Using Buriedness To Improve Discrimination between Actives and Inactives in Docking

Noel M. O'Boyle,*,† Suzanne C. Brewerton,‡ and Robin Taylor†

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, U.K., and Astex Therapeutics, Ltd., 436 Cambridge Science Park, Milton Road, Cambridge CB4 0QA, U.K.

Received February 11, 2008

A continuing problem in protein−ligand docking is the correct relative ranking of active molecules versus inactives. Using the ChemScore scoring function as implemented in the GOLD docking software, we have investigated the effect of scaling hydrogen bond, metal−ligand, and lipophilic interactions based on the buriedness of the interaction. Buriedness was measured using the receptor density, the number of protein heavy atoms within 8.0 Å. Terms in the scaling functions were optimized using negative data, represented by docked poses of inactive molecules. The objective function was the mean rank of the scores of the active poses in the Astex Diverse Set (Hartshorn et al. *J. Med. Chem.*, **2007**, *50*, 726) with respect to the docked poses of 99 inactives. The final four-parameter model gave a substantial improvement in the average rank from 18.6 to 12.5. Similar results were obtained for an independent test set. Receptor density scaling is available as an option in the recent GOLD release.

INTRODUCTION

Protein−ligand docking software has the potential to identify promising lead compounds at an early stage of the drug discovery pipeline. It aims to computationally mimic the biological process of a ligand binding to a protein, to provide either an estimate of the binding affinity or simply a score reflecting the predicted strength of binding. These programs can be thought of as consisting of two principal parts: a search algorithm, which places the potential ligand in multiple positions or poses in the binding site, and a scoring function,[1] which calculates a score for each pose.

When docking software is used for virtual high-throughput screening, the practical problem is to correctly rank the true ligands (actives) with respect to nonbinders (inactives). This is still a difficult problem, as shown by a recent large-scale docking study by Warren et al.[2] Several factors contribute to this difficulty: scoring functions use empirical formulas and approximations in the interest of speed; their success is based upon the assumption that binding affinity can be described as a sum of independent terms; and entropic effects are largely disregarded.[3] The alternative to using an empirical scoring function is to calculate binding energies from first principles, but despite advances in this area[4] such calculations are too slow for a virtual screen and require accurate calculation of a small difference between two very large energies (the energies of the protein and ligand separately and as a complex).[5]

Another reason for the difficulty in ranking actives with respect to inactives is that scoring functions are usually trained only with positive data, that is, information relating to known binders. Negative data, or information relating to inactive molecules or inactive poses, is rarely included. To use terminology from the field of Quantitative Structure−Activity Relationships (QSAR), inactive molecules are outside the Domain of Applicability of the scoring function.[6]

Graves et al.[7] have highlighted the fact that negative data can be used to improve docking algorithms by identifying particular weaknesses of an algorithm. They refer to two types of negative data: 'geometric decoys', incorrect poses of a known active which score better than the crystallographic pose; and 'hit list decoys', molecules which experimentally do not bind to a particular target but which are ranked highly in a virtual screen. Smith et al.[8] were the first to use negative data to optimize terms in a scoring function. For negative data, they used 200 diverse poses of 120 decoy molecules for 20 proteins in a training set. Using a genetic algorithm, they optimized the coefficients in a scoring function where the objective function was the average rank of the score of the crystallographic pose of the native ligand with respect to the scores of the decoy poses. Whereas the previous studies focused on improving the results from rescoring docked poses, Pham and Jain[9] optimized the weights of two penalty terms in the Surflex-Dock scoring function[10] through an iterative docking process. Their objective function included the deviation from the experimental binding affinity for the known actives (positive data) as well as the scores of high scoring decoy molecules (negative data). This method allowed the estimation of parameter values which had previously received little weight when the training involved only positive data. Negative data has also been used in the development of knowledge-based scoring functions. Huang and Zou[11] used an iterative procedure to optimize their knowledge-based scoring function ITScore to rank the correct pose ahead of decoy poses of the same molecule.

Stahl[12] described a modification of the hydrogen bond term in the FlexX scoring function,[13] where a sigmoidal function was used to scale hydrogen bonds scores according to a solvent accessibility value for the protein atom. Parameters for the scaling function were chosen empirically based on docking performance on a series of in-house data sets. The

* Corresponding author phone: +44-1223762531; fax: +44-1223336033; e-mail:oboyle@ccdc.cam.ac.uk.
† Cambridge Crystallographic Data Centre.
‡ Astex Therapeutics, Ltd.

10.1021/ci8000452 CCC: $40.75
Published on Web 06/06/2008

effect of the scaling function was to reduce the contribution of hydrogen bonds to the overall score. The tailored scoring function showed improved enrichment when applied to a thrombin data set.

Here we focus on improving the discrimination between actives and inactives for the ChemScore scoring function[14,15] used by the docking software GOLD.[16,17] We introduce new terms into the scoring function whose effect is to scale the calculated scores for particular interactions depending on the degree of buriedness. Our hypothesis is that the introduction of these new terms will allow scores to be adjusted to take into account entropic or enthalpic effects empirically; interactions deep in the active site should receive an increased score, while those close to the solvent should be scored less favorably. Our measure of buriedness is the receptor density, the number of protein heavy atoms within a particular distance of the interaction location. We incorporate negative data using a method similar to that of Smith et al.[8] The new parameters in our scaling function were chosen to minimize the mean rank of the actives with respect to inactives for a data set of 85 proteins. This allows us to directly train the scoring function to improve the discrimination between actives and inactives.

METHODS

ChemScore. The ChemScore scoring function as implemented in GOLD is an empirical scoring function consisting of a sum of several positive terms relating to binding affinity as well as some negative terms[18]

$$\text{score} = \Delta G_{\text{binding}} + E_{\text{clash}} + E_{\text{int}} + E_{\text{cov}} \quad (1)$$

where E_{clash} is an energetic penalty for ligand-protein clashes, E_{int} is a penalty for the internal energy of a ligand, and E_{cov} is a term used to handle covalently bound ligands. $\Delta G_{\text{binding}}$ is an estimate of the free energy of binding and is defined as follows

$$\Delta G_{\text{binding}} = \Delta G_0 + \Delta G_{\text{hbond}} S_{\text{hbond}} + \Delta G_{\text{lipo}} S_{\text{lipo}} + \Delta G_{\text{metal}} S_{\text{metal}} + \Delta G_{\text{rot}} H_{\text{rot}} \quad (2)$$

where S_{hbond}, S_{lipo}, and S_{metal} are the calculated total scores for hydrogen bond, lipophilic, and metal interactions formed by the molecule. H_{rot} is the calculated value for the loss of configurational entropy of the ligand on binding to the protein. ΔG_0 and the ΔG coefficients are constants whose values were determined by Eldridge et al. by multilinear regression against the binding affinities of a training set of 82 protein−ligand complexes.[14] The values of S_{hbond}, S_{lipo}, and S_{metal} are typically calculated as a sum of scores from several hydrogen bond, lipophilic, and metal interactions formed by a molecule:

$$S_{\text{hbond}} = \sum s_{\text{hbond}} \quad S_{\text{lipo}} = \sum s_{\text{lipo}} \quad S_{\text{metal}} = \sum s_{\text{metal}} \quad (3)$$

Receptor density, ρ. We use the receptor density, ρ, as a measure of buriedness. We define the receptor density at a point $\mathbf{x}$, $\rho_r(\mathbf{x})$, as the number of protein heavy atoms within a distance r of $\mathbf{x}$. The receptor density value for a protein−ligand hydrogen bond interaction was calculated at the protein donor or acceptor coordinates. In the ChemScore implementation in GOLD, the contribution of lipophilic interactions to the score is precalculated on a grid over the active site, so the receptor density at the ligand atom involved in the lipophilic interaction was used.

Note that the S_{hbond}, S_{lipo}, and S_{metal} scores for each molecule may be a sum of several interactions (see eq 3) involving different atoms each with its own receptor density value.

Astex Diverse Set. Protein−ligand structures were taken from the test set published by Hartshorn et al.,[19] which is referred to here as the Astex Diverse Set. This set contains 85 high-quality crystal structures of a diverse range of proteins with druglike ligands. In the data set, waters have been removed, and protein hydrogens have been added to the active sites. The protein and ligand structures were used as provided, without any additional processing. This data set is freely available for download from http://www.ccdc.cam.ac.uk/free_services/free_downloads/. Where PDB identifiers are referred to in the Results and Discussion sections below, the corresponding protein name and ligand can be found in Table 2 of ref 19. Figures of ligands in the protein active site were created with Jmol.[20]

Investigation of Difference in Receptor Density Values between Actives and Inactives. A data set of active and inactive poses was created by docking each ligand in the Astex Diverse Set to its corresponding protein using the ChemScore scoring function in a development version of GOLD 4.0 (30000 genetic algorithm operations). The 'diverse solutions' option was used to generate 10 diverse poses with a minimum pairwise similarity of greater than 1.5 Å rmsd. All those poses within 2 Å rmsd of the crystal structure geometry were considered 'active poses' or positive data, while all those more than 4 Å rmsd from the crystal structures were considered 'inactive poses' or negative data. The value of 2 Å rmsd is widely used to measure success rates of pose prediction in docking comparisons,[2,19] while 4 Å rmsd is greater than the 3 Å value used by Graves et al.[7] to detect incorrect poses ('geometric decoys'). Proteins without at least one active and one inactive pose were discarded from the final set which consisted of 64 proteins and an average of 3.5 active poses and 3.9 inactive poses per protein.

For each pose, the scores of all lipophilic and hydrogen bond interactions were collated, along with the corresponding values of ρ_r for several values of r from 2.5 to 9.0 Å. These data were analyzed to determine the value of r with the optimum potential to discriminate actives from inactives.

Optimization of Receptor Density Scaling Functions Using Negative Data. To incorporate the receptor density into the ChemScore scoring function, we introduced two functions f_{HB}, for scaling individual hydrogen bond interactions (s_{hbond} in eq 3), and f_{L}, for scaling lipophilic interactions (s_{lipo} in eq 3), of the form

$$f(\rho, \rho_1, \rho_2, S) = \begin{cases} 0 & \rho \le \rho_1 \\ S\dfrac{\rho - \rho_1}{\rho_2 - \rho_1} & \rho_1 < \rho \le \rho_2 \\ S & \rho > \rho_2 \end{cases} \quad (4)$$

where ρ_1, ρ_2, and S are constants whose values were determined as described below. A constant function $g(\rho) = S$ was also tested for comparison (g_{L} for lipophilic interac-

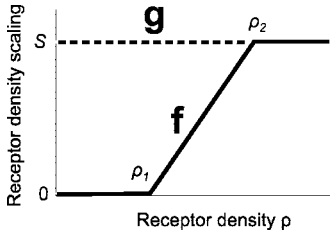

Figure 1. The shape of the functions $f(\rho, \rho_1, \rho_2, S)$ and $g(\rho, S)$.

tions, g_{HB} for hydrogen bonds). Figure 1 depicts the shape of these functions.

A training set of actives and inactives was prepared to help determine the best combination of f or g functions and suitable parameter values for ρ_1, ρ_2, and S. Protein structures and actives were taken from the Astex Diverse Set. For the active molecules, the geometries of the crystal structures were not used directly as they have an artifactually low score compared to the docked inactives. Instead, an exhaustive docking was carried out using GOLD (100 poses, diverse solutions differing by 0.1 Å rmsd, cluster size of 3), and the highest-scoring docked pose within 2.0 Å of the crystal structure was used. For those proteins where no pose within 2.0 Å was found (6 proteins: 1jje, 1t9b, 1uml, 1xm6, 1xoq, and 1ygc), the crystal structures were optimized using the ChemScore scoring function as follows: protein and ligand rotatable hydrogens were optimized in the presence of the crystal structures ("run_flag = LOCAL_OPT") and then the crystal structures were geometry optimized using the simplex method in GOLD ("run_flag = RESCORE retrieve"), a procedure which involves a local optimization of the ChemScore score of the ligand in the presence of a rigid protein (the mean rmsd of the resulting ligand geometries was 0.2 Å). These poses were used as positive data.

For each protein, a set of 99 inactives was selected from the Astex in-house ATLAS database of compounds available for purchase. Artificial enrichment can occur where an inactive data set is chosen with distinct physicochemical properties compared to actives.[21] In order to avoid this problem, we first selected molecules within a Euclidean distance of 2 of each active in terms of a fingerprint consisting of the number of acceptors, the number of donors, and the number of heavy atoms. Of these molecules, we selected 99 which were topologically dissimilar from the actives. Here similarity was measured by a Tanimoto coefficient of hashed element and bond order path fingerprints (256 bits) and a value of 0.6 (empirically determined) was used for maximum allowed similarity. For nine of the actives, insufficient inactives were found with this procedure, and it was repeated using a Euclidean distance of 4. CORINA was used to generate the geometry of the molecule and to generate the protonation state of the molecule at pH 7.5 (using a fixed set of rules).[22]

Negative data were generated by docking each inactive against its respective target. For each protein, the rank of the active was then determined by ordering it and the 99 inactives into descending order of scoring function value; that is, the molecule with the highest score (most likely to be a true active) was ranked 1, whereas that with the lowest score was ranked 100.

The parameter values ρ_1, ρ_2, and S for each function were determined by a brute-force optimization procedure where the objective function was the average rank of active molecules. The initial optimization was carried out using the Python module *scipy.optimize.brute*, part of the SciPy library,[23] followed by a parameter sweep. A number of different combinations of f_L, f_{HB}, g_L, and g_{HB} were optimized as described in the Results section and shown in Table 1. The final chosen model is referred to as the receptor density scaling (RDS) model.

Validation of the Receptor Density Scaling Model. To test for overtraining, we investigated whether we could get equal improvements in mean rank if we replaced the true active with an inactive and then repeated the optimization procedure. This is similar to the idea behind *y*-scrambling in the field of QSAR (see Rücker et al.[24] and references therein). Test Set A was created by generating false actives as follows: each active in the training set was swapped with an inactive that was a neighbor in terms of rank (when scored by the original ChemScore). Odd-numbered actives were swapped with the inactive that scored one rank higher, while even-numbered actives were swapped with the inactive that scored one rank lower. The optimization procedure described in the previous section was repeated with Test Set A for the best performing model identified for the training set (a combination of f_{HB} and g_L).

Based on the results of the original optimization procedure using the training set, the best receptor density scaling (RDS) model was identified. It was tested using Test Set B, which consisted of inactives selected in the same way as the training set but ensuring that there was no overlap between the test and training sets for the same protein.

The best RDS model was also tested using Test Set C, which consisted of a broader molecular weight range of inactives for each protein. All of the inactives from all of the proteins from the training set and Test Set B were pooled and duplicates removed. Next, for each protein, a set of 99 inactives was chosen randomly without replacement from the subset of inactives topologically distinct from the active. Here, topologically distinct means that the Tanimoto similarity of the Daylight-type fingerprints of the active and inactive was less than 0.7 (as measured by the Open Babel library[25]). This gave a set of inactives of broader molecular weight range compared to that used for the optimization procedure and in Test Set B.

RESULTS

Comparison between Receptor Density Values for Interactions Formed by Actives versus Those Formed by Inactives. To simplify our initial study of whether receptor density could be used to distinguish between actives and inactives, we used correct and incorrect poses of the same (active) ligand as our actives and inactives, respectively. This meant that possible confounding effects such as the influence of molecular weight or the number of hydrogen bond donors or acceptors were eliminated. For each pose, the receptor density was calculated at each hydrogen bond and lipophilic interaction point for several values of r from 2.5 Å to 9.0 Å.

Figure 2 is a plot of interaction score versus receptor density for $r = 8.0$ Å ($\rho_{8.0}$) for lipophilic and hydrogen bond interactions. The lines of best fit for both hydrogen bond and lipophilic interactions show that the average value at a

Table 1. Parameter Values for Receptor Density Functions That Minimize the Mean Rank of Actives

receptor density functions used	optimized mean rank of actives	hydrogen bond function term(s)			lipophilic function term(s)		
		ρ_1	ρ_2	S	ρ_1	ρ_2	S
			Training Set				
none	18.6	–	–	–	–	–	–
f_{HB} and f_L	11.5	19	162	3.24	64	146	2.01
f_L	13.9	–	–	–	44	126	0.974
f_{HB}	13.0	31	120	4.98	–	–	–
g_{HB} and g_L	14.0	–	–	1.80	–	–	0.70
f_{HB} and g_L	12.5	13	105	1.80	–	–	0.52
			Test Set A				
none	18.8						
f_{HB} and g_L	18.6	−40	0	0.991	–	–	1.09

particular value of ρ is the same (or only slightly different) whether the interaction is formed by an active or an inactive. It is also clear that the relationship between interaction score and ρ is quite different for hydrogen bond interactions compared to lipophilic interactions. The hydrogen bond score decreases slightly with increasing ρ due to the difficulty of forming favorable hydrogen bond geometries at high ρ. On the other hand, the strength of a lipophilic interaction tends to increase with increasing ρ. Buried portions of the active site tend to be lipophilic. Since ChemScore calculates the lipophilic contribution as a sum over many pairwise interactions, any lipophilic interactions at high ρ tend to have a higher score.

Figure 3 shows the same data as a normalized sum of scores binned by $\rho_{8.0}$ (top) and a normalized cumulative sum (bottom). Each point on the curves in Figure 3 (top) represents the total score of interactions occurring at that receptor density value divided by the total number of interactions. As a result of this normalization procedure, the integrand of each curve is the average score of those interactions. The same data are represented as cumulative plots in Figure 3 (bottom). The final values of the active and inactive curves are the average values for an active and inactive interaction, respectively.

Although inactives and actives both have the same average score for hydrogen bond interactions, Figure 3 shows that the location of the interactions in terms of $\rho_{8.0}$ is different; those interactions formed by the inactives tend to occur at lower values of $\rho_{8.0}$. The cumulative graph is slightly different for lipophilic interactions, as active molecules have a higher average score than inactive molecules. Nevertheless, the same trend in terms of $\rho_{8.0}$ can be seen. These results were the same for a broad range of values of r from 7.0 Å to 9.0 Å. For all subsequent work a value for r of 8.0 Å was used, and references to ρ will indicate $\rho_{8.0}$ unless otherwise stated.

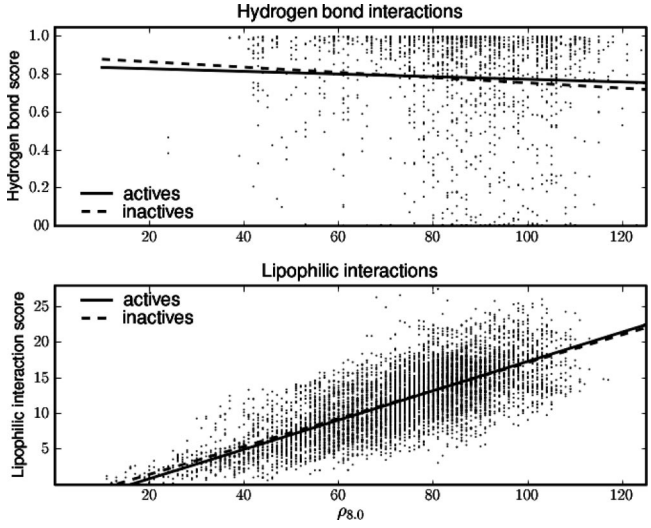

Figure 2. The scores for (top) individual hydrogen bond interactions and (bottom) individual lipophilic interactions at different values of $\rho_{8.0}$. Best fit lines through the points are shown separately for interactions in actives and inactives.

Determination of the Optimal Form and Parametrization of the Receptor Density Scaling Functions. In order to take advantage of the difference between actives and inactives shown by Figure 3, we introduced the scaling function $f(\rho, \rho_1, \rho_2, S)$ (eq 4), for hydrogen bond interactions (f_{HB}) and lipophilic interactions (f_L). These terms adjust the score assigned to a particular interaction depending on the value of ρ at the coordinates of the interaction as shown in Figure 1.

To determine appropriate values for ρ_1, ρ_2, and S, we prepared a training data set containing correctly docked poses of the actives in the Astex Diverse Set and docked poses of 99 inactives per protein. Unlike our initial study, the inactives used here were distinct molecules from the active, selected from a set of compounds available for purchase. The inactives were chosen so that they had similar physicochemical properties as the actives. We then applied a brute-force optimization procedure to optimize the mean rank of the score of each active with respect to the scores of its inactives by varying the values of the parameters in f_{HB} and f_L.

The initial mean rank of the actives (using the original ChemScore function) was 18.6. Here, and in subsequent discussions of rank, a rank of 1 indicates the highest scoring molecule (that is, the molecule predicted to be most active). The optimized value of the mean rank was 11.5 (Table 1). The associated values for the parameters were $\rho_1 = 19$, $\rho_2 = 162$, and $S = 3.24$ for f_{HB} and $\rho_1 = 64$, $\rho_2 = 146$, and $S = 2.01$ for f_L.

To find out which of f_{HB} or f_L was the greater contributor to the improvement in the rank of the actives, the optimization procedure was repeated but only including one of f_{HB} and f_L. As shown in Table 1, optimization of f_{HB} improved the ranking to 13.0, whereas optimizing f_L only improved the ranking to 13.9. This indicates that the formation of hydrogen bonds at high receptor density values is a key feature that differentiates actives from inactives in ChemScore docked poses.

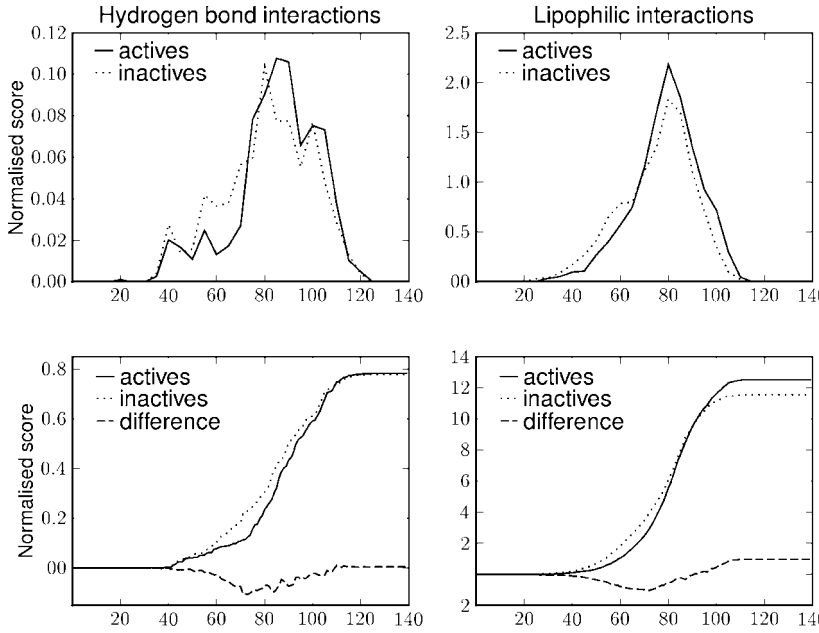

Figure 3. The scores for hydrogen bond and lipophilic interactions in terms of $\rho_{8.0}$ represented as a sum of scores binned by ρ (top) and a cumulative plot (bottom). Each point on the curves in the top diagrams represents the sum of the scores of those interactions in a bin width of 5 ρ units. These sums are normalized separately for actives and inactives by dividing by the total number of active or inactive interactions. The curves in the cumulative plots are also divided by the total number of active or inactive interactions. The difference between the cumulative curve for the actives and the inactives is also shown.

In the optimization of the parameters of f_{HB} and f_L, the final value of S for hydrogen bond interactions is almost five times that for lipophilic interactions. To quantify to what extent the relative scaling of the hydrogen bond and lipophilic interactions contributes to the improvement in active ranks, we repeated the optimization procedure using constant functions, $g(\rho) = S$, instead of the interpolating functions used previously. The final value for the objective function, 14.0, was almost as good as that obtained when f_L was optimized on its own. This indicates that most of this improvement was simply due to the optimization of the ratio of hydrogen bond to lipophilic contributions.

As a result, we decided to replace the lipophilic function f_L by a constant function g_L whose value was optimized in concert with the hydrogen bond function f_{HB}. This four-parameter model gave slightly poorer results than the original six-parameter model (12.5 compared to 11.5), but the reduction in the number of parameters meant that it was less likely to overfit to the training data.

When the four-parameter model was applied to the training set, not all protein actives improved in rank. Of the five proteins whose active ranks worsened the most, two (1p2y and 1jje) involved interactions between the active molecule and a metal atom. This suggested that we should reward deeply buried metal interactions in the same way as hydrogen bond interactions. Since there are far fewer metal interactions in our data set (only 14 of the 85 proteins contain an interaction between the active molecule and a metal), the simplest solution was to scale metal interactions using the same function used for hydrogen bond interactions, f_{HB}. On incorporating metal scaling, the performance for the subset of 14 proteins whose actives interacted with a metal improved from 8.9 to 7.0. In particular, where the rank of the active for 1p2y had previously moved from 13 to 40 on applying RDS, it now only moved from 13 to 18.

This then was the model which we selected to be used for receptor density scaling ('the RDS model'). The form of the final equation is as follows

$$\Delta G_{binding} = \Delta G_0 + \Delta G_{hbond} S'_{hbond} + \Delta G_{lipo} S'_{lipo} + \Delta G_{metal} S'_{metal} + \Delta G_{rot} H_{rot} \quad (5)$$

where the modified S' terms include an additional scaling function f_{HB} or a constant scaling of g_L (compare to eq 3):

$$S_{hbond} = \sum f_{HB}(\rho) s_{hbond} \quad S_{lipo} = \sum g_L(\rho) s_{lipo} \quad S_{metal} = \sum f_{HB}(\rho) s_{metal} \quad (6)$$

The values of the f_{HB} parameters are $\rho_1 = 13$, $\rho_2 = 105$, and $S = 1.05$ (see eq 4). The g_L parameter is $S = 0.52$. Figure 4 shows how the ranks of the actives change when the RDS model is applied to the training set.

Validation of the RDS Model. Given that there are four free parameters, we wished to test whether the improvement in mean rank by 6 places was genuinely due to improved discrimination between active and inactives. As described in the Methods section, we generated false actives by swapping the true active with an inactive of similar score to create Test Set A and repeated the optimization procedure. Since we choose an inactive of similar score, the initial rank of the false actives, 18.8, was almost identical to that of the true actives, 18.6. However, the optimization procedure was only able to improve this ranking slightly to 18.6 (Table 1), indicating that the original optimization was based on true differences between actives and inactives in the training set, rather than overfitting to noise in the data.

To test the robustness of the RDS model to unseen data, we created a test set of new inactives (Test Set B) chosen in the same way as the inactives in the training set. After docking, the poses were rescored with the RDS model and compared to the scores for the active poses from the training

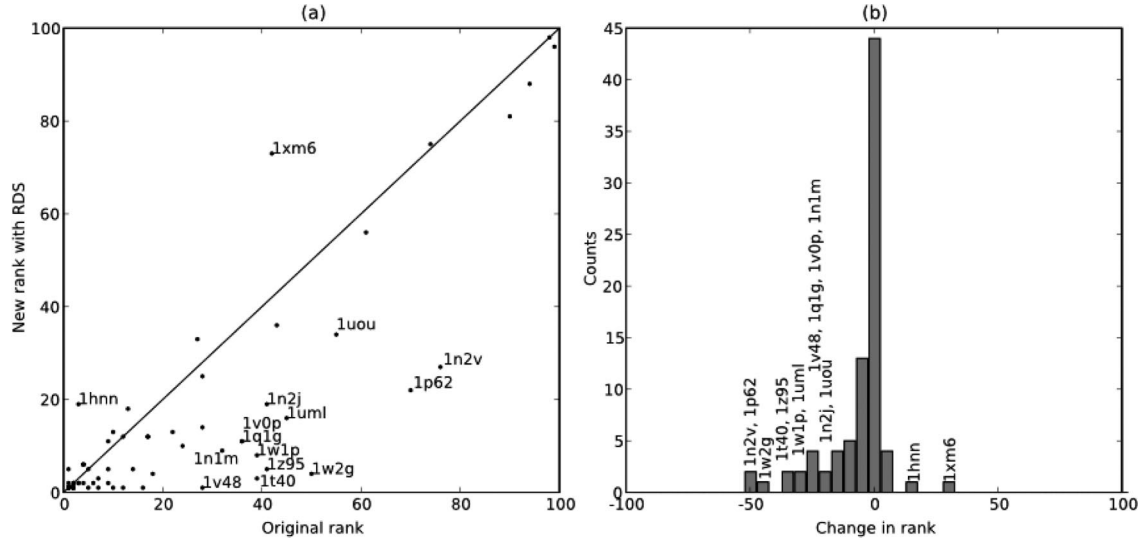

Figure 4. A comparison of the rank of the actives before and after applying the RDS model. In (a) a rank of 1 is the top rank. Protein actives below the $x = y$ line improved in rank when RDS was applied, while those above worsened. (b) is a histogram which focuses on the change of rank (bin size of 5). Negative values correspond to an improvement in rank.

Table 2. Mean Rank of Actives before and after Scaling with the Chosen RDS Model[a]

	mean rank of actives	
data set	before scaling	after scaling
training set	18.6	12.5
Test Set B	18.8	12.6
Test Set C	20.2	11.9

[a] This model uses the receptor density functions f_{HB} and g_L (parameter values given in Table 1), where the function f_{HB} is applied to both hydrogen bond interactions and metal interactions and g_L is applied to lipophilic interactions.

set. The mean rank of the actives was 12.6, which is as good as that obtained for the training set itself (Table 2).

Effect of Molecular Weight on the Performance of the RDS Model. Pan et al.,[26] among others, have pointed out that the greater the number of atoms in a molecule, the more likely it is to score highly in a docking study irrespective of whether it is an active or an inactive. This effect is simply due to the fact that the larger the molecule is, the more hydrogen bond and lipophilic interactions it is likely to make. This can lead to trivial discrimination between actives and inactives where a set of decoys has a different molecular weight distribution than the actives, as described by Verdonk et al.[21] For this reason, our training was performed using sets of inactives with a narrow molecular weight distribution centered around the active.

Since the RDS model alters the relative influence of lipophilic and hydrogen bond interactions, we wanted to investigate whether our ability to identify low molecular weight actives among high molecular weight inactives had worsened. We created a Test Set C consisting of 99 inactives per protein, where the inactives were randomly selected from the pool of all inactives for all proteins in the training set and Test Set B. Figure 5 shows the molecular weight distributions for Test Set B and Test Set C for the two proteins at the extremes of our data set in terms of molecular weights of the actives, 1n2j and 1kzk. The molecular weight distributions of the inactives for Test Set C are much broader than for Test Set B, and, in addition, they are centered around

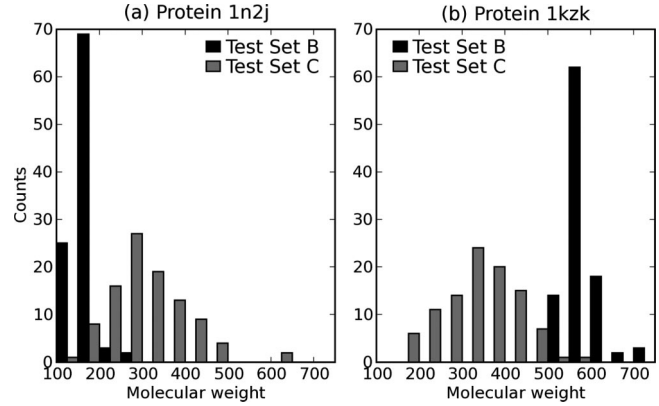

Figure 5. Histograms comparing the molecular weight distributions for the inactives in Test Set B and Test Set C for (a) 1n2j and (b) 1kzk (bin width of 50). The molecular weights of the corresponding actives are 147 and 575, respectively.

the mean molecular weight of all of the inactives rather than around the molecular weight of the active for that protein. Note that although the Test Set C inactives for 1n2 j and 1kzk have been drawn from the same pool, their distributions appear slightly different due to the small sample size.

Based on the difference in molecular weight distributions, we should expect low molecular weight actives such as 1n2j (Figure 5(a)) to perform poorer with Test Set C compared to Set Test B, as the inactives have a higher molecular weight in Test Set C. Similarly, the reverse should be true for high molecular weight actives such as 1kzk (Figure 5(b)). This is indeed the case in general, as shown in Figure 6(a) where a naïve analysis shows a weak linear correlation between the molecular weight of the active and the difference in its rank when using Test Set B versus Test Set C (R^2 is 0.21, $p < 10^{-5}$).

Applying RDS to Test Set C, the average rank of the actives improves from 20.2 to 11.9, a value slightly better than that obtained for Test Set B (see Table 2). When the improvement in rank per active is compared to that obtained with Test Set B, a molecular weight dependence is again observed (Figure 6(b), R^2 is 0.35, $p < 10^{-6}$). Low molecular weight actives show a greater improvement when RDS is

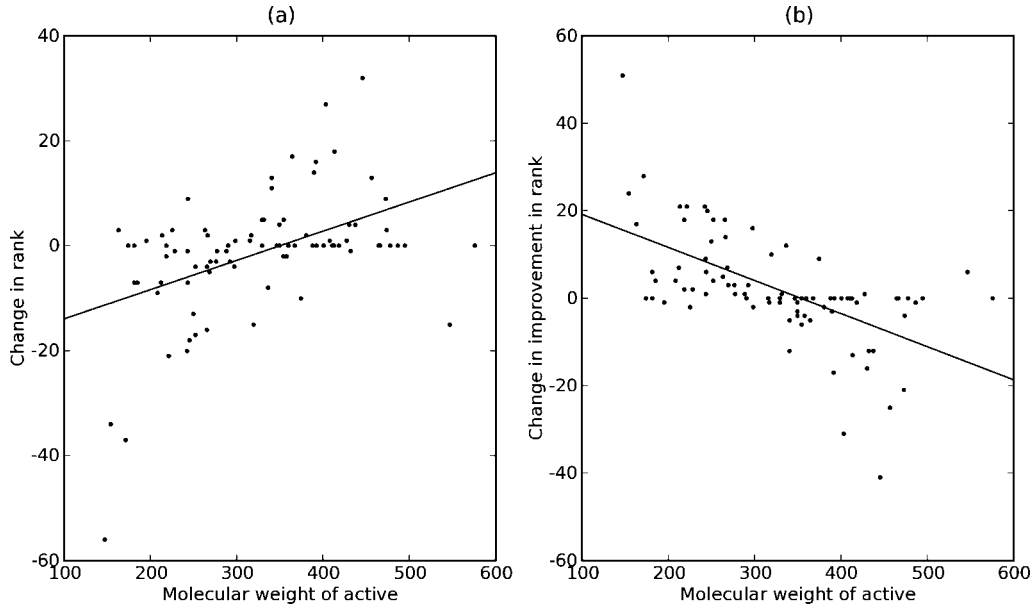

Figure 6. (a) The active ranks versus Test Set B inactives compared to versus Test Set C inactives. A positive value indicates that the rank of the active was higher when using Test Set C. (b) The difference between the improvement obtained when using RDS for Test Set B inactives versus Test Set C inactives. A positive value indicates that a greater improvement was obtained using Test Set C. Both graphs include lines of best fit.

applied to Test Set C, whereas the opposite is true for high molecular weight actives. The correlations in Figure 6(a),(b) are not unrelated; the poorer the initial rank, the greater the potential there is for improvement and vice versa. Despite our original concerns, it appears that low molecular weight actives are actually easier to identify when using RDS.

DISCUSSION

This work originates from an observation that the poses generated by ChemScore for inactive molecules often occur close to the protein surface (that is, rather than deep in the binding site). In order to investigate and exploit this tendency of inactive poses to occur close to the surface, we introduced a measure of buriedness, the receptor density, ρ. This is the number of protein heavy atoms within a certain radius r of the location of a particular interaction. This is easily computed and increases with the distance to the protein surface up to a maximum value where an interaction occurs in the interior of a protein. This measure also implicitly penalizes (or scores as less buried) interactions with flexible walls of an active site; that is, if the side of the active site is not in the protein interior but formed by a peptide loop, for example, the number of protein atoms within the sphere defined by r will be reduced.

As part of our initial studies we also looked at whether other measures of buriedness gave better discrimination between actives and inactives. Coleman et al.[27] and Petøek et al.[28] both measure buriedness as the distance to the convex hull of the protein (the smallest convex polyhedron that encloses the protein). Coleman et al.'s Travel Depth method calculates the shortest distance to the convex hull that does not pass through the molecular surface. We implemented the method used by Petøek et al.'s CAVER software which calculates the shortest distance to the convex hull as measured along the center of the path through the binding site. Despite the increased sophistication of this method, for our purposes it was poorer at discriminating actives from inactives (results not shown).

The initial investigation of the difference between the features of active and inactive poses gave a result contrary to our expectations. We had thought that inactives were scoring highly in some instances due to the formation of lipophilic interactions around the edge of the active site. Although the results depicted in Figure 3 support this idea to some extent, the difference in the distribution of hydrogen bond interactions with respect to ρ is more pronounced. Figure 2 shows that hydrogen bond strengths are independent of ρ; that is, hydrogen bond interactions close to the surface of the cavity score equally as well as those deep in the cavity. In reality, closer to the solvent-exposed surface there is more competition with solvent for the formation of favorable bonds, so that formation of such bonds has a greater energetic penalty compared to formation of a bond deep in the active site. This provides a physical explanation for why RDS works. Lipophilic interactions can be considered as already incorporating a measure of receptor density since the S_{lipo} term (eq 2) consists of a sum of pairwise interactions between protein−ligand atoms that has a distance dependence that includes contributions from atoms up to 7 Å away.

Our optimization procedure improved the average rank of the actives from 18.6 to 12.5 with respect to their inactives. As shown by Figure 4(b), while many proteins improved in rank, only two proteins significantly worsened. Those that improved the most include 1n2v, 1p62, 1w2g, 1t40, 1z95, and 1w1p, whose actives improved in rank by 49, 48, 46, 36, 36, and 31 positions, respectively. These changes in rank are mainly due to the active forming multiple hydrogen bonds at the very bottom of the pocket. For example, the actives of 1n2v, 1w2g, and 1p62 all form three hydrogen bonds at high ρ value (Figures 7, 8, and 9).

The two proteins whose actives significantly worsened in rank are 1xm6 and 1hnn, for which the ranks worsened by 31 and 16 positions, respectively. The removal of water molecules as part of the protocol used by Hartshorn et al.[19] in the preparation of the Astex Diverse Set explains some of these results. In fact, Hartshorn et al. discuss 1xm6 in the

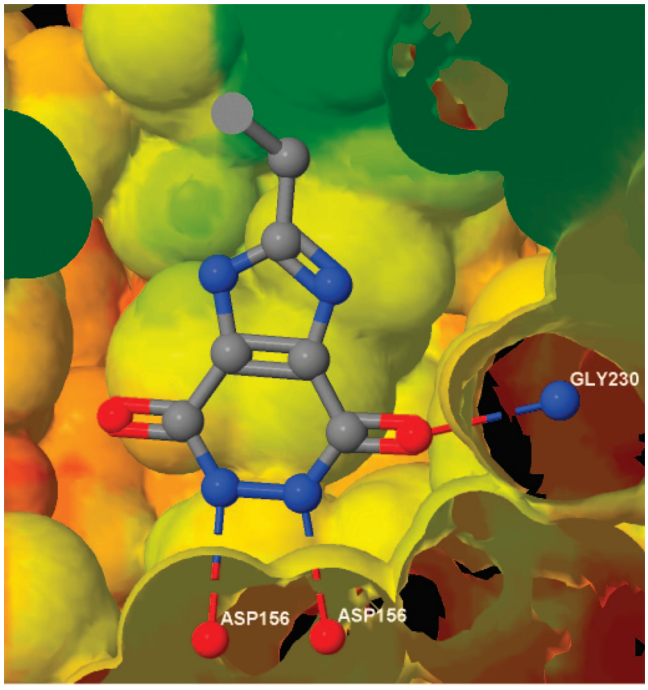

Figure 7. A molecular surface of the active site of 1n2v colored by receptor density value (red is high, blue is low) and sliced by a cutting plane to show the location of the ligand, which forms three hydrogen bonds at high receptor density value. Hydrogens are not shown for clarity.

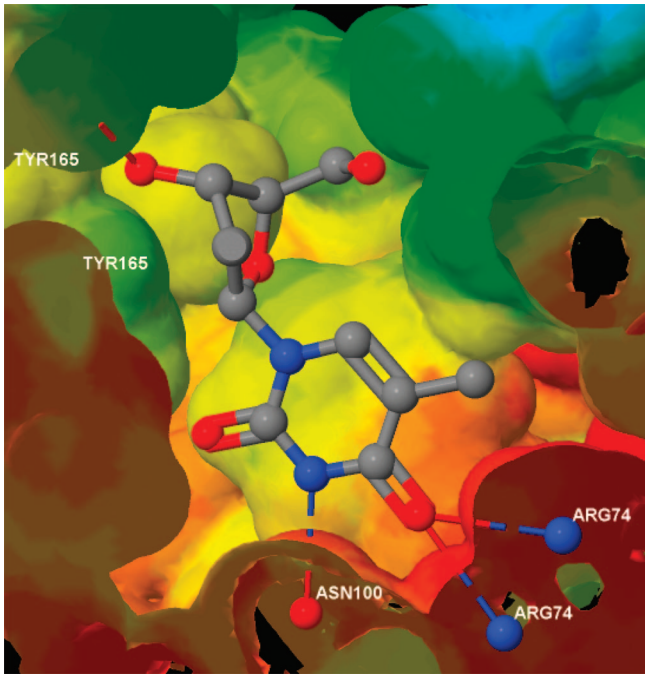

Figure 8. A molecular surface of the active site of 1w2g colored by receptor density value (red is high, blue is low) and sliced by a cutting plane to show the location of the ligand, which forms three hydrogen bonds at high receptor density value. Hydrogens are not shown for clarity.

context of errors in pose prediction. In the crystal structure of 1xm6, a zinc ion is present in the pocket but is not coordinated by the active ligand; instead, two water molecules coordinate to the ion (see Figure 10). As the water molecules are removed in the docking run, the metal ion is free to coordinate to the docked inactives. The score resulting

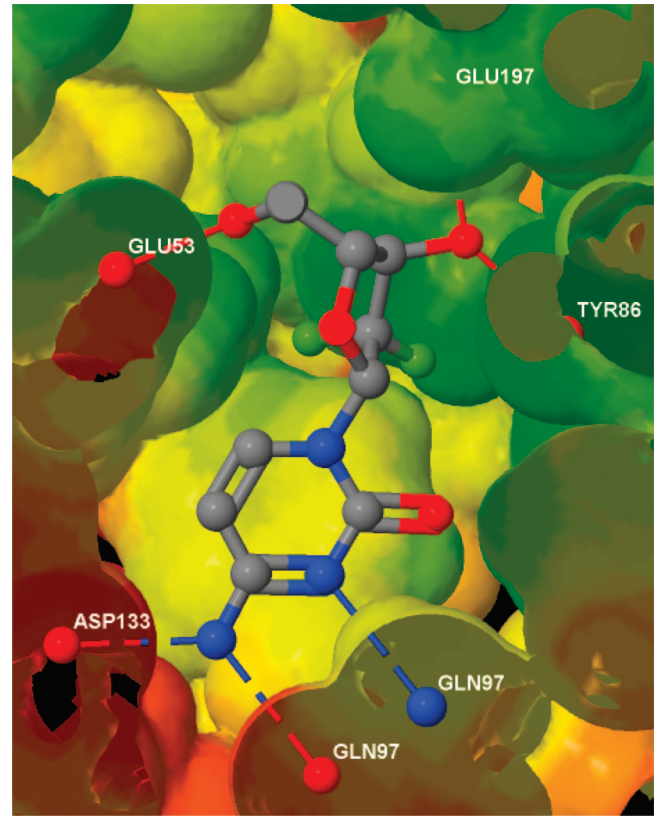

Figure 9. A molecular surface of the active site of 1p62 colored by receptor density value (red is high, blue is low) and sliced by a cutting plane to show the location of the ligand, which forms three hydrogen bonds at high receptor density value. Hydrogens are not shown for clarity.

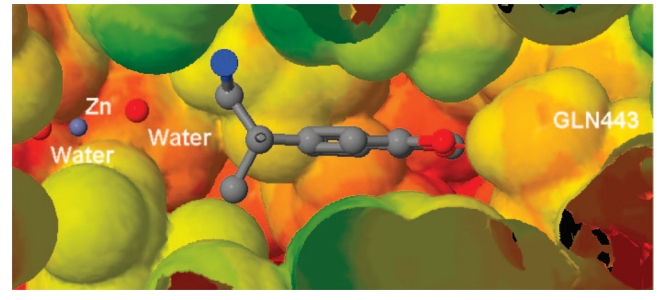

Figure 10. A molecular surface of the active site of 1xm6 colored by receptor density value (red is high, blue is low) and sliced by a cutting plane to show the location of the ligand in relation to a Zn atom in the active site. Shown also are two water atoms which are coordinated to the zinc atom in the crystal structure. Hydrogens are not shown for clarity.

from these interactions is further increased when RDS is applied, thus penalizing the active.

For 1hnn, the binding site is very enclosed, and almost all interactions occur at medium to high receptor density values. The active ligand in the crystal structure forms a hydrogen bond from a charged secondary amine to a water molecule that is also bound to two acidic protein residues and which is highly conserved among related crystal structures. The absence of this water molecule during scoring causes the active to score poorly. In addition, some inactives score highly as they form multiple hydrogen bonds in a pocket with high ρ value which, in the crystal structure, is filled by a hydrogen-bonded network of two water molecules, one of which is highly conserved among related crystal structures (see Figure 11).

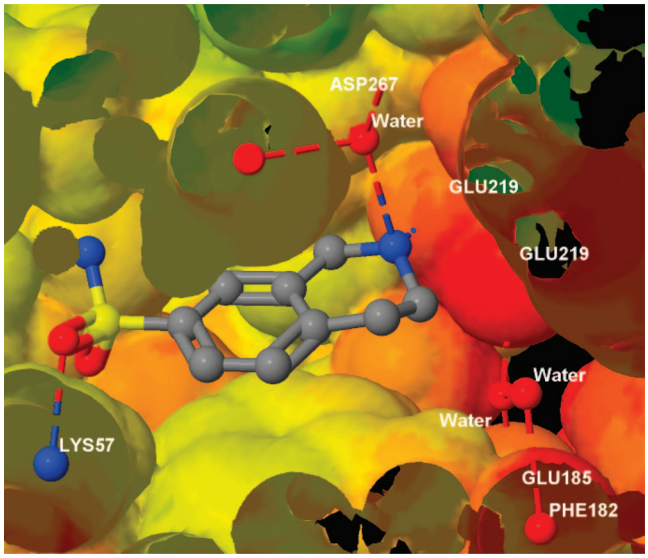

Figure 11. A molecular surface of the active site of 1hnn colored by receptor density value (red is high, blue is low) and sliced by a cutting place to show the location of the ligand in relation to a pocket of high receptor density value occupied by two crystalline waters. A third water is shown located above an amino group in a ring of the ligand. Hydrogens are not shown for clarity.

For a small number of proteins, the rank of the active was very poor whether or not RDS was applied: 1gm8 remained unchanged at 98; 1sq5 moved from 99 to 96; 2br1 from 94 to 88; 1t9b from 94 to 88; 1xoq from 74 to 75. Several of these involve hydrophobic active sites. The active poses for 1gm8, 1t9b, and 1xoq involve π-π stacking which does not receive any explicit score in ChemScore. The active in 2br1 has an intramolecular hydrogen bond which also is unscored.

Using the mean of the rank of the actives as our objective function has some consequences. Since the best and worst ranks are bounded by one and the number of inactives, improvement of the scores of actives that perform poorly has much more of an influence than improvement of those that already do quite well ("the prodigal son effect"). This however is in keeping with our objectives, to develop an improved scoring function with good performance across a wide range of protein targets.

Although we have shown that improved discrimination between actives and inactives is possible by rescoring with RDS, it does not follow that this RDS model can be trivially used during the docking process to yield similar improved results. Preliminary results show that applying RDS during docking causes inactive molecules to be driven down to the bottom of binding sites where their interactions are scored highly despite the presence of unfavorable clashes. We intend to carry out further work to improve docking results.

CONCLUSION

We have identified a consistent difference between active and inactive poses based on the receptor density value of the location of hydrogen bond and lipophilic interactions. By exploiting this difference using receptor density scaling, we were able to substantially improve the average ranking of actives across a training set of 85 proteins from 18.6 to 12.5.

Optimization of the performance of a scoring function using negative data is a practical method of improving scoring functions that directly addresses their deficiencies. Since we know that inactive molecules have features that allow the protein to distinguish them from actives, we can take advantage of this by incorporating negative data into the training of scoring functions.

Rescoring using receptor density scaling is available as an option in GOLD 4.0.

Abbreviations: RDS, receptor density scaling.

ACKNOWLEDGMENT

N.M.O.B. and R.T. thank a number of people for their critical and lively discussions of this work: Jason Cole, John Liebeschuetz, Chris Murray, and Marcel Verdonk. In particular they thank Martin Harrison for the original idea and definition of receptor density scaling.

REFERENCES AND NOTES

(1) Jain, A. N. Scoring functions for protein-ligand docking. *Curr. Prot. Pept. Sci.* **2006**, *7*, 407–420.
(2) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
(3) Schulz-Gasch, T.; Stahl, M. Scoring functions for protein-ligand interactions: a critical perspective. *Drug. Discovery Today Technol.* **2004**, *1*, 231–239.
(4) Foloppe, N.; Hubbard, R. Towards predictive ligand design with free-energy based computational methods. *Curr. Med. Chem.* **2006**, *13*, 3583–3608.
(5) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **2006**, *49*, 5851–5855.
(6) Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A stepwise approach for defining the applicability domain of SAR and QSAR models. *J. Chem. Inf. Model.* **2005**, *45*, 839–849.
(7) Graves, A. P.; Brenk, R.; Shoichet, B. K. Decoys for docking. *J. Med. Chem.* **2005**, *48*, 3714–3728.
(8) Smith, R.; Hubbard, R. E.; Gschwend, D. A.; Leach, A. R.; Good, A. C. Analysis and optimization of structure-based virtual screening protocols (3). New methods and old problems in scoring function design. *J. Mol. Graphics. Modell.* **2003**, *22*, 41–53.
(9) Pham, T. A.; Jain, A. N. Parameter estimation for scoring protein-ligand interactions using negative training data. *J. Med. Chem.* **2006**, *49*, 5856–5868.
(10) 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.
(11) Huang, S.-Y.; Zou, X. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. *J. Comput. Chem.* **2006**, *27*, 1866–1875.
(12) Stahl, M. Modification of the scoring function in FlexX for virtual screening applications. *Perspect. Drug Discovery Des.* **2000**, *20*, 83–98.
(13) 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.
(14) 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.
(15) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible docking using tabu search and an empirical estimate of binding affinity. *Proteins* **1998**, *33*, 367–382.
(16) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43–53.
(17) 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.
(18) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins* **2003**, *52*, 609–623.

(19) 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.
(20) *Jmol*: an open-source Java viewer for chemical structures in 3D, v11.3.53 (prerelease). http://www.jmol.org/ (accessed April 14, 2008).
(21) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using protein-ligand docking: Avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793–806.
(22) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000–1008.
(23) Jones, E.; Oliphant, T.; Peterson, P. SciPy: Open source scientific tools for Python, v0.5.2. http://www.scipy.org (accessed April 14, 2008).
(24) Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and its variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357.
(25) *Open Babel* - a molecular informatics toolkit, v2.1.1. sf.net http://openbabel (accessed April 14, 2008).
(26) Pan, Y.; Huang, N.; Cho, S.; MacKerell, A. D., Jr. Consideration of molecular weight during compound selection in virtual target-based database screening. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 267–272.
(27) Coleman, R. G.; Sharp, K. A. Travel Depth, a new shape descriptor for macromolecules: Application to ligand binding. *J. Mol. Biol.* **2006**, *362*, 441–458.
(28) Petøek, M.; Otyepka, M.; Banáš, P.; Košinová, P.; Koèa, J.; Damborský, J. CAVER: a new tool to explore routes from protein clefts, pockets and cavities. *BMC Bioinformatics* **2006**, *7*, 316.

CI8000452