Docking Ligands into Flexible and Solvated Macromolecules. 5. Force-Field-Based Prediction of Binding Affinities of Ligands to Proteins

Pablo Englebienne and Nicolas Moitessier*

Department of Chemistry, McGill University, 801 Sherbrooke Street West, Montreal, Quebec, Canada H3A 2K6

Received July 17, 2009

We report herein our efforts in the development of three empirical scoring functions with application in protein−ligand docking. A first scoring function was developed from 209 crystal structures of protein−ligand complexes and a second one from 946 cross-docked complexes. Tuning of the coefficients for the different terms making up these functions was performed by an iterative approach to optimize the correlations between observed activities and calculated scores. A third scoring function was developed from libraries of known actives and decoys docked to six different protein conformational ensembles. In the latter case, the tuning of the coefficients was performed so as to optimize the area under the curve of a receiver operating characteristic (ROC) for the discrimination of actives and inactives. The newly developed scoring functions were next assessed on independent sets of protein−ligand complexes for their ability to predict binding affinities and to discriminate actives from inactives. In the first validation the first function, which was trained on active compounds only, performed as well as other commonly used ones. On a high-throughput virtual screening validation on five protein conformational ensembles, the third scoring function that included data from inactive compounds performed significantly better. This validation showed that the inclusion of data from inactive compounds is critical for performance in virtual high-throughput screening applications.

INTRODUCTION

For the past decade or so, docking methods have been widely used in structure-based drug design.[1] However, although many successful studies have been reported, scoring of the docked ligands is still often poorly predictive and highly target dependent.[2] In practice, the prediction of binding affinities in protein−ligand complexes can be achieved with various levels of accuracy and speed.[3] On one side of the spectrum, molecular dynamics (MD) based methods (e.g., free energy perturbation[4] and linear interaction energy (LIE)[5,6]) take into account the average interaction energy of a Boltzmann distribution of conformers in explicit or implicit aqueous media. On the other end of the spectrum, scoring functions implemented in docking programs and/or used in virtual screening studies deal with a single ligand pose and often in vacuo.[7,8] This certainly yields a different time frame for the calculation with a concomitant decrease in accuracy. These scoring functions are traditionally classified as being either force-field based, empirical (regression-based), or based on a potential of mean force.[2]

The application of molecular mechanical force fields to the scoring of docked poses (i.e., scoring functions implemented in docking programs) has been limited to the AMBER force field[9−11] (AutoDock,[12] Dock,[13] Fitted[14]), the Tripos force field[15] (Gold[16]), as well as the Dreiding[17] and CFF force fields[18] (LigandFit[19]). In these cases, the selection of a specific force field was often based on its availability (e.g., AMBER parameters are publicly available[20]). To date, force fields have never been assessed in detail to evaluate their ability to reproduce binding free energies of ligands to proteins. In fact, although force fields have been developed to reproduce a number of thermodynamic and kinetic properties of molecules, they have not been specifically developed to predict binding free energies and, when not combined with other terms, often overestimate them.[5]

We recently developed a docking program (Fitted) that accounts for protein flexibility and bridging water.[14,21,22] The scoring function used with this software is a force-field-based scoring function reported earlier.[23] Although this scoring function was found to extract new active compounds from large libraries[21] and rank compounds by activity with accuracy similar to the state-of-the art scoring functions,[24] we believe that the number of false positives can be reduced by the use of a more advanced scoring function. We report herein our efforts toward the development of new force-field-based scoring functions starting from an exhaustive comparative study of the ability of popular force fields to predict the binding affinity of ligands to proteins. In the following sections, we describe the development of a scoring function derived from crystallographic structures and docked complexes as well as another scoring function trained from virtual screening data. Finally, the validation of the developed scoring functions on benchmark sets of protein−ligand complexes and sets of active compounds and decoys is described. As with our docking program, we focused on a scoring function that would be suited for flexible proteins including bridging water molecules.

RESULTS AND DISCUSSION

RankScore. Our previous version of RankScore (referred to as RankScore 1 in this publication) was derived from a

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

10.1021/ci900251k CCC: $40.75
Published on Web 10/30/2009

set of docked ligands and crystal structures of BACE-1 inhibitors.[23] In the present study, we envisioned three different approaches. A scoring function as implemented in docking programs can have two major applications. First, scoring functions can be used to rank potential ligands by their predicted binding affinity (i.e., as in a lead optimization problem). Second, they can be used at an earlier drug discovery stage to discriminate active from inactive compounds (i.e., as a strategy for hit identification).

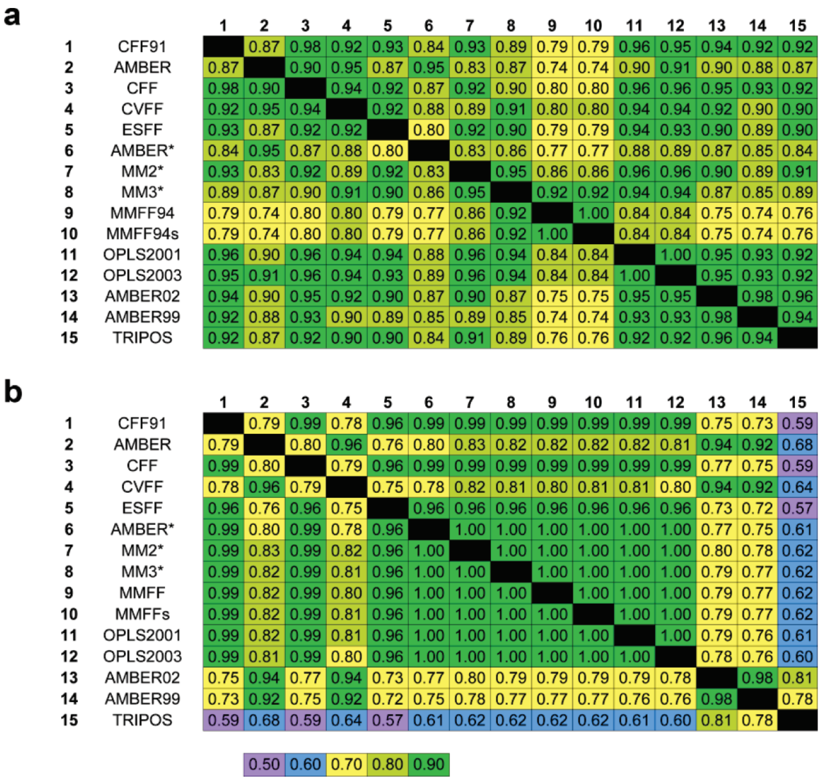
On the basis of these premises, we exploited our previously reported set of 209 protein/ligand crystal structures[24] to derive a scoring function (referred to as RankScore 2). This second version of RankScore was, consequently, developed from active compounds (exhibiting weak to strong binding affinities) and has potential application in the ranking of actives. This scoring function was based on the assumption that the protein conformation is fine tuned to each of the ligands. However, even with a large number of conformations from experimental data or simulations, docking methods do not allow to fine-tune the protein conformations and simulate a true induced fit. To address this issue, we also developed a scoring function, namely, RankScore 3, from our set of 946 cross-docked ligand/protein complexes.[24] Finally, we will describe the development of a third scoring function (RankScore 4) from large sets of active and inactive compounds. This last variant will have application in virtual screening (discrimination between binders and nonbinders).

Screened Force Fields. When considering the development of a force-field-based scoring function, there is no rationale for the selection of one force field over another. The first goal of this research project was therefore to assess various class I and class II molecular mechanics force fields to identify the one(s) that would be better suited to make a quick yet accurate estimation of the binding energy between a ligand and a receptor and to compare their accuracy to that of commercially available scoring functions. We selected five force fields implemented in Discover (Accelrys), namely CVFF, CFF91, CFF, ESFF, and AMBER84, seven force fields implemented in MacroModel (Schrödinger), namely, OPLS2001, OPLS2003, MMFF94, MMFF94s, MM2*, MM3*, and AMBER*, and three force fields implemented in Sybyl (Tripos), namely, Tripos force field, AMBER99, and AMBER02 (which is the only polarizable force field used in this study). Although some of these force fields are overall very similar, their parametrization is inherently different (e.g., experimental data such as microwave, NMR, and neutron diffraction spectroscopy for AMBER[9] and high quality, MP2/6-31G*, ab initio data for MMFF[25]), and the mathematical functions used vary from one force field to another. For instance, the van der Waals interaction energy is often computed using a Lennard−Jones (LJ) potential, most commonly with 6−12 exponents, but 6−9 (CFF, ESFF) and buffered 7−14 (MMFF) exponents are also used. All force fields use the Coulomb equation to calculate the electrostatic interaction between point charges centered on the nuclei. Older versions of AMBER (such as AMBER84 implemented in Discover) and the MMx* family add an explicit term for hydrogen bonding in the form of a 10−12 Lennard−Jones potential.

Training Sets and Scoring Function. The starting point for our comparative study was the two training sets of protein−ligand structures reported in the preceding manuscript of this series.[24] Efforts were devoted to the development of unbiased training sets, showing little correlation between binding affinities and ligand molecular weights. In this previous report, we also applied 18 commonly used scoring functions to evaluate their accuracy with these testing sets. This study shed light on the poor to moderate accuracy of some of these scoring functions when a challenging testing set is used. It also set the lower limit for the development of an accurate scoring function as XScore and ChemScore were found to be the most accurate with Kendall τ coefficients never exceeding 0.37.[24]

General Considerations. A first set of calculations was performed with the force fields as they were shipped in the corresponding software packages, which in some cases prevented the calculation to proceed because of the lack of parameters. With the addition of parameters to the force-field definitions (see Experimental Section), all systems were run with all the force fields. As expected, preliminary runs with HIV-1 protease inhibitors have shown that the inclusion of the key water molecule (the so-called water 301) was critical for a greater predictive power of the method, and only these results will be discussed. In addition, a water molecule was added to a structure where it was clearly missing (see Experimental Section). The hydrogen-bond terms arising from the force-field energy were taken into account in the cases where they existed (AMBER84, MM2*, MM3*). The first step of the computation was to reconcile the structures with the assessed force fields. These relaxation steps when performed with each the force fields led to large deviation in some of the cases. In order to address this issue, the ligands were energy-minimized in the binding sites with the ligand heavy atoms constrained to the crystal structure coordinates. As expected, the resulting structures were much closer to the crystal structures and further processed with a global average RMSD of 0.32 Å, a standard deviation of 0.20 Å, and a maximum RMSD of 1.6 Å for 1okl when minimized with the Amber99 force field. A closer look at the energy-minimized structures indicated that the deviations were only slightly force-field dependent, with the Sybyl-implemented force fields exhibiting larger RMSDs than the Discover and MacroModel ones. However, it was seen that the distribution was highly complex-dependent: 81 out of the 209 ligands had RMSDs of less than 0.5 Å with all the force fields, while only 14 ligands displayed RMSD's higher than 1.0 Å with at least one force field. This may indicate either weaknesses of the force fields to reproduce crystal structures, the impact of computation in vacuo, inaccuracies in the fitting of the models to the electron density in the crystal structures, or the effect of crystal packing. In crystals, protein−ligand complexes are packed and using a single structure may lead to a misinterpretation of the binding mode. For example, we may believe that some of the ligands are exposed to the aqueous medium (which may exhibit a high ionic strength), while they may in fact interact with a second complex (in the next cell of the crystal). Observing a large change in the potentially solvent-exposed portions of ligands is therefore expected.

Force-Field Energy Terms. The first step of our study was to compare the various force fields and their implementations for their ability to reproduce binding affinities. We therefore looked at the correlations between van der



a

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	CFF91		0.87	0.98	0.92	0.93	0.84	0.93	0.89	0.79	0.79	0.96	0.95	0.94	0.92	0.92
2	AMBER	0.87		0.90	0.95	0.87	0.95	0.83	0.87	0.74	0.74	0.90	0.91	0.90	0.88	0.87
3	CFF	0.98	0.90		0.94	0.92	0.87	0.92	0.90	0.80	0.80	0.96	0.96	0.95	0.93	0.92
4	CVFF	0.92	0.95	0.94		0.92	0.88	0.89	0.91	0.80	0.80	0.94	0.94	0.92	0.90	0.90
5	ESFF	0.93	0.87	0.92	0.92		0.80	0.92	0.90	0.79	0.79	0.94	0.93	0.90	0.89	0.90
6	AMBER*	0.84	0.95	0.87	0.88	0.80		0.83	0.86	0.77	0.77	0.88	0.89	0.87	0.85	0.84
7	MM2*	0.93	0.83	0.92	0.89	0.92	0.83		0.95	0.86	0.86	0.96	0.96	0.90	0.89	0.91
8	MM3*	0.89	0.87	0.90	0.91	0.90	0.86	0.95		0.92	0.92	0.94	0.94	0.87	0.85	0.89
9	MMFF94	0.79	0.74	0.80	0.80	0.79	0.77	0.86	0.92		1.00	0.84	0.84	0.75	0.74	0.76
10	MMFF94s	0.79	0.74	0.80	0.80	0.79	0.77	0.86	0.92	1.00		0.84	0.84	0.75	0.74	0.76
11	OPLS2001	0.96	0.90	0.96	0.94	0.94	0.88	0.96	0.94	0.84	0.84		1.00	0.95	0.93	0.92
12	OPLS2003	0.95	0.91	0.96	0.94	0.93	0.89	0.96	0.94	0.84	0.84	1.00		0.95	0.93	0.92
13	AMBER02	0.94	0.90	0.95	0.92	0.90	0.87	0.90	0.87	0.75	0.75	0.95	0.95		0.98	0.96
14	AMBER99	0.92	0.88	0.93	0.90	0.89	0.85	0.89	0.85	0.74	0.74	0.93	0.93	0.98		0.94
15	TRIPOS	0.92	0.87	0.92	0.90	0.90	0.84	0.91	0.89	0.76	0.76	0.92	0.92	0.96	0.94	

b

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	CFF91		0.79	0.99	0.78	0.96	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.75	0.73	0.59
2	AMBER	0.79		0.80	0.96	0.76	0.80	0.83	0.82	0.82	0.82	0.82	0.81	0.94	0.92	0.68
3	CFF	0.99	0.80		0.79	0.96	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.77	0.75	0.59
4	CVFF	0.78	0.96	0.79		0.75	0.78	0.82	0.81	0.80	0.81	0.81	0.80	0.94	0.92	0.64
5	ESFF	0.96	0.76	0.96	0.75		0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.73	0.72	0.57
6	AMBER*	0.99	0.80	0.99	0.78	0.96		1.00	1.00	1.00	1.00	1.00	1.00	0.77	0.75	0.61
7	MM2*	0.99	0.83	0.99	0.82	0.96	1.00		1.00	1.00	1.00	1.00	1.00	0.80	0.78	0.62
8	MM3*	0.99	0.82	0.99	0.81	0.96	1.00	1.00		1.00	1.00	1.00	1.00	0.79	0.77	0.62
9	MMFF	0.99	0.82	0.99	0.80	0.96	1.00	1.00	1.00		1.00	1.00	1.00	0.79	0.77	0.62
10	MMFFs	0.99	0.82	0.99	0.81	0.96	1.00	1.00	1.00	1.00		1.00	1.00	0.79	0.77	0.62
11	OPLS2001	0.99	0.82	0.99	0.81	0.96	1.00	1.00	1.00	1.00	1.00		1.00	0.79	0.76	0.61
12	OPLS2003	0.99	0.81	0.99	0.80	0.96	1.00	1.00	1.00	1.00	1.00	1.00		0.78	0.76	0.60
13	AMBER02	0.75	0.94	0.77	0.94	0.73	0.77	0.80	0.79	0.79	0.79	0.79	0.78		0.98	0.81
14	AMBER99	0.73	0.92	0.75	0.92	0.72	0.75	0.78	0.77	0.77	0.77	0.76	0.76	0.98		0.78
15	TRIPOS	0.59	0.68	0.59	0.64	0.57	0.61	0.62	0.62	0.62	0.62	0.61	0.60	0.81	0.78	

Figure 1. Correlation of van der Waals (a) and electrostatic (b) terms for the different force fields in use.

Waals and Coulombic energies as computed with each of the force fields. We first calculated the respective components of the binding affinity as the difference between the nonbonded interactions observed in the complex and the ligand and protein and then computed the Pearson correlation for each pair of values. As can be seen in Figure 1, the van der Waals and electrostatic terms of all the investigated force fields considered were highly correlated. The only force field that appeared to have less of a correlation is Tripos. This was not unexpected as the recommended charge treatment for its use is through formal charges, while all other force fields consider some kind of partial charges. MMFF94 also appeared to produce van der Waals energies that are less correlated to the other force fields. The van der Waals functional form, a buffered 14−7 Lennard−-Jones, is different from the more traditional Lennard−Jones 12−6 or 9−6 used by the other force fields.

RankScore 2.0, 3.0, and 4.0, Novel Scoring Functions. From that initial study, it became clear that any force field would most probably perform as well when implemented in a more complex scoring function. We therefore turned our attention to the publicly available AMBER/GAFF force field already implemented in our docking program FITTED 2.6.[14] An ideal scoring function would accurately predict the binding affinity of any given ligand for any given protein or nucleic acid target. In practice, this can be reduced to the prediction of the free energy of binding ($\Delta G_{\text{binding}}$) representing the complex formation illustrated in Figure 2. When developing such a function, many researchers relied on the additivity of contributions.[23,26−30] Following this approximation, we broke apart the free energy of binding into the change in entropy and enthalpy measured upon complex formation. As this Michaelis complex forms in water, these contributions are accompanied by a change in solvation (eqs 1 and 2)

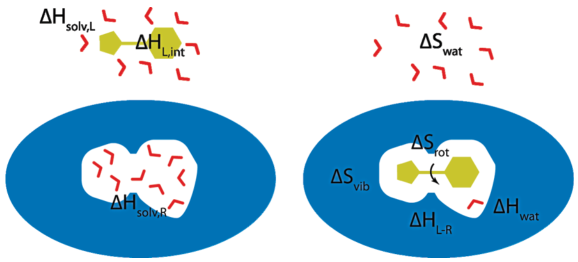


Figure 2. Ligand (yellow)/protein (blue) complex formation. $\Delta H_{\text{L-R}}$ indicates the direct interaction between ligand and receptor; $\Delta H_{\text{solv,L}}$ and $\Delta H_{\text{solv,R}}$ denote the (de)solvation energy of the ligand and the receptor, respectively; $\Delta H_{\text{L,int}}$ indicates the internal strain of the ligand; ΔH_{wat} refers to the stabilization provided by bridging water molecules; ΔS_{wat}, ΔS_{rot}, and ΔS_{vib} refer to the entropic costs of reorganizing water molecules and freezing rotational and vibrational degrees of freedom, respectively.

$$\Delta G_{\text{binding}} = \Delta G_{\text{complex formation}} + \Delta G_{\text{solvation}} \tag{1}$$

$$\Delta G_{\text{binding}} = \Delta H_{\text{complex}} - \Delta H_{\text{ligand}} - \Delta H_{\text{protein}} - T\Delta S_{\text{complex}} + T\Delta S_{\text{ligand}} + T\Delta S_{\text{protein}} + \Delta G_{\text{solvation}} \tag{2}$$

RankScore Formalism. Each of these contributions to the free energy of binding had next to be computed as accurately and quickly as possible. To do so, we implemented various approaches into FITTED and tried a number of combinations. The first three terms of the right-hand side of eq 2 were computed using the GAFF force field (eqs 3−5). In the present study, we assume the protein potential energy (internal energy) to be constant (C in eq 5) as the energetic contribution of protein conformational changes would be difficult to evaluate accurately with high throughput. It is clear that C should be different from one complex to the next as the protein may adjust its conformation upon ligand binding. However, considering the large number of atoms and degrees of freedom within a protein, an accurate evaluation of the protein internal energy would be very time consuming. It is worth recalling that discriminating active from inactive compounds requires prediction of the free energy of binding with an error below a very few kcal/mol, while the potential energy of a protein would be orders of magnitude higher. We have not investigated this term further.

$$\Delta H_{\text{complex}} = E^{\text{FF}}_{\text{complex}} \tag{3}$$

$$\Delta H_{\text{ligand}} = E^{\text{FF}}_{\text{ligand(unbound)}} \tag{4}$$

$$\Delta H_{\text{protein}} = E^{\text{FF}}_{\text{protein+water}} + C \tag{5}$$

In practice, these contributions were computed following these steps. (i) Optimization through conjugate gradient energy minimization of the ligand pose within the protein binding site using FITTED. (ii) Computation of the ligand internal energy and protein+water molecules/ligand intermolecular energy (eqs 3 and 5). (iii) Optimization of the ligand in vacuo and computation of the ligand potential energy (eq 4). As the developed scoring function is to be used with FITTED, we thought that optimizing the pose using the function implemented in FITTED would be more representative of a docked pose.

The fourth term of eq 2 was computed by penalizing the number of rotatable bonds (N_{rot} in eq 6) that are frozen upon binding, defined as all single bonds not in rings. Many scoring functions give each rotatable bond the same penalty; however, as already mentioned by Eldrigde et al., we considered that some bonds become "more frozen" than others when a ligand binds.[29] First, it is well known that a long hydrophobic chain does not move as freely as a polar chain in water and is less frozen than hydrogen-bonded groups when bound to a protein. Second, a portion of a ligand in close contact with a protein cannot move as much as a solvent-exposed group. In order to account for these two aspects, we developed a function shown in eq 6 where the penalty given to the ligand is a function of the number of rotatable bonds, the polarity of the bonds, and their buriedness in the protein:

$$\Delta S_{\text{lig}} = \sum_{i=0}^{N_{\text{rot}}} \text{polarity}_i \times \text{frozen}_i \tag{6}$$

with N_{rot} specifying the number of rotatable bonds in the ligand and for a given bond i:

$$\text{polarity}_i = \begin{cases} 1.0 \text{ if} & \text{polar} \\ 0.5 \text{ if} & \text{semipolar} \\ 0.0 \text{ if} & \text{apolar} \end{cases}$$

$$\text{frozen}_i = \text{frozen}_{i,\text{atom1}} + \text{frozen}_{i,\text{atom2}}$$

$$\text{frozen}_{i,\text{atom}j} = \begin{cases} 0.5 \\ 0.25 \\ 0.0 \end{cases}$$

Two strategies accounting for bond polarity were evaluated. Either the atom types of the rotatable bonds (identified as "polar", "semipolar", or "apolar") or the solvation energy (GB/SA) of the entire ligand were used as polarity descriptors. Only the first of these two descriptors led to an increase in accuracy when added to the scoring function.

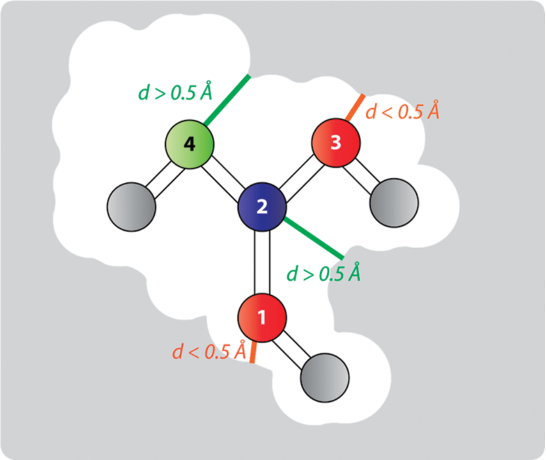


Figure 3. Frozen bond determination. Atoms 1 and 3 have van der Waals contacts with the protein (distance between van der Waals surfaces < 0.5 Å), while atoms 2 and 4 do not. In this case, the bond between atoms 1 and 2 would have a frozen value of 0.75, while the bond between atoms 2 and 4 would have a value of 0.25.

The freezing of the bond is next evaluated and defined by the value "frozen". Once more, two options were evaluated: (i) the number of protein atoms within a certain distance of the "frozen" bond (e.g., 10 Å) or (ii) the presence or absence of a contact with the protein. The former value was used as a descriptor of the buriedness of the rotatable bond in the protein binding site, while the latter is a measure of the freezing effect of the neighboring protein atoms. Although the former had a negligible effect on the scoring function accuracy, the latter significantly increased it. In this formalism, a bond was defined as completely frozen (frozen = 1.0) if the two atoms making this bond were within 0.5 Å of any protein atom van der Waals surface, each atom contributing 0.5 to frozen. If one of the atoms is not in close contact with the protein, the atoms covalently bound to atom 1 are examined. If at least one of these connected atoms is within 0.5 Å of any protein atom van der Waals surface, then the atom contributes 0.25 to frozen. Otherwise, if none of the atoms connected are in close contact with the protein, the atom does not contribute to frozen. In the example in Figure 3, atoms 1 and 3 have contacts with the protein, while atoms 2 and 4 do not. When considering the frozen value of bond 1−2, atom 1 would contribute 0.5 while atom 2 would contribute 0.25 (for being attached to atom 3, having a contact with the protein), yielding a frozen value of 0.75 for that bond. Analogously, bond 2−4 would have a frozen value of 0.25, stemming from the contribution of atom 2 alone.

Conformational entropy loss upon ligand binding has two major components. First, the potential energy surface well in which a given conformation lies may get narrower. Second, some wells may disappear in the presence of the protein. Clearly, although strategies we investigated may lead to a more accurate description of the ligand entropy change than a function of N_{rot} alone, they evaluate the first component and not the second one, which would require a more exhaustive search of the ligand potential energy surface. This has indeed been proposed and implemented within AutoDock.[31]

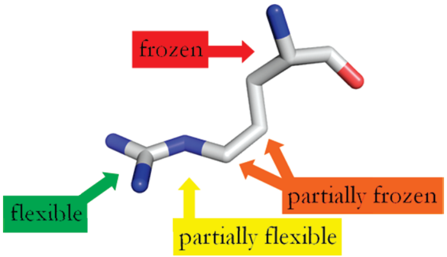


Figure 4. Flexibility of side-chain atoms. The further away from the peptide backbone an atom is, the more flexible it is considered; therefore, the entropic penalty upon binding will be larger.

The change in entropy of the protein upon binding (sixth term of eq 2) was accounted for by reducing the interaction between ligand atoms and flexible side chains. This approach discussed previously[14,23] has been shown to increase the accuracy of the previous version of RankScore.[23] In the previous implementation, each side chain atom was assigned a scaling factor (λ in eq 7) ranging from 0.6 to 1.0, which was next used to scale down the nonbonded interactions. In this early implementation, the value of λ was residue-dependent (0.6 for arginine and 1 for alanine). However, interaction of the ligand with the beta carbon of an arginine does not lead to the freezing of the entire side chain and should not be penalized much, while interactions with the guanidinium group of this same side chain significantly reduce the mobility of the whole side chain. To account for this fact, the scaling factor λ is now dependent on the location of the atom on the side chain (Figure 4).

$$\Delta S_{\text{prot}} = \sum_{\text{non-bondpairs}} \lambda E^{\text{FF}}_{\text{prot-lig}} \tag{7}$$

Finally, the solvation/desolvation contribution was evaluated using a generalized Born (GB) approach combined with an evaluation of the change in solvent-accessible surface area, known to be proportional to the apolar change in solvation (GB/SA[32]). For this purpose, GB/SA has been fully implemented in FITTED.

$$\Delta G_{\text{solvation}} = f(\text{SASA}, \text{GB}) \tag{8}$$

Development of RankScore 2 and RankScore 3. The value for each of these terms was computed for the 209 complexes of the self-docking set 1 and for the nearly 1000 complexes of the cross-docking set 2.[24] Next, random weights between 0 and 1 were generated for each term, and the Kendall τ measuring the correlation between the predicted and observed rankings was computed (Figure 5). This strategy will therefore lead to a scoring function optimized to predict the ranking of compounds and not to reproduce binding free energies. These steps were repeated 25 000 times, and the corresponding table with 25 000 entries was sorted by decreasing τ. In order to ensure a better transferability, we did not immediately select the best one but instead opted for an iterative approach. To do so, statistical analysis of the sets of weights leading to the top 4% (1000 entries) correlation coefficients was carried out (Figure 6). On the basis of this information, the ranges for the different coefficients were constrained to the weights with the higher chance of leading to a better correlation with experimental binding affinities, and the protocol described in Figure 5 reiterated with the new ranges. For example, in a second iteration the coefficients for the van der Waals interaction were randomly generated between 0.00 and 0.40.

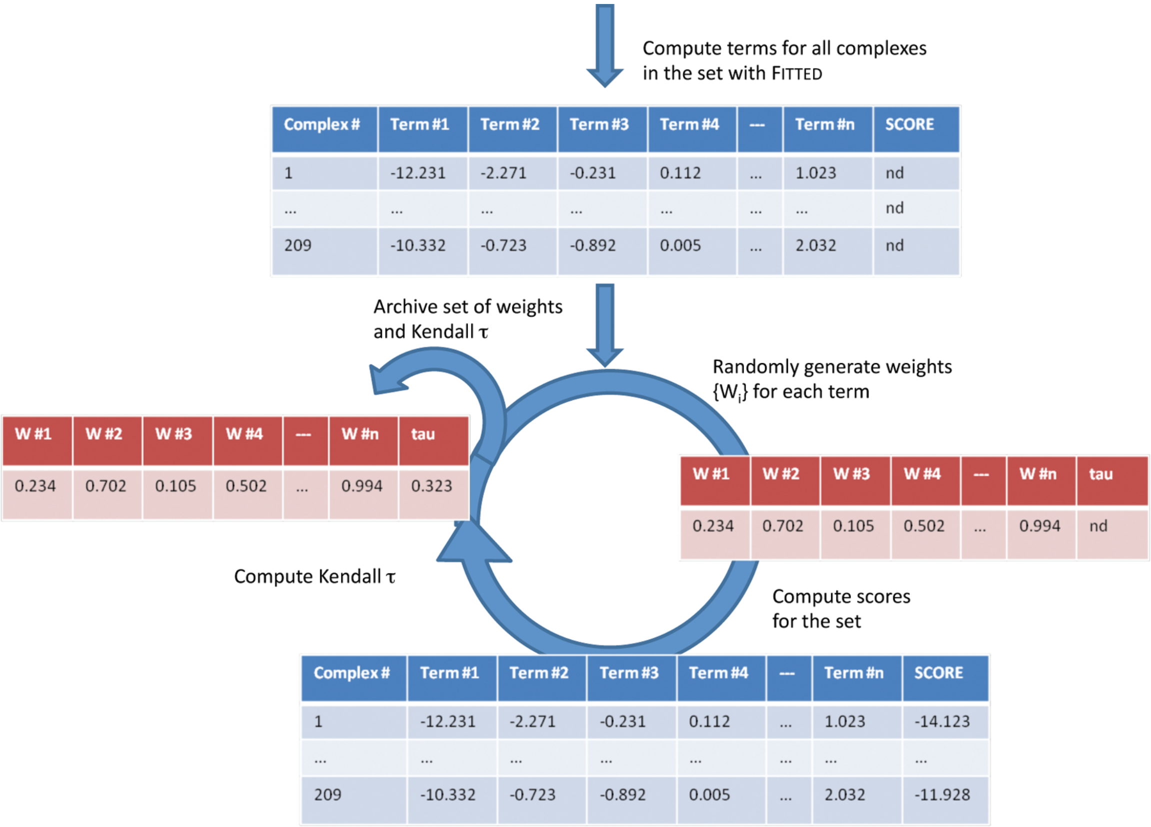
After five iterations of the entire protocol, RankScore 2 was produced (eq 9) and a Kendall τ of 0.37 was computed. This value is similar to those obtained previously with the same training set and the most accurate scoring functions assessed.[24] The same procedure was carried out with set 2 (cross-docked structures). Expectedly, as cross-docked structures are not as accurate as crystal structures (i.e., the signal might be buried under the noise of the data), the trends (as the one shown in Figure 6) were not as marked and the function (referred to as RankScore 3) derived from this set is not expected to be as accurate as RankScore 2. In fact, most of the terms did not show any preferred range of values, and none of the random set of weights led to Kendall τ as high as those observed with the previous set and RankScore 2. Interestingly, these two functions (RankScore 2 and 3) were found to be very similar and RankScore was not considered further

$$\begin{aligned}\text{RankScore 2} = {} & 0.680E^{\text{vdW}}_{\text{complex}} + 0.040E^{\text{elec}}_{\text{complex}} + 0.100E^{\text{HB}}_{\text{complex}} \\ & + 0.000\Delta G_{\text{GB}} + 0.100\Delta G_{\text{SASA}} + 0.040N_{\text{wat}} + \\ & + 0.450(N_{\text{rot}} + 2\sum_{\text{bonds}} f(N_{\text{rot}}, \text{polarity}, \text{contact}))\end{aligned} \tag{9}$$

with N_{wat} being the number of captured water molecules and N_{rot} being the number of rotatable bonds. Interestingly, the weight for the polar contribution to the solvation was found to be very low, and setting it to zero reduces the computation time necessary to compute a score while not affecting the accuracy. This observation is consistent with our previous report.[23]

Development of RankScore 4. The last scoring function was developed using a slightly different approach. Six proteins (purine nucleoside phosphorylase, PNP; acetylcholinesterase, AC; neuraminidase, NA; estrogen receptor, ER; trypsin; P38 map kinase, P38) and the corresponding decoys and active compounds were selected from the DUD set.[33] For each of the proteins, three or four conformations cocrystallized with different ligands were considered. All the compounds were docked using FITTED in the flexible protein mode and the iterative protocol illustrated in Figure 5 applied. However, instead of computing τ for the correlation between scores and binding affinities, we computed the area under a receiver operating characteristic (ROC) curve for the retrieval of known actives.

Once more, the protocol was iterated five times with increasingly smaller ranges of coefficient values. At the end of this procedure RankScore 4 was derived (eq 10) with ROC values of 0.85 (PNP), 0.47 (AC), 0.73 (NA), 0.87 (ER), 0.95 (trypsin), and 0.67 (P38) with an average of 0.79. A close look of the functional forms of RankScore 2 and 4 leads to some interesting observations. While in RankScore 1, 2, and 3, the electrostatic interaction term was almost turned off, it became a major term in RankScore 4. Conversely, the van der Waals coefficient in RankScore 4 is roughly one-fourth of the one used in the other scoring functions. Thus, it clearly appears that RankScore 2 and 4 are capturing very different information. In fact, when RankScore 2 was applied to these same six proteins and ligand sets (see Table 1), the ROC values were much lower (0.82, 0.40, 0.42, 0.67, 0.73, and 0.47) with an average of 0.56. In addition, RankScore 2


Compute terms for all complexes in the set with FITTED
Archive set of weights and Kendall τ
Randomly generate weights {W$_i$} for each term
Compute Kendall τ
Compute scores for the set

Complex #	Term #1	Term #2	Term #3	Term #4	---	Term #n	SCORE
1	-12.231	-2.271	-0.231	0.112	...	1.023	nd
...	...	...	...	...	...	...	nd
209	-10.332	-0.723	-0.892	0.005	...	2.032	nd

W #1	W #2	W #3	W #4	---	W #n	tau
0.234	0.702	0.105	0.502	...	0.994	0.323

W #1	W #2	W #3	W #4	---	W #n	tau
0.234	0.702	0.105	0.502	...	0.994	nd

Complex #	Term #1	Term #2	Term #3	Term #4	---	Term #n	SCORE
1	-12.231	-2.271	-0.231	0.112	...	1.023	-14.123
...	...	...	...	...	...	...	...
209	-10.332	-0.723	-0.892	0.005	...	2.032	-11.928

Figure 5. Procedure used to derive RankScore 2 and RankScore 3. In each case, 25 000 loops were performed, yielding an identical amount of sets of weights with their associated Kendall τ value.

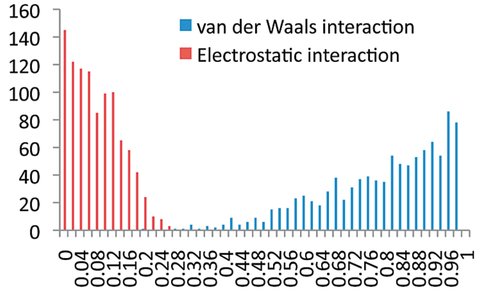


Figure 6. Distribution of scaling factors for van der Waals (blue) and electrostatic (red) interactions.

provided rankings that were not better than random for AC, NA, and P38, while only AC remained problematic even after extensive training of RankScore 4.

$$\text{RankScore 4} = 0.184\cdot E^{\text{vdW}}_{\text{complex}} + 0.746\cdot E^{\text{elec}}_{\text{complex}} + 0.595\cdot E^{\text{HB}}_{\text{complex}} + +0.000\cdot\Delta G_{\text{GB}} + 0.160\cdot\Delta G_{\text{SASA}} + 0.050\cdot N_{\text{wat}} + +0.664\cdot\sum_{\text{bonds}} f(N_{\text{rot}}, \text{polarity}, \text{contact}) \quad (10)$$

Application of the RankScore Scoring Functions to Benchmark Sets. In order to evaluate the predictive power and compare RankScore 2 to other available functions, we applied it to Wang's set of 100 protein/ligand complexes.[34] To our surprise, a close look at this set revealed some discrepancies and prompted us to curate it. Covalent inhibitors were removed, and metal ions were added to metalloenzymes. With this set in hand, we applied 15 scoring functions including some that Wang and co-workers used as well as ligand molecular weight, RankScore 2, and RankScore 4. In a previous report, we found that the accuracy of the scoring functions both Wang et al. and us looked at correlated well except for ChemScore.[24] With the set cleaned, accuracies obtained with both sets now correlate well. As shown in Figure 7, RankScore 2 stands within the best scoring functions, behind X-Score and DrugScore, and within range of ChemScore, DockScore, and PLP2. More interestingly, the correlation of the scores calculated with RankScore 4 was among the least predictive of all the scoring functions considered. It is worth recalling that RankScore 2 has been developed to rank-order active compounds (as in this set), while RankScore 4 has been trained to discriminate between active and inactive compounds.

As an additional validation, RankScore 2 and RankScore 4 were also applied to the screening of libraries of known actives and decoys against thymidine kinase (TK), HIV-1 protease (HIVP), thrombin (THR), cyclin-dependent kinase 2 (CDK2), and HIV-1 reverse transcriptase (HIVRT). As can be seen in Table 1, RankScore 4, developed for this specific purpose, is much more accurate than RankScore 2, which was developed to reproduce binding affinities. These discrepancies indicate that distinct scoring functions for hit identification and lead optimization should be developed.

CONCLUSIONS

It is well established that molecular mechanical force-field energy on a single conformation is often not sufficient to provide a predictive tool for the fast estimation of the binding energy of ligands to proteins. However, addition of other terms simulating other aspects of the energetics of binding to the equation provides more reliable methods. In a first

Table 1. Area Under the Curve of a Receiver-Operating Characteristic for the Docking of Libraries Containing About 1000 Ligands and Decoys to 11 Proteins (6 in the training set and 5 in the testing set)

	training set							testing set				
	PNP	AC	NA	ER	Trypsin	P38	TK	HIVP	THR	CDK2	HIVRT	average
RankScore 2	0.82	0.40	0.42	0.67	0.73	0.47	0.56	0.63	0.63	0.65	0.61	0.60
RankScore 4	0.85	0.47	0.73	0.87	0.95	0.67	0.88	0.87	0.90	0.83	0.65	0.79

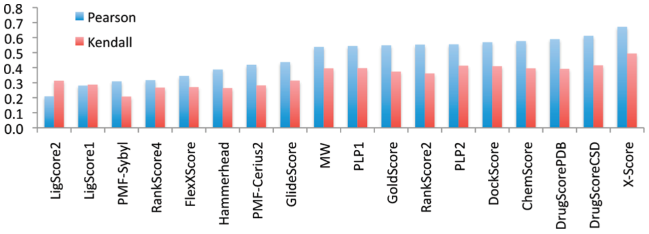


Figure 7. Pearson correlation and Kendall τ for a variety of scoring functions when applied to the testing set of 93 complexes.

section we have shown that the intermolecular potential energies computed with many common force fields are highly correlated, showing that any force field would potentially perform as well at predicting ligand binding affinities. We next developed RankScore 2 and RankScore 3, built around the general Amber force field (GAFF), from an iterative process that optimized the scoring function weights in order to maximize the correlation between the ranked lists of calculated scores and experimental binding affinities. Validation of RankScore 2 against an independent set of protein−ligand complexes indeed found it to perform better or as well as many commonly used scoring functions. We then trained RankScore 4 to discriminate between active and inactive compounds, as required for a high-throughput virtual screening campaign, in an analogous iterative fashion. Testing of RankScore 2 and RankScore 4 on other unrelated libraries of ligands and proteins revealed that RankScore 4 performs significantly better than RankScore 2 at discriminating between actives and decoys, indicating that scoring functions for VS applications should not be developed from active compounds only.

EXPERIMENTAL SECTION

Preparation of the Training Set Structures. The preparation of the training set has been described in a previous report.[24] Succinctly, it involved (i) removal of all water molecules except the ones making bridging interactions (at least 3 hydrogen bonds) with both ligand and protein, (ii) assignment of appropriate protonation states to both ligand and protein side chains, and (iii) constrained optimization by energy minimization of the ligand with a force field.

Derivation of Additional Parameters for Force Fields. Some of the force fields did not contain some parameters that were relevant to the calculations, so *ad hoc* parameters were derived for them. Some of the ligands featured moieties that were not included in the original force-field parametrization; however, they were deemed fairly rigid and, starting from a crystalline structure, not much optimization would be necessary. Parameters (mostly bond stretch, and torsions) for these moieties were derived by defining the new parameters so as to conserve the values observed in the crystal structure. For bond stretches, the equilibrium distance (r_0) was defined as the average of the interatomic distance observed in all the molecules containing the particular moiety in the training set, with a stretch constant defined by analogy with another pair of atoms existing in the force-field definition, or by default a large stretch constant to keep the bond stiff. The same strategy was used for bending and torsional parameters when needed. This study focuses on nonbonded interactions; hence, the guessed parameters should not have much impact on the final result.

Force-Field Charges. The force fields included for use with Discover use bond-charge increments to assign partial charges, while Macromodel uses a bond dipole definition. When charge definitions were missing from the force-field definition, semiempirical calculations using the AMPAC module in Insight II were performed on model molecules (e.g., for carbamates, *N*-methyl-methoxycarbamate was used) to determine appropriate bond increments. The bond increments were defined appropriately to reproduce the charge distribution observed by the semiempirical method.

Development and Validation of RankScore 2 and RankScore 4. The sets used to develop RankScore 2 and RankScore 3 are those previously reported and were used with no further modifications. The set of protein−ligand complexes reported by Wang et al.[34] used to validate RankScore 2 was curated by (a) removing covalent complexes (1a46, 1a5g, 1ba8, 1bb0, 1exw, 1yyy, 1zzz). In addition, missing metal atoms in metalloproteins (1af2, 1bzm, 1cbx, 1e96, 1mnc, 1tlp, 1tmn, 2ctc, 2tmn, 2xim, 2xis, 3cpa, 3tmn, 4tln, 4xia, 5p21, 5tln, 7tln, 8xia) were readded from the respective Protein Data Bank[35] entries. Scores were recalculated following the protocols recommended by the developers of the different scoring functions and described previously.[24] Proteins, active compounds, and decoys used to derive and validate RankScore 4 were retrieved from the DUD library.[33] A maximum of 1000 decoys were selected for each of the 11 proteins considered (AC: 1e66, 1gpn, 1h22, 1h23; CDK2: 1pxp, 1dm2, 1aq1, 1pxn; ER: 1sj0, 1err, 3ert; HIVP: 1b6l, 1hpo, 1hpv, 1pro; HIVRT: 1vrt, 1fk9, 1rt1, 1c1b; NA: 1f8d, 1f8e, 2qwe, 2qwf; P38: 1a9u, 1w7h, 1w82, 1w84; PNP: 1b8n, 1b8o, 1v48; THR: 1dwc, 1etr, 1tmt, 1ett; TK: 1e2k, 1ki3, 1of1, 2ki5; Trypsin: 1f0u, 1ghz, 1o2h, 1qb9).

The ligands from the various sets were docked using the FITTED 2.6 suite, and all the energy terms were output and tabulated. Python scripts were used to randomly generate the weights, compute the scores, and either Kendall τ (RankScore 2) or area under ROC curves (RankScore 4). Another Python script was used to analyze the generated data and define the range for the next cycle (see Figure 3).

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. P.E. was supported by a scholarship from the CIHR (Strategic Training Initiative in Chemical Biology) and a J. W. McConnell Memorial Fellowship while performing the research.

Supporting Information Available: Correlation values for all scoring functions considered in the validation set. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303*, 1813–1818.

(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) Mobley, D. L.; Dill, K. A. Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get. *Structure* **2009**, *17*, 489–498.

(4) Kollman, P. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* **1993**, *93*, 2395–2417.

(5) 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.

(6) Hansson, T.; Marelius, J.; Åqvist, J. Ligand binding affinity prediction by linear interaction energy methods. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 27.

(7) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 409–443.

(8) Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11*, 580–594.

(9) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765–784.

(10) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.* **1986**, *7*, 230–252.

(11) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz Jr, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(12) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639–1662.

(13) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411.

(14) 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.

(15) Clark, M.; Cramer, R. D., III; Opdenbosch, N. V. Validation of the general purpose Tripos 5.2 force field. *J. Comput. Chem.* **1989**, *10*, 982–1012.

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

(17) Mayo, S. L.; Olafson, B. D.; Goddard, W. A., III. DREIDING: A generic force field for molecular simulations. *J. Phys. Chem.* **1990**, *94*, 8897–8909.

(18) Ewig, C. S.; Berry, R.; Dinur, U.; Hill, J.-R.; Hwang, M.-J.; Li, H.; Liang, C.; Maple, J.; Peng, Z.; Stockfisch, T. P.; Thacher, T. S.; Yan, L.; Ni, X.; Hagler, A. T. Derivation of class II force fields. *J. Comput. Chem.* **2001**, *22*, 1782–1800.

(19) Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: A novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graphics Modell.* **2003**, *21*, 289–307.

(20) Case, D. A.; Darden, T. A.; T.E. Cheatham, I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER 10*; University of California, San Francisco: San Francisco, CA, 2008; http://ambermd.org.

(21) 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.

(22) 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.

(23) 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.

(24) Englebienne, P.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 4. Are Popular Scoring Functions Accurate for this Class of Proteins. *J. Chem. Inf. Model.* **2009**, *49*, 1568–1580.

(25) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.

(26) 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.

(27) Jain, A. N. Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 427–440.

(28) 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.

(29) 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.

(30) 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.

(31) Lee, J.; Seok, C. A statistical rescoring scheme for protein-ligand docking: Consideration of entropic effect. *Proteins: Struct., Funct., Genet.* **2008**, *70*, 1074–1083.

(32) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.

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

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

(35) 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.

CI900251K