions that are likely more predictive than the MW descriptor. Similarly, ChemScore demonstrates very good accuracy with HIV-1 protease with $\tau = 0.62$, while it is not predictive at all with thrombin ($\tau < 0.1$) or the serine proteases ($\tau = 0.27$). It is worth noting that affinities of these subset ligands are more MW-dependent than the complete set and that only three of the assessed functions (RankScore, GoldScore, and DockScore) were consistently more accurate than MW (although within the computed errors) used as a descriptor with these families of proteins. In contrast to Wang’s report, we found XScore and ChemScore reliable with HIV-1 protease. In fact, many scoring functions were at least marginally predictive against this protein class, while MW was likely a chance correlation. We relate this better accuracy to the care taken to prepare the systems: the protonation state of this aspartic protease was carefully assigned to each of the complexes and the essential water molecule was kept when necessary (see Experimental Section). These two features ensured that the scores reflect the binding energies of the complexes. Interestingly, FlexXScore was found among the most accurate with HIV-1 protease and the least accurate with the thrombin and serine protease ligands, further demonstrating the poor transferability of some scoring functions and the need for a broad set when carrying a comparative study or developing a scoring function or for protein-specific scoring functions. When we attempted to expand this analysis to other families of proteins, we found that their representation in our set prevented us from making statistically significant predictions (that is, discarding the possibility of a chance correlation).

Hydropathicity and Accuracy. When a computational medicinal chemist starts a docking study, a recurrent question is always about the selection of the docking/scoring program best suited for the ongoing study. To partially answer this question, we previously investigated the accuracy of docking programs as a function of the hydropathicity of the proteins,[11] while herein we looked at the accuracy of scoring functions. The hydropathicity of the binding sites was evaluated by considering all the residues within 6.0 Å of the ligands on a combination of the Hopp-Woods, Kyle-Doolittle, and Grantham scales of hydrophobicity.[34−36] From the 209 initial complexes, the 85 most hydrophobic and the 93 most hydrophilic were selected (Figure 6). Three classes of scoring functions emerged from this study. First, XScore performed well with both hydrophobic and hydrophilic protein classes. Second, RankScore, DrugScore, PLP1, and the eHiTS scoring functions were found to be more predictive with hydrophilic proteins. This is an indication for further improvement of Fitted, as we have also found that Fitted docks small molecules very accurately to hydrophilic proteins but performed poorly with hydrophobic proteins.[37] These

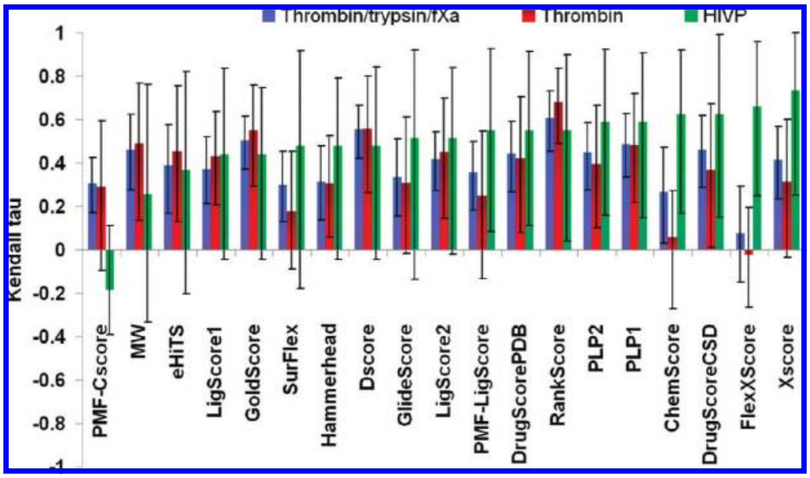


Figure 5. Accuracy (τ) of the scoring functions on 3 subsets of set 1.

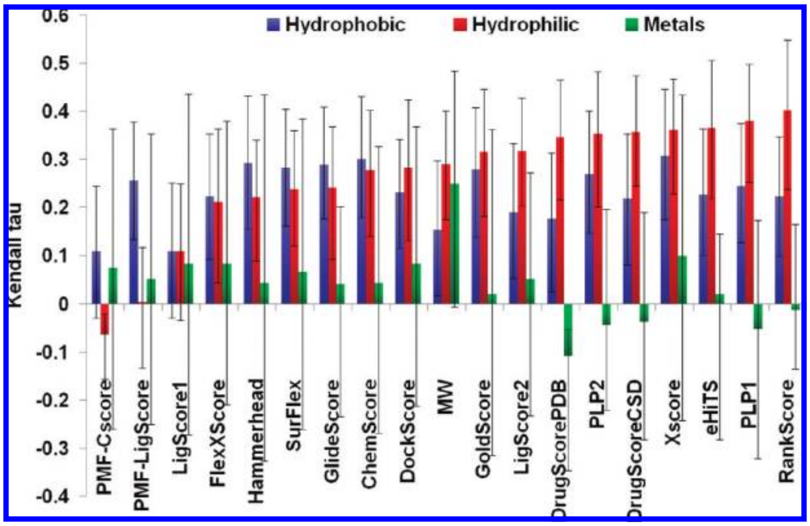
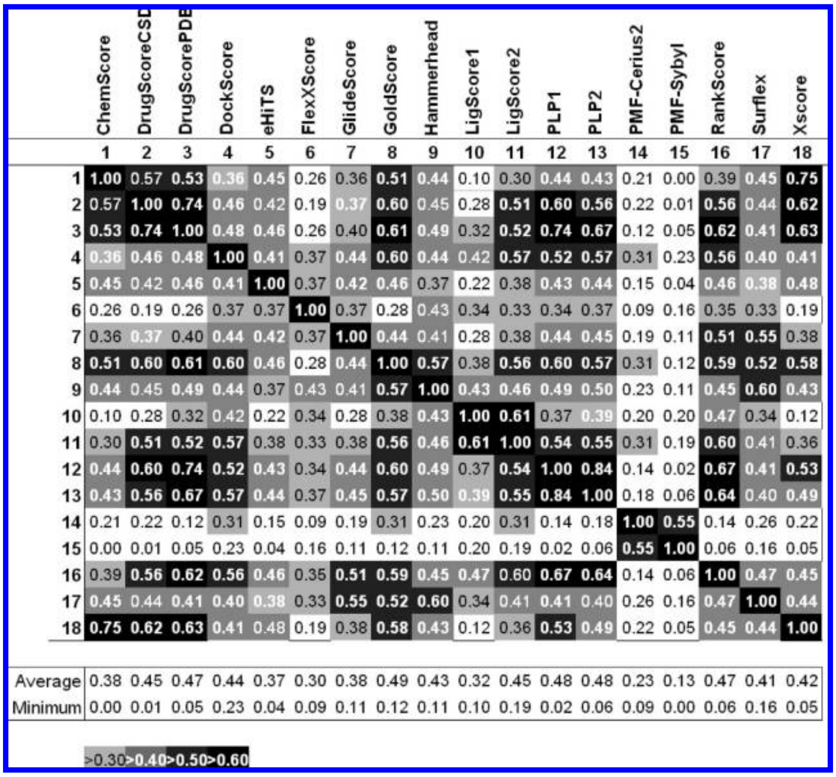


Figure 6. Accuracy (Kendall tau correlation) of the scoring functions on three subsets of Set 1, classified by the hydropathicity of the binding sites. Error bars calculated with the bootstrap method.

observations are most likely related to the prediction of hydrophobic interactions on one side and hydrogen bonds and ionic interactions on the other side, the latter being easier to identify and quantify. At the other end of the spectrum, GlideScore is more accurate with hydrophobic proteins. Nevertheless, as these subsets demonstrate different dependence between MW and affinities, drawing conclusions might require considering larger sets. Finally, the data collected for the 23 metalloenzymes do not allow one to rule out a chance correlation for none of the scoring functions, except for DrugScore$^{\mathrm{PDB}}$ which exhibited an unexpected modest anticorrelation. These observations clearly confirmed that current scoring functions are not reliable when metal chelation is the key interaction, as none of the scoring functions considered were more predictive than the ligand molecular weight.

Scoring Function Correlation. As discussed above, the accuracy of scoring functions varies significantly from one protein class to another and is very dependent on the testing set used (as shown by the error bars). A closer look at the data revealed that some of the scoring functions varied following the same patterns. In order to further investigate this trend, we computed the correlations between scoring functions regardless of the protein target and binding affinities (Figure 7). Some of the scoring functions are highly correlated, with values of τ computed for each pair often over 0.60, with a maximum of 0.75 between XScore and ChemScore. In these cases, they are more correlated to each other than to the observed binding energies, as the best coefficient obtained between scoring functions and observed binding affinities was 0.37 (with XScore). In fact, the four scoring functions previously identified as the most accurate in Table 3 (XScore, GoldScore, ChemScore, and DrugScore$^{\mathrm{CSD}}$) are all highly correlated (τ ranging from 0.50 to 0.75). Our scoring function, RankScore, correlates with 8 scoring functions with τ greater than 0.50, while GoldScore correlates with 11 scoring functions at the same threshold of τ. On the other side, PMF and FlexXScore, which were found to be poorly accurate, are not highly correlated with the other scoring functions assessed. It is striking that scoring functions derived by different groups in very different manners exhibit this high level of correlation. It is worth recalling that DrugScore is a knowledge-based scoring function, while XScore is an empirical/consensus scoring function and ChemScore is empirical. In addition, they are made up of terms accounting for different properties (e.g., no entropy term in DrugScore). Interestingly, GlideScore has been originally derived from ChemScore but does not show a high degree of correlation with it.



	ChemScore	DrugScoreCSD	DrugScorePDB	DockScore	eHiTS	FlexXScore	GlideScore	GoldScore	Hammerhead	LigScore1	LigScore2	PLP1	PLP2	PMF-Cerius2	PMF-Sybyl	RankScore	Surflex	Xscore
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1.00	0.57	0.53	0.36	0.45	0.26	0.36	0.51	0.44	0.10	0.30	0.44	0.43	0.21	0.00	0.39	0.45	0.75
2	0.57	1.00	0.74	0.46	0.42	0.19	0.37	0.60	0.45	0.28	0.51	0.60	0.56	0.22	0.01	0.56	0.44	0.62
3	0.53	0.74	1.00	0.48	0.46	0.26	0.40	0.61	0.49	0.32	0.52	0.74	0.67	0.12	0.05	0.62	0.41	0.63
4	0.36	0.46	0.48	1.00	0.41	0.37	0.44	0.60	0.44	0.42	0.57	0.52	0.57	0.31	0.23	0.56	0.40	0.41
5	0.45	0.42	0.46	0.41	1.00	0.37	0.42	0.46	0.37	0.22	0.38	0.43	0.44	0.15	0.04	0.46	0.38	0.48
6	0.26	0.19	0.26	0.37	0.37	1.00	0.37	0.28	0.43	0.34	0.33	0.34	0.37	0.09	0.16	0.35	0.33	0.19
7	0.36	0.37	0.40	0.44	0.42	0.37	1.00	0.44	0.41	0.28	0.38	0.44	0.45	0.19	0.11	0.51	0.55	0.38
8	0.51	0.60	0.61	0.60	0.46	0.28	0.44	1.00	0.57	0.38	0.56	0.60	0.57	0.31	0.12	0.59	0.52	0.58
9	0.44	0.45	0.49	0.44	0.37	0.43	0.41	0.57	1.00	0.43	0.46	0.49	0.50	0.23	0.11	0.45	0.60	0.43
10	0.10	0.28	0.32	0.42	0.22	0.34	0.28	0.38	0.43	1.00	0.61	0.37	0.39	0.20	0.20	0.47	0.34	0.12
11	0.30	0.51	0.52	0.57	0.38	0.33	0.38	0.56	0.46	0.61	1.00	0.54	0.55	0.31	0.19	0.60	0.41	0.36
12	0.44	0.60	0.74	0.52	0.43	0.34	0.44	0.60	0.49	0.37	0.54	1.00	0.84	0.14	0.02	0.67	0.41	0.53
13	0.43	0.56	0.67	0.57	0.44	0.37	0.45	0.57	0.50	0.39	0.55	0.84	1.00	0.18	0.06	0.64	0.40	0.49
14	0.21	0.22	0.12	0.31	0.15	0.09	0.19	0.31	0.23	0.20	0.31	0.14	0.18	1.00	0.55	0.14	0.26	0.22
15	0.00	0.01	0.05	0.23	0.04	0.16	0.11	0.12	0.11	0.20	0.19	0.02	0.06	0.55	1.00	0.06	0.16	0.05
16	0.39	0.56	0.62	0.56	0.46	0.35	0.51	0.59	0.45	0.47	0.60	0.67	0.64	0.14	0.06	1.00	0.47	0.45
17	0.45	0.44	0.41	0.40	0.38	0.33	0.55	0.52	0.60	0.34	0.41	0.41	0.40	0.26	0.16	0.47	1.00	0.44
18	0.75	0.62	0.63	0.41	0.48	0.19	0.38	0.58	0.43	0.12	0.36	0.53	0.49	0.22	0.05	0.45	0.44	1.00
Average	0.38	0.45	0.47	0.44	0.37	0.30	0.38	0.49	0.43	0.32	0.45	0.48	0.48	0.23	0.13	0.47	0.41	0.42
Minimum	0.00	0.01	0.05	0.23	0.04	0.09	0.11	0.12	0.11	0.10	0.19	0.02	0.06	0.09	0.00	0.06	0.16	0.05

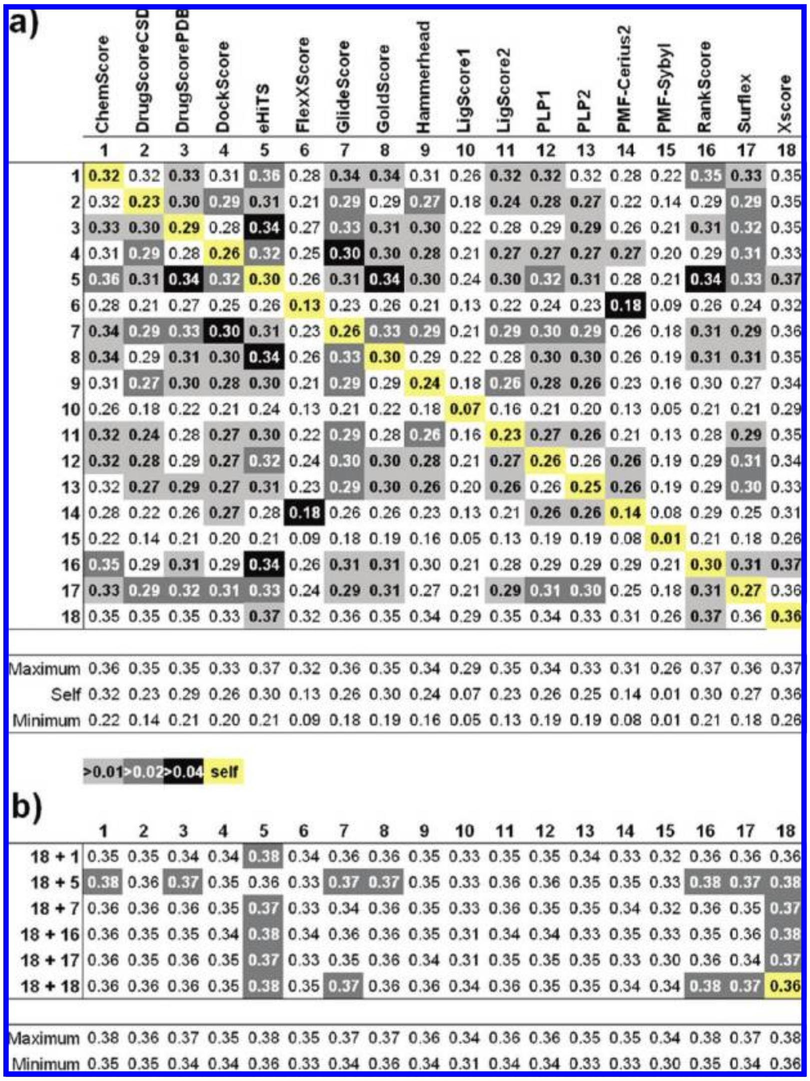
>0.30 >0.40 >0.50 >0.60

Figure 7. Ranked-list correlation coefficients (τ) calculated between predicted ranking lists of SFs; the darker the shading, the higher the correlation (see key). The numbers **1**-**18** represent the scoring functions as specified in the top row.

Consensus Scoring. A scoring function aims to predict the binding affinities of ligands for proteins and/or to compute the free energy of binding. The results from the previous section demonstrated that the best performing scoring functions capture the same information, showing high correlation between their scores (often with $\tau > 0.60$), while the moderate correlation ($0.20 < \tau < 0.35$) between these functions and the observed binding affinities also indicates that these functions may disregard the same aspects of the binding process. From these conclusions, we hypothesized that consensus scoring may not lead to significantly better accuracy and that only a very few scoring functions can be considered in this context. In order to test this hypothesis, the accuracy for each pair of scoring functions was assessed. In order to normalize the data, we combined the ranks computed with each scoring function and not the scores. As illustrated in Figure 8, most of the combinations (shown in white) led to τ coefficients that are not better than each of the τ coefficients computed for each scoring function. More interestingly, there was no case where the predictiveness of a pair of functions was lower than both individual scoring functions. In all cases, combinations that led to the best τ values included either XScore, the eHiTS scoring function, RankScore, or ChemScore, which were already found to be among the four most predictive scoring functions with τ greater than 0.30. These data validate our hypothesis and demonstrate that consensus scoring using a combination of traditional scoring functions can at best provide a moderate increase in accuracy and that consensus scoring should be developed from more different scoring approaches[37] or include additional information.[38]

Impact of Protein Conformation and Water Molecules. Cross-docking is a more appropriate experiment than self-docking when one wants to mimic virtual screening studies. A set of cross-docked ligands (i.e., Set 2) was therefore assembled and used to assess scoring functions in this context. With this second set, we aim to assess the impact of the selected protein conformation and presence/absence of water molecules on scoring function accuracy. Each cross-docked complex was first optimized in presence of water molecules and then assigned two scores: one corresponding to the scoring considering water molecules (wet/wet) and one corresponding to the scoring with no waters included (wet/dry). As a third subset, we performed the local optimization of the ligands without the water molecules and scored the resulting complexes without any water molecules (dry/dry). This resulted in three subsets: one with the key water molecules retained for both docking and scoring, one with all the key water molecules retained for docking but removed for scoring, and one with all the water molecules removed for both docking and scoring. As each of the nearly 1000 complexes can be found in the three subsets, each complex was assigned three scores. With all these data in hand, one can simulate the displacement of the water molecule, by selecting the best score out of the three for each complex (water displaceable).

At this stage, each of the 92 ligands had been scored with all the structures (native and non-native) of the same protein. In order to evaluate the impact of the selection of the protein conformation on the scoring accuracy, scripts were written to evaluate the correlation coefficients with various random selections of cross-docked complexes. For this purpose, a protein structure (e.g., HIV-1 protease 1a30) out of the non-native protein structures of the same protein was randomly selected for each ligand (e.g., HIV-1 protease ligand 1b6l) and the resulting complex scored. From this set of predictions, the ranking of ligands by predicted binding affinity was computed, compared to the observed ranking and a value



a)

	ChemScore	DrugScoreCSD	DrugScorePDB	DockScore	eHiTS	FlexXScore	GlideScore	GoldScore	Hammerhead	LigScore1	LigScore2	PLP1	PLP2	PMF-Cerius2	PMF-Sybyl	RankScore	Surflex	Xscore
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	0.32	0.32	0.33	0.31	0.36	0.28	0.34	0.34	0.31	0.26	0.32	0.32	0.32	0.28	0.22	0.35	0.33	0.35
2	0.32	0.23	0.30	0.29	0.31	0.21	0.29	0.29	0.27	0.18	0.24	0.28	0.27	0.22	0.14	0.29	0.29	0.35
3	0.33	0.30	0.29	0.28	0.34	0.27	0.33	0.31	0.30	0.22	0.28	0.29	0.29	0.26	0.21	0.31	0.32	0.35
4	0.31	0.29	0.28	0.26	0.32	0.25	0.30	0.30	0.28	0.21	0.27	0.27	0.27	0.27	0.20	0.29	0.31	0.33
5	0.36	0.31	0.34	0.32	0.30	0.26	0.31	0.34	0.30	0.24	0.30	0.32	0.31	0.28	0.21	0.34	0.33	0.37
6	0.28	0.21	0.27	0.25	0.26	0.13	0.23	0.26	0.21	0.13	0.22	0.24	0.23	0.18	0.09	0.26	0.24	0.32
7	0.34	0.29	0.33	0.30	0.31	0.23	0.26	0.33	0.29	0.21	0.29	0.30	0.29	0.26	0.18	0.31	0.29	0.36
8	0.34	0.29	0.31	0.30	0.34	0.26	0.33	0.30	0.29	0.22	0.28	0.30	0.30	0.26	0.19	0.31	0.31	0.35
9	0.31	0.27	0.30	0.28	0.30	0.21	0.29	0.29	0.24	0.18	0.26	0.28	0.26	0.23	0.16	0.30	0.27	0.34
10	0.26	0.18	0.22	0.21	0.24	0.13	0.21	0.22	0.18	0.07	0.16	0.21	0.20	0.13	0.05	0.21	0.21	0.29
11	0.32	0.24	0.28	0.27	0.30	0.22	0.29	0.28	0.26	0.16	0.23	0.27	0.26	0.21	0.13	0.28	0.29	0.35
12	0.32	0.28	0.29	0.27	0.32	0.24	0.30	0.30	0.28	0.21	0.27	0.26	0.26	0.26	0.19	0.29	0.31	0.34
13	0.32	0.27	0.29	0.27	0.31	0.23	0.29	0.30	0.26	0.20	0.26	0.26	0.25	0.26	0.19	0.29	0.30	0.33
14	0.28	0.22	0.26	0.27	0.28	0.18	0.26	0.26	0.23	0.13	0.21	0.26	0.26	0.14	0.08	0.29	0.25	0.31
15	0.22	0.14	0.21	0.20	0.21	0.09	0.18	0.19	0.16	0.05	0.13	0.19	0.19	0.08	0.01	0.21	0.18	0.26
16	0.35	0.29	0.31	0.29	0.34	0.26	0.31	0.31	0.30	0.21	0.28	0.29	0.29	0.29	0.21	0.30	0.31	0.37
17	0.33	0.29	0.32	0.31	0.33	0.24	0.29	0.31	0.27	0.21	0.29	0.31	0.30	0.25	0.18	0.31	0.27	0.36
18	0.35	0.35	0.35	0.33	0.37	0.32	0.36	0.35	0.34	0.29	0.35	0.34	0.33	0.31	0.26	0.37	0.36	0.36
Maximum	0.36	0.35	0.35	0.33	0.37	0.32	0.36	0.35	0.34	0.29	0.35	0.34	0.33	0.31	0.26	0.37	0.36	0.37
Self	0.32	0.23	0.29	0.26	0.30	0.13	0.26	0.30	0.24	0.07	0.23	0.26	0.25	0.14	0.01	0.30	0.27	0.36
Minimum	0.22	0.14	0.21	0.20	0.21	0.09	0.18	0.19	0.16	0.05	0.13	0.19	0.19	0.08	0.01	0.21	0.18	0.26

>0.01 >0.02 >0.04 self

b)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
18 + 1	0.35	0.35	0.34	0.34	0.38	0.34	0.36	0.36	0.35	0.33	0.35	0.35	0.34	0.33	0.32	0.36	0.36	0.36
18 + 5	0.38	0.36	0.37	0.35	0.36	0.33	0.37	0.37	0.35	0.33	0.36	0.36	0.35	0.35	0.33	0.38	0.37	0.38
18 + 7	0.36	0.36	0.36	0.35	0.37	0.33	0.34	0.36	0.35	0.33	0.36	0.35	0.35	0.34	0.32	0.36	0.35	0.37
18 + 16	0.36	0.35	0.35	0.34	0.38	0.34	0.36	0.36	0.35	0.31	0.34	0.34	0.33	0.35	0.33	0.35	0.36	0.38
18 + 17	0.36	0.35	0.36	0.35	0.37	0.33	0.35	0.36	0.34	0.31	0.35	0.35	0.35	0.33	0.30	0.36	0.34	0.37
18 + 18	0.36	0.36	0.36	0.35	0.38	0.35	0.37	0.36	0.36	0.34	0.36	0.35	0.35	0.34	0.34	0.38	0.37	0.36
Maximum	0.38	0.36	0.37	0.35	0.38	0.35	0.37	0.37	0.36	0.34	0.36	0.36	0.35	0.35	0.34	0.38	0.37	0.38
Minimum	0.35	0.35	0.34	0.34	0.36	0.33	0.34	0.36	0.34	0.31	0.34	0.34	0.33	0.33	0.30	0.35	0.34	0.36

Figure 8. τ calculated for combinations of scoring functions. The numbers **1**-**18** on the second row and leftmost columns represent the scoring functions as described in the top row of part (a). (a) Pairs of scoring functions: darker boxes represent an increase of the combined scoring functions over the individual scoring functions (see key; the value in the key indicates the increase of the value of τ over the value for the individual scoring function in the same row). The yellow boxes correspond to the individual scoring functions. (b) Groups of three scoring functions: darker boxes represent increase with respect to XScore alone (yellow box).

of τ was generated. Then the process was reiterated with a different random selection (e.g., HIV-1 protease ligand 1b6l alternatively cross-docked to all the HIV-1 protease conformations but the 1b6l native protein conformation). This process was iterated 10,000 times with a different population of cross-docked structures each time, thus providing a range of values for the correlation coefficient τ (see Figure 9). The median values for the τ obtained under the different conditions are given in Table S5 (see the Supporting Information) and graphed in Figure 9.

Although the presence of water molecules bridging the intermolecular interaction between ligands and biomacromolecules is known to be critical for binding in some cases, docking and scoring programs do not commonly handle water molecules. Two aspects can be affected by water molecules: the binding mode can be optimized differently whether the waters are kept or not, and the score of the same pose can be different if the scoring function scores the water-mediated interactions. As can be seen in Figure 9 (and Table S5 in the Supporting Information), XScore, DrugScore, and the eHiTS scoring function demonstrated the best correlations when the waters are kept, although within errors from more than half of the other scoring functions. When comparing the Kendall coefficients for the different water treatments, it is clear that none of the scoring function is significantly affected by the presence or absence of water molecules. It is worth mentioning that RankScore has been optimized to account for the presence of water molecules and demonstrates slightly enhanced accuracy when displaceable water molecules are used. Overall, making the water molecules displaceable increases the accuracy of most of the scoring functions as expected although by only a small increment.

Next we looked at the drop in accuracy when going from native protein structures to cross-docked complexes. It clearly appears (Figure 10, see also Table S6 in the Supporting Information) that some of the scoring functions including GlideScore and RankScore are greatly affected by the protein structures considered. The selection of the protein conformation when more than one crystal structure is available should be done with care. Once more, XScore and the eHiTS scoring function were the most accurate, likely an outcome of the larger training sets used in the parametrization of these

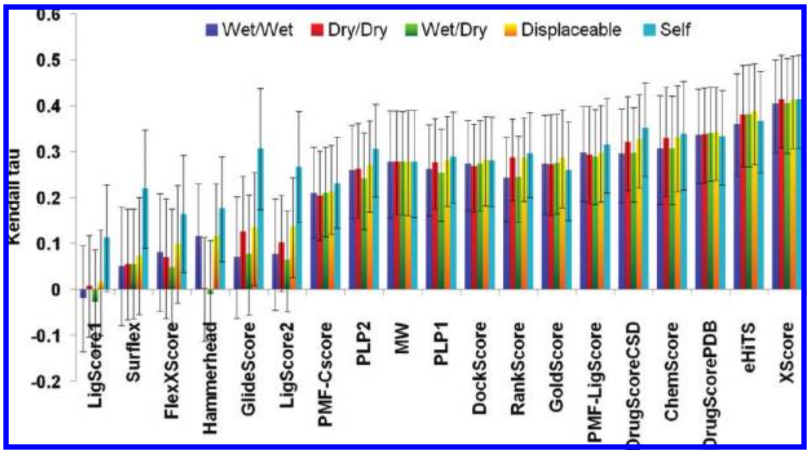


Figure 9. Accuracy (τ) of the scoring functions on set 2 with different water considerations. Waters were kept for docking and scoring (wet/wet, blue), kept for docking and removed for scoring (dry/dry, red), removed for docking and scoring (dry/dry, green), or made displaceable (yellow). For comparison, the correlation of the scores obtained for the native structures in wet/wet conditions is shown (light blue).

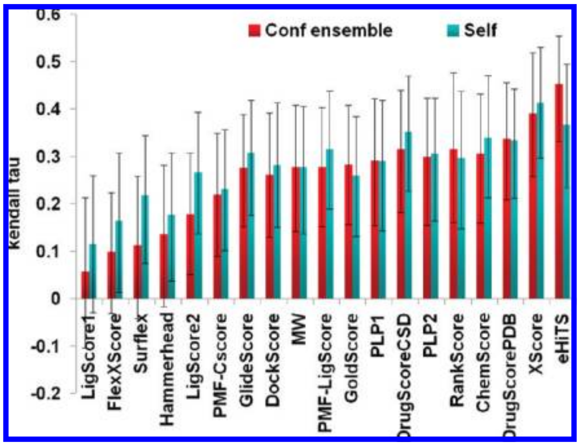


Figure 10. Accuracy of the scoring functions (τ) on the complete set 2 when protein conformational ensembles *(best scored complex among the cross-docked ones)* were considered.

scoring functions. Alternatively, each ligand was docked to all the non-native protein conformations available for each protein and the best score was retained. The selected score corresponds to a docking experiment carried out on a conformational ensemble of protein structures, therefore considering the protein flexibility (conformational ensemble, see Figure 10). By moving to conformational ensembles, we expected to restore part of the accuracy lost when moving from self- to cross-docked structures. In fact, while most of the scoring functions are negatively affected by the use of conformational ensembles, the accuracy of the eHiTS scoring function significantly increases when conformational ensembles are used in place of the native protein conformation. A closer look at the data did not allow any explanation for this behavior. We believe that scoring functions based on soft proteins are less sensitive to the protein conformation. In fact, the steep Lennard-Jones 12−6 employed by some of the scoring functions (e.g., RankScore) is very sensitive to subtle moves, while the soft function used by eHiTS scoring function is not (Figure 10).

CONCLUSIONS

We have carefully developed two sets of protein/ligand complexes with reduced correlation between MW and binding affinities. The first set was scanned for accuracy in scoring of native poses (analogous to self-docking) using a large number of available scoring functions. This screening revealed the good accuracy of XScore, the eHiTS scoring function, DrugScore (PDB- and CSD-derived), GoldScore, and ChemScore. Analysis of the results on subsets of complexes indicated a large dependence on the protein, analogously to the one observed previously with docking programs.[11] This can be in part explained by the training sets used in the derivation of these scoring functions: eHiTS and XScore are the empirical scoring functions with the largest training sets considered in this study. In particular, the eHiTS scoring function implements specific parameters based on the protein class involved, which might (at least in part) account for this enhanced accuracy. In a subsequent section, we demonstrated that consensus scoring can only lead to moderate increase in accuracy and that other strategies should be proposed (i.e., smart postprocessing of poses) or novel (i.e., more predictive) scoring functions should be developed to address the scoring issue in docking methods. Finally, we have shown that some scoring functions lose some predictive power when applied to cross-docked structures, demonstrating the need to incorporate protein flexibility in docking programs and scoring functions. From this information we draw the conclusion that using softer proteins or conformational ensembles should lead to more predictive scoring functions for use in docking-based virtual screening. More surprisingly, the consideration of displaceable water molecules has been shown to improve the docking accuracy, although the present work shows that it does not affect the scoring significantly. However, most of these scoring functions were trained on "dry" proteins and often do not consider the water molecules even if present. Our scoring function RankScore performed well in the self-docking experiment and when using conformational ensembles (as does our docking program Fitted), but poorly with cross-docked structures. In subsequent work, we will develop a more accurate scoring function for these situations.

EXPERIMENTAL SECTION

Generalities. Scripts were required in order to automate many repetitive tasks in each of the interfaces; to this effect Python scripts were used in Maestro, SPL scripts in Sybyl, and BCL scripts in InsightII. Awk, shell, and Python scripts were written to pre- and postprocess the structures and input and output files from each of the scoring functions as necessary. Calculations were run on SGI Fuel workstations with a single R16000 processor and Linux workstations (AMD Opteron and/or Intel Core2 processors).

Selection of the Training Set Structures. 1. PDB. Queries on the Protein Data Bank[39] were performed looking for X-ray crystal structures of either HIV-1 protease (HIVP), thrombin, or matrix metalloproteases (MMPs) in complexes with small molecule ligands. The structures found were filtered by keeping the ones with resolution better than 2.5 Å, and among those, the ones containing noncovalent ligands with reported activity. A set of 20 complexes for each protein was selected ensuring chemical diversity and a homogeneous representation of 7 orders of magnitude of K_i. For HIVP, the 20 complexes included 11 wild-type structures and 9 mutants, with all the point mutations located away from the first layer of residues in contact with the ligand. For thrombin, the 20 structures correspond to wild-type human thrombin; the MMP training set is composed of 6 structures of MMP-3 (stromelysin-1), 2 structures of MMP-7 (matrilysin), 9 structures of MMP-8 (collagenase-2), and 3 structures of MMP-13 (collagenase-3). 2. PDBBind. Two hundred and twenty complexes from the refined PDBBind database[18,19] were selected for chemical diversity and even distribution of affinity spanning 9 orders of magnitude (from $pK_d = 3$ to 12). Naturally, the affinity ranges between $pK_d = 5$ to 10 (the most common affinity of a good lead) are the better represented. The compounds selected have a leadlike molecular weight of between 250 and 600, and the complexes for which there was more than one and different affinities reported within the PDBBind database were not selected.

Preparation of the Training Set - Generalities. Preparation of the complexes for further calculations was performed with Maestro 7.0 (Schrödinger, Inc.) and MacroModel 9.0. (Schrödinger, Inc.). Succinctly, it involved completion of the side chains missing from the PDB structure (exposed to solvent); capping of the protein termini as either ammoniums or carboxylates; assignment of bond orders and protonation states in the ligand and active site residues; and removal of all extraneous molecules (e.g., ions far away from the ligand binding site, ethyleneglycol). In the cases where more than one pose for the ligand was present in the PDB file, the one with the highest occupancy was chosen; otherwise the first pose described was used. Water molecules were treated as described in the following section. Hydrogen atoms were added and minimized with all other atoms fixed (MMFFs94, up to 500 steps of conjugate gradient). The binding mode of the ligands was relaxed by an energy minimization in which only ligand atoms and hydrogens bound to heteroatoms were allowed to move (MMFFs94, up to 2000 steps of conjugate gradient), and heavy atoms were constrained by a harmonic potential ($k = 5$ kcal mol^{-1} Å^{-1}) to their crystal structure positions.

Treatment of Water Molecules. All explicit water molecules were removed from the complexes, except for the ones that were contained in the intersection of a volume 3.0 Å around all atoms of the ligand and a volume 3.0 Å around all the atoms of the receptor. These remaining water molecules (varying in number between 0 and 20) were examined more closely, and the ones capable of forming at least 3 hydrogen bonds with both the ligand and the receptor were kept.

Assignment of Protonation States for Aspartyl Proteases. The two catalytic aspartyl residues in HIV-1 protease (Asp25) are considered to exist in a monoprotonated (monoionized) state for catalytic activity.[20] The exception has to be made for ligands binding to the catalytic dyad by means of a 1,2-diol, where NMR and X-ray data points to both Asp25 being protonated, formed a tight hydrogen bonding network with the two hydroxyls in the ligand,[21] as well as ligands with an ammonium group facing the catalytic dyad, which might stabilize the dideprotonated state.[22] A careful observation of the environment around the Asp25, defining how the hydrogen bond network could be formed, together with the coplanar (or not) orientation of the Asp25 carboxylates led us to define the protonation state of each of the complexes.

Thrombin. Crystal structures for thrombin usually contain 3 chains: the 2 chains (L and H) resulting from the self-cleavage of the protein plus a hirugen peptide, which binds to an alternate binding site in the protein. Given that the hirugen peptide and the low-molecular weight chain of thrombin lie far away (more than 15 Å) from any atoms of the ligand, they were removed in all cases, and only the heavy chain was used for the calculations. The protonation state of all residues in the protein was assigned as expected at pH 7, except for the artificial terminal groups resulting from the missing loop in the crystal structures between residues 146 and 149E, which were considered neutral (COOH and NH_2) for the sake of not adding artificial charges in the binding site. All ligands contain at least one protonated basic moiety (ammonium, amidinium, guanidinium), which interacts with Asp45 in the receptor.

Metalloenzymes. Zinc-containing proteins (e.g., MMP-3, MMP-8, carbonic anhydrase) were prepared by breaking all bonds between the metal ion and heteroatoms in ligands and protein and specifying a formal charge of +2.

PDBBind Complexes. The complexes retrieved from the PDBbind database were prepared in a way analogous to the previous complexes. Ligand protonation states were checked and corrected where applicable (e.g., in some cases nitrogens attached to aryl groups where incorrectly protonated); aspartyl proteases (e.g., penicillopepsin, endothiapepsin, SIV and HIV-1 protease) were identified as such, and the catalytic dyad was treated as described above; metalloenzymes' Zn atoms were treated as in MMPs. In the case of multimeric complexes, the minimum number of chains necessary to describe the complex was kept. The number of atoms for the proteins was kept under 10,000 by removing residues far away (i.e., > 20 Å) from the ligand binding site if necessary.

Preparation of Cross-Scoring Set (Set 2). Proteins represented with at least 5 complexes in set 1 were selected (with the exception of PNP, for which not all proteins were from the same source species). Crystal structures of the complexes were prepared as described above, and all complexes from the same protein were superimposed to the α carbon trace of one of them. FITTED was used in local search mode to adjust the binding mode of the ligands to each protein binding site separately, with each protein being treated rigidly. Due to the presence of flexible side chains, the maximum allowed translation in the generation of the initial population was set as 5 Å. The resulting binding modes were used as input for all scoring methods.

Scoring: CScore. The stand-alone CScore module from Sybyl v7.3 was used, with default parameters for the DScore (DockScore), GScore (GoldScore), PMF, and ChemScore scoring functions. **GlideScore.** Glide v4.5 was used in all calculations. Grids were generated in a box of 20 Å around the ligand, with default parameters. Scoring was performed in place. **eHiTS.** The score.sh script supplied with eHiTS v6.2 was used. Two ligands, 1h22 and 1h23, were not assigned a score by eHiTS for having more than 10 rotatable bonds in a linear fragment; a few others failed to be optimized, hence the nonoptimized score was considered. **Surflex.** Surflex v2.301 was used. Protomol files for the protein structures were first generated with the "proto" option, and then scores were calculated with the "score_list" option. The nonoptimized score was considered, as the optimized one gave poorer correlations. **Cerius2.** Cerius2 v4.10 was used on all calculations. PDB-formatted protein structure files and SD-formatted ligand files were used as input for the LigScore1/2, PLP1/2, PMF, and Jain (Hammerhead) scoring functions. **XScore.** X-Tool v.1.2.1 was used. Protein and ligand structures were first prepared with the "-fixpdb" and "-fixmol2" options, respectively, prior to running the score computation with the "-score" option. **DrugScore.** Executables of DrugScore^{PDB} and DrugScore^{CSD} v1.2 were used under IRIX. **RankScore.** The scores were calculated with the FITTED 2.6 docking program, after preprocessing the protein and ligand structures with PROCESS and SMART, respectively.

Bootstrap Analysis. The scores for each protein/ligand complex with every scoring function as well as the molecular weight and the experimental binding affinities were organized in a CSV file and processed with a Python script. A random subset of observations of the same size as the original (repetition was allowed) was selected, and the correlation coefficients for each scoring function were calculated using the functions provided in the SciPy module; this process was iterated 10,000 times. For each scoring function, the range of correlation coefficients spanning 95% of the obtained in the previous trials was taken as the uncertainty in the correlation coefficient for the original set. In the case of the cross-docked structures, a random protein was selected for each ligand on each iteration, and the correlation was calculated with one complex for each ligand.

ACKNOWLEDGMENT

We thank Virochem Pharma as well as the Canadian Foundation for Innovation (New Opportunities Fund) and the Canadian Institutes for Health Research (CIHR) for financial support. Operational help and discussions with developers from Simbiosys (Zsolt Zsoldos) and BioSolveIT (Chris Lemmen) as well as academic licenses for the use of DrugScore (G. Klebe. U. Marburg, Germany), XScore (S. Wang, U. Michigan), Surflex (A. Jain, UCSF), and eHiTS (Simbiosys, Toronto, Canada) are acknowledged. P.E. was supported by a scholarship from CIHR (Strategic Training Initiative in Chemical Biology) and a J. W. McConnell Memorial fellowship while performing the research.

Supporting Information Available: Complete training set description (PDB codes, experimental ligand binding affinities) and detailed data for Figures 5, 6, 9, and 10. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. J. Hits, Leads and Artifacts from Virtual and High Throughput Screening. *Molecular Informatics: Confronting Complexity, May 13th−16th 2002*; 2002.

(2) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. *Br. J. Pharmacol.* **2008**, *153*, S7–S26.

(3) Rester, U. Dock around the clock - Current status of small molecule docking and scoring. *QSAR Comb. Sci.* **2006**, *25*, 605–615.

(4) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein-ligand docking: Current status and future challenges. *Proteins: Struct., Funct., Genet.* **2006**, *65*, 15–26.

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

(6) Kirchmair, J.; Distinto, S.; Schuster, D.; Spitzer, G.; Langer, T.; Wolber, G. Enhancing drug discovery through in silico screening: Strategies to increase true positives retrieval rates. *Curr. Med. Chem.* **2008**, *15*, 2040–2053.

(7) Aqvist, J.; Medina, C.; Samuelsson, J. E. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **1994**, *7*, 385–391.

(8) Pham, T. A.; Jain, A. N. Customizing scoring functions for docking. *J. Comput.-Aided Mol. Des.* **2008**, 1–18.

(9) Corbeil, C. R.; Englebienne, P.; Moitessier, N. Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. *J. Chem. Inf. Model.* **2007**, *47*, 435–449.

(10) Corbeil, C. R.; Englebienne, P.; Yannopoulos, C. G.; Chan, L.; Das, S. K.; Bilimoria, D.; L'Heureux, L.; Moitessier, N. Docking ligands into flexible and solvated macromolecules. 2. Development and application of FITTED 1.5 to the virtual screening of potential HCV polymerase inhibitors. *J. Chem. Inf. Model.* **2008**, *48*, 902–909.

(11) Corbeil, C. R.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 3. Impact of Input Ligand Conformation, Protein Flexibility, and Water Molecules on the Accuracy of Docking Programs. *J. Chem. Inf. Model.* **2009**, *49*, 997–1009.

(12) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287–2303.

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

(14) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726–741.

(15) Verdonk, M. L.; Mortenson, P. N.; Hall, R. J.; Hartshorn, M. J.; Murray, C. W. Protein-ligand docking against non-native protein conformers. *J. Chem. Inf. Model.* **2008**, *48*, 2214–2225.

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

(17) Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133–139.

(18) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977–2980.

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

(20) Kulkarni, S. S.; Kulkarni, V. M. Structure Based Prediction of Binding Affinity of Human Immunodeficiency Virus-1 Protease Inhibitors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1128–1140.

(21) Yamazaki, T.; Nicholson, L. K.; Wingfield, P.; Stahl, S. J.; Kaufman, J. D.; Eyermann, C. J.; Hodge, C. N.; Lam, P. Y. S.; Torchia, D. A.; et al. NMR and X-ray Evidence That the HIV Protease Catalytic Aspartyl Groups Are Protonated in the Complex Formed by the Protease and a Non-Peptide Cyclic Urea-Based Inhibitor. *J. Am. Chem. Soc.* **1994**, *116*, 10791–10792.

(22) Czodrowski, P.; Sotriffer, C. A.; Klebe, G. Atypical Protonation States in the Active Site of HIV-1 Protease: A Computational Study. *J. Chem. Inf. Model.* **2007**, *47*, 1590–1598.

(23) 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.

(24) 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.

(25) Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M. LigScore: A novel scoring function for predicting binding affinities. *J. Mol. Graphics Modell.* **2005**, *23*, 395–407.

(26) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425–445.

(27) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 609–623.

(28) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(29) Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S. B.; Johnson, A. P. eHiTS: A new fast, exhaustive flexible ligand docking system. *J. Mol. Graphics Modell.* **2007**, *26*, 198–212.

(30) Jain, A. N. Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 281–306.

(31) Moitessier, N.; Therrien, E.; Hanessian, S. A method for induced-fit docking, scoring, and ranking of flexible ligands. Application to peptidic and pseudopeptidic β-secretase (BACE 1) inhibitors. *J. Med. Chem.* **2006**, *49*, 5885–5894.

(32) Englebienne, P.; Fiaux, H.; Kuntz, D. A.; Corbeil, C. R.; Gerber-Lemaire, S.; Rose, D. R.; Moitessier, N. Evaluation of Docking Programs for Predicting Binding of Golgi alpha-Mannosidase II Inhibitors: A Comparison with Crystallography. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 160–176.

(33) Wang, R.; Lu, Y.; Fang, X.; Wang, S. An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114–2125.

(34) Grantham, R. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* **1974**, *185*, 862–864.

(35) Hopp, T. P.; Woods, K. R. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. U. S. A.* **1981**, *78*, 3824–3828.

(36) Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105–132.

(37) Renner, S.; Derksen, S.; Radestock, S.; Morchen, F. Maximum Common Binding Modes (MCBM): Consensus Docking Scoring Using Multiple Ligand Information and Interaction Fingerprints. *J. Chem. Inf. Model.* **2008**, *48*, 319–332.

(38) Bar-Haim, S.; Aharon, A.; Ben-Moshe, T.; Marantz, Y.; Senderowitz, H. SeleX-CS: A New Consensus Scoring Algorithm for Hit Discovery and Lead Optimization. *J. Chem. Inf. Model.* **2009**, *49*, 623–633.

(39) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Bourne, P. E.; Shindyalov, I. N. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

CI8004308