Docking and Scoring Articles

Parameter Estimation for Scoring Protein−Ligand Interactions Using Negative Training Data

Tuan A. Pham and Ajay N. Jain*

Cancer Research Institute, Department of Biopharmaceutical Sciences, and Department of Laboratory Medicine, University of California, San Francisco, 2340 Sutter Street, San Francisco, California 94143-0128

Received January 14, 2005

Surflex-Dock employs an empirically derived scoring function to rank putative protein−ligand interactions by flexible docking of small molecules to proteins of known structure. The scoring function employed by Surflex was developed purely on the basis of *positive* data, comprising noncovalent protein−ligand complexes with known binding affinities. Consequently, scoring function terms for improper interactions received little weight in parameter estimation, and an ad hoc scheme for avoiding protein−ligand interpenetration was adopted. We present a generalized method for incorporating synthetically generated *negative* training data, which allows for rigorous estimation of all scoring function parameters. Geometric docking accuracy remained excellent under the new parametrization. In addition, a test of screening utility covering a diverse set of 29 proteins and corresponding ligand sets showed improved performance. Maximal enrichment of true ligands over nonligands exceeded 20-fold in over 80% of cases, with enrichment of greater than 100-fold in over 50% of cases.

Introduction

Discovery of novel lead compounds through virtual screening of chemical databases against protein structures is well established,[1] but there is still much room for improvement in key aspects of algorithm performance. Many methods have been published that vary primarily two components: scoring functions[2−8] and search methods[9−15] (for a more complete review, see Bissantz et al.[16] and Jain[17]).

The primary criteria for evaluating docking strategies are geometric docking accuracy, screening utility, scoring accuracy, and speed. Geometric docking accuracy measures a docker's ability to generate and recognize the native conformation and alignment (*pose*) of a ligand bound to its cognate protein beginning from an arbitrary initial pose. This is typically reported as the fraction of cases where the docker's top-scoring ligand pose is within 2.0 Å rmsd from the experimentally determined binding geometry. Screening utility measures a docker's ability to rank cognate ligands of a protein above random ligands, as is desired in typical virtual screening applications. Methods for quantifying screening utility varies, but most frequently it is characterized by constructing virtual screening libraries that contain some small number of known active molecules for a protein under study along with a large number of randomly selected compounds typical of a screening library. Following docking of a such a virtual library to a protein, the resulting ranking of the ligands is used to compute the observed true positive rates (percentage of known ligands found) at various false positive rates (essentially the percentage of the database that must be experimentally assayed, assuming a low rate of "hits" in a screening library). Alternatively, screening enrichment is reported, which is the ratio of the proportion of true hits found to the expected proportion based on the composition of the library, computed for a fixed small percentage of the top-ranked ligands of the screening library or reported as a maximum value over all possible percentages of the ranked library. Scoring accuracy is the degree to which a docker's quantitative scoring of ligand binding matches experimentally determined values. This can be very important in focused medicinal chemistry exercises, but the thrust of this paper is on large-scale virtual screening, so the methodological evaluation focused most strongly on screening utility. One important recent trend in the docking literature has been the use of publicly available benchmarks for assessing the performance of methods. Rognan's group has been at the forefront of this trend,[16,18] and others have made use of the benchmarks developed there, both for docking accuracy and for screening utility. In particular, reports on Surflex[19] and GLIDE[13,20] have made direct use of those benchmarks. In addition, reports of the performance of GOLD have been very important in establishing benchmarks of docking accuracy.[11,21]

The issue of docking accuracy has been extensively tested by many groups, and the data sets are sufficiently large that the reports of different groups largely agree as to performance of the most widely used methods. The broadest recent study directly compared eight methods: DOCK, FlexX, FRED, Glide, GOLD, SLIDE, Surflex, and QXP.[18] The four most successful methods achieved very similar results, ranging from 50% to 55% success in returning top-ranked poses within 2.0 Å rmsd of the experimental results: FlexX, GLIDE, GOLD, and Surflex. Recent methods-focused reports on GOLD, Surflex, and GLIDE contained benchmarking information on docking accuracy as well, and these results largely agreed with the independent work of Rognan's group,[19−21] suggesting comparable accuracy among these methods. Additional details of these benchmark results can be found in a recent review.[17]

With respect to screening utility, the situation is more complex. First, there is a very limited set of publicly available benchmarks (e.g., the two cases from Rognan's group[18]). Recent work by Perola et al.[22] made use of proprietary data, and other

* To whom correspondence should be addressed. Voice: (415) 502-7242. Fax: (650) 240-1781. E-mail: ajain@jainlab.org.

10.1021/jm050040j CCC: $33.50
Published on Web 06/29/2005

recent reports have focused on single proteins[23,24] or small sets of different proteins.[13] Second, the performance of methods is significantly more variable than for docking accuracy. While it appears that the methods that perform best in terms of docking accuracy generally outperform other methods with respect to screening utility,[18] there is still a multifold difference in screening enrichment on the common benchmarks among the different methods,[13,17−19] The critical difference that distinguishes the successful methods listed above with respect to docking accuracy from other methods is the use of empirically derived scoring functions. Both GOLD and Glide make use of modified versions of the ChemScore scoring function.[4] FlexX makes use of a function based on Bohm's work,[3] and Surflex makes use of the Hammerhead[9] scoring function.[2]

In the development of these scoring functions, only *positive* data were used, encompassing protein−ligand complexes with known binding affinities. One consequence of this choice is that repulsive terms, which include effects such as improper steric clashes, same charge atomic interactions, and desolvation penalties, can receive little weight. This is because the training ligands generally fit well within protein active sites, do not typically make same charge close contacts, and do not bury hydrophobic ligand surfaces against hydrophilic protein surfaces (or vice versa). In this paper, we introduce the idea of employing *negative* data in training scoring functions for molecular docking. By this we mean making use of putative nonbinding ligands that do not fit a protein active site, make inappropriate polar interactions, or violate aspects of protein−ligand complementarity that should result in desolvation penalties. There are three obvious constraints that can be used in the context of estimating parameters for a scoring function: (1) that the computed scores for ligands of known geometry correspond closely to the known affinities of the ligands, (2) that the computed scores for the highest-scoring poses of nonligands be poor relative to some value, and (3) that the computed scores for geometrically incorrect dockings of ligands be poorer than the score for poses that are very close to correct. The first constraint is the typical use of positive training data, recently made more robust by the availability of large numbers of such complexes from databases such as PDBbind.[25] The second constraint carries with it two questions. What should be the source of nonligands, and what should be the value of the bound on score? The third constraint offers a direct method for tuning scoring functions using protein−ligand complexes where the binding affinity is *not known*. Where a particular scoring function incorrectly identifies a ligand pose as scoring better than the correct pose, a dynamic constraint can be computed to penalize the incorrectly scored pose.

In this paper, we combined the first two types of constraints to re-estimate parameters for the Surflex-Dock scoring function. For the positive data, we employed the same 34 complexes used originally in the parameter estimation (we did not make use of larger, newer data sets in order to avoid complications with testing the method). For the negative data, we screened a random compound library against each of the proteins in the positive data set and retained ligands that scored better than a predetermined value but that, on the basis of molecular similarity, did not look at all like the native ligands of the proteins. In employing the negative data, we imposed a penalty on the objective function for parameter optimization if negative ligands exceeded a fixed score.

We developed a large set of screening test cases, totaling 29 sets of proteins with associated true positives, which cover a very diverse set of protein active sites and corresponding ligand properties. The newly formulated scoring function obviated the need for ad hoc treatment of improper clashes, and screening enrichment remained the same or improved in $^{21}/_{29}$ cases. Maximal enrichment of true ligands over nonligands exceeded 20-fold in over 80% of cases, with enrichment of greater than 100-fold in over 50% of cases. In the six cases of poorest performance by the new scoring function, use of multiple protein conformations exhibited promise in improving screening enrichment. We also established that docking accuracy was essentially unchanged with the new scoring function using a set of 81 protein−ligand complexes.

By simply adding automatically generated negative data to the training of the Surflex-Dock scoring function, we were able to estimate parameters that previously received so little weight that ad hoc terms were required to make use of the function in docking. The new scoring function yielded excellent performance over a wide variety of test cases, both in terms of docking accuracy and in terms of screening utility, without requiring knowledge-based postprocessing of docking scores to incorporate interpenetration values. Further generalization of this approach to estimate additional parameters (e.g., involving desolvation effects), with the inclusion of more positive *and* negative data, should yield more complex scoring functions for molecular docking with substantially improved performance.

Surflex-Dock is available free of charge to academic researchers for noncommercial use (see http://www.jainlab.org/downloads.html for details on obtaining the software). The data sets used for benchmarking in this paper are freely available to all researchers via the same Web site.

Methods

The focus of the paper is on improving the treatment of the repulsive terms of the Surflex-Dock scoring function. We will briefly review the scoring function, since additional details are presented elsewhere.[2] We employed multiple sources of data to construct test cases for screening enrichment, and we employed our previous benchmark of 81 complexes to assess docking accuracy. The following reviews the scoring function, the data sets and preparation, the optimization procedure for retuning the scoring function, and the procedures for assessment of performance.

Scoring Function. The Surflex-Dock scoring function (originally used within Hammerhead[9]) was tuned to predict the binding affinities of 34 protein−ligand complexes, with its output being represented in units of $-\log(K_d)$.[2] The range of ligand potencies in the training set ranged from 10^{-3} to 10^{-14} and represented a broad variety of functional classes. The parametrization of the function models the noncovalent interactions of organic ligands with proteins, including proteins with bound metal ions in their active sites. The function is continuous and piecewise differentiable with respect to ligand pose, which is important for the gradient-based optimization procedures employed within Surflex-Dock. The terms, in rough order of significance, are hydrophobic complementarity, polar complementarity, entropic, and solvation (negligible). The full scoring function is the sum of each of these terms.

The dominant terms are the hydrophobic contact term and a polar contact term that has a directional component and is scaled by formal charges on the protein and ligand atoms. These functional terms are parametrized on the basis of distances between van der Waals surfaces, with negative values indicating interpenetration. Each atom on the protein and ligand is labeled as being nonpolar (e.g., the H of a C−H) or polar (e.g., the H of an N−H or the O of a C═O), and polar atoms are also

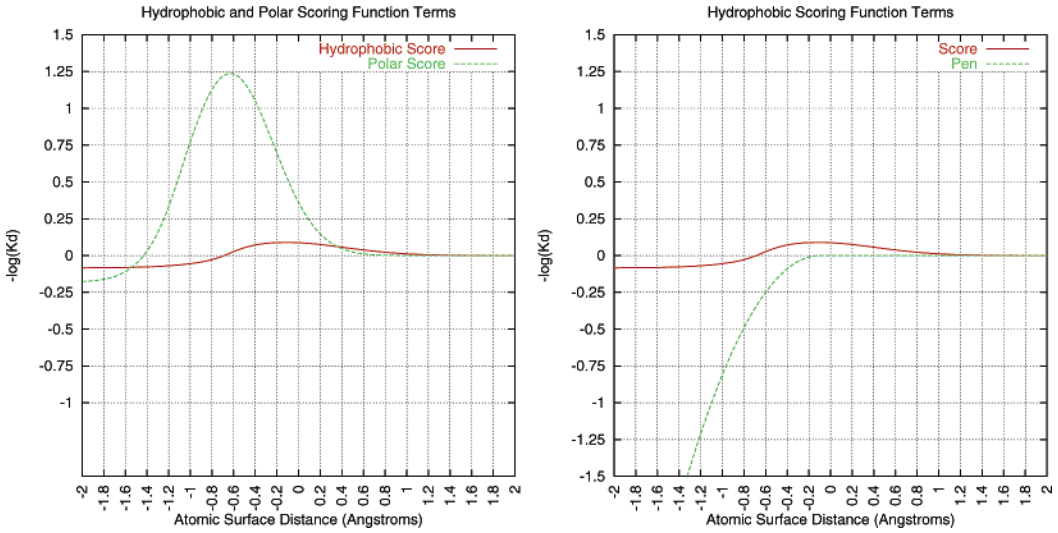

Figure 1. Left: hydrophobic and polar terms of the scoring function. The hydrophobic term peaks at approximately 0.1 units with a slight surface interpenetration. The polar term for an ideal hydrogen bond peaks at 1.25 units. Right: hydrophobic scoring function term plotted with the ah hoc penetration term. The learned penalties for improper interpenetration were insignificant because of the use of only positive training data. Docking with this scoring function is carried out by optimization of the score, with the added constraint of the penetration term during gradient-based refinement of ligand scores.

assigned a formal charge, if present. Figure 1 shows plots of the hydrophobic term and the polar term for a hydrogen bond. The hydrophobic term (bottom curve in red) yields approximately 0.1 unit per ideal hydrophobic atom−atom contact. A perfect hydrogen bond yields about 1.2 units and has a peak corresponding to 1.97 Å from the center of a donor proton to the center of an acceptor oxygen (learned entirely on the basis of the empirical data and corresponding quite closely to the expected value range). Despite the large difference in the value of a single hydrophobic contact versus a single polar contact, the hydrophobic term accounts for a *larger* total proportion of ligand binding energy on average. This is because there are many more hydrophobic contacts than ideal polar contacts in a typical protein−ligand interaction.

Apart from the hydrophobic and polar terms, the remaining important terms include the entropic term and the solvation term. The entropic term includes a penalty that is linear in the number of rotatable bonds in the ligand, intended to model the entropic cost of fixation of these bonds, and a term that is linearly related to the log of the molecular weight of the ligand, intended to model the loss of translation and rotational entropy of the ligand. The solvation terms are linearly related to a count of the number of missed opportunities for appropriate polar contacts at the protein−ligand interface.

However, neither the solvation term nor any of the terms intended to guard against improper clashes received much weight in the original training (the solvation term was, in fact, 0.0). This was due to the fact that no negative data were employed; only ligands with their cognate proteins were used in parameter estimation. Thus, there were essentially no data from which to induce such penalty terms. In particular, the linear weights on the terms for improper steric clashes, noncomplementary polar contacts, and solvation effects were, respectively, −0.08 (l_1 in the original paper), −0.15 (l_5), and 0.0 (l_6). All of these were very small relative to, for example, the magnitude of a single ideal hydrogen bond (1.23). To make use of the scoring function for molecular docking, it was necessary to superimpose a term to prevent atomic overlap between the protein and the ligand (and within the ligand itself): $-10.0(d_{ij} + \delta)(d_{ij} + \delta)$. In this term, d_{ij} is the distance between atomic surfaces (negative for surfaces that interpenetrate) and δ was 0.1 for all contacts except those between complementary polar atoms, where δ was 0.7. In the reimplementation of the Hammerhead scoring function for Surflex-Dock, this term was normalized to a value called "pen" by multiplying by 4.0 and dividing by the number of atoms in the ligand. A docked ligand yielded two values: score and pen. The user was required to choose a cutoff for pen beyond which a ligand was rejected (or alternatively construct a combination score by weighting the two terms). The formulation of the term was unsatisfying because the parameters were chosen in a largely arbitrary fashion and the requirement for selecting a threshold for interpenetration made for an extra methodological complexity.

In this work, we sought to address this term in a systematic fashion by making use of negative training data. However, this required that the new penetration term be treated in an absolute sense, both with respect to protein−ligand interactions and with respect to ligand self-interpenetration. The latter required a change in the internal computation of self-clashing, eliminating atom pairs from consideration if they were connected by nonrotatable bonds. Without this modification, ligands with, for example, bicyclic ring systems were at a disadvantage relative to other ligands because of the inherent nominal clashing among atoms within constrained covalent systems. Earlier versions of Surflex-Dock used an heuristic method to estimate the best possible self-penetration for each ligand and normalized the self-penetration by subtracting this value, but this estimate was not sufficient for systematic parameter tuning. Also, to obtain the most reliable final scores for docked ligands, the final gradient-based ligand pose optimization was enhanced in thoroughness to ensure convergence of the scoring function. Incomplete convergence would effectively add noise to the scores of ligands during scoring function optimization as well as during the evaluation of the methodology.

Software Versions. Surflex, version 1.24, was in widest circulation prior to the current work, and was used in data set generation and for certain control experiments. This version implemented the original Hammerhead scoring function, as described above, with the ad hoc interpenetration treatment. Versions up to 1.28 continued to use this formulation. The modified scoring function, with the new treatment of penetration values, more aggressive gradient-based pose optimization, and

a switch to select the original scoring function, begins with Surflex-Dock versions 1.31 and higher.

Negative Data Sets. Two sources were used for nominal negative ligands. The screening data set from the comparative paper of Bissantz et al.[16] was used, as in our previous report.[19] The original data set included 990 randomly chosen nonreactive organic molecules chosen from the Available Chemicals Directory (ACD) ranging from 0 to 41 rotatable bonds. The data set was used with two modifications. First, all ligands were subjected to an automatic protonation procedure and energy minimization in order to eliminate differential bias between positive and negative ligands (positives were treated the same; see below). Second, ligands with greater than 15 rotatable bonds were eliminated, resulting in 861 negative ligands. This eliminated decoys that were clearly not druglike and better reflected the composition of the positive ligands.

The second source was ZINC (see http://blaster.docking.org/zinc). We randomly selected 1000 compounds from the druglike subset (1 847 466 total) of the 07-26-2004 version of the database. These compounds had molecular weights of $\leq$500, with computed $\log_p$ of $\leq$5, h-bond donors of $\leq$5, and h-bond acceptors of $\leq$10. The compounds were processed identically to the ligands above, and the number of rotatable bonds in the set ranged from 0 to 12. In the remainder of the paper, "negative ligand set" refers to the 861-compound set derived from Bissantz et al.[16] unless otherwise noted specifically as the "ZINC negative set".

Training Data Set. Re-estimation of the scoring function parameters relating to improper interactions required *both* positive and negative data; otherwise, the scoring function could be trivially modified to include very large penalty terms. To simplify evaluation of the new function, we employed the 34-complex training set that was used in constructing the original Hammerhead scoring function.[2] All of the complexes dated from 1992 or earlier, reducing the possibility that the training set could contain information relevant to our tests, which were based largely on more recent data. The original complexes were converted from PDB to Sybyl mol2 format and protonated per expectation at physiological pH, with active site rotamers of hydroxyls and thiols and tautomers of imidazoles optimized for cognate ligand interactions.

For the negative data, we employed the negative ligand set above (restricted, for computational efficiency, to the 600 least flexible molecules). For each of the protein structures of the 34-complex positive data set, we docked all negative ligands using Surflex-Dock, version 1.24, using default parameters. Ligands that scored greater than 4.0 (in units of pK_d, ignoring penetration values) were presumed to be false positives. The value 4.0 was chosen because in our experience a K_d or K_i of 100 μM is at the limit of what is generally considered to be a specific ligand of a protein. We believed that it was quite likely that ligands with nominal scores greater than 4.0 pK_d from our small random library were false hits, based on the expectation that true hit proportions in a typical library are roughly $^1/_{1000}$ to $^1/_{10000}$. Note, however, that several of the positive ligands have pK_d less than 4.0, so there are clearly examples of weak ligands that specifically bind proteins with quite weak affinities. To reduce the likelihood of including a true ligand as a negative in the training set, we further screened the ligands based on molecular similarity to the bound pose of the native ligand for each protein using Surflex-Sim.[26] Those ligands that scored worse than 0.5 on a scale from 0 to 1 were retained as negatives for the purpose of parameter estimation. Table 1 lists the PDB codes, true ligands, and number of negative ligands for each protein in the training data set. The total number of negative ligands was 2274, with 34 positive ligand examples.

Table 1. Training Data Set

complex	ligand	*N* negative	pK_d
7cpa	ZFV^p(O)F	78	14
1stp	biotin	46	13.4
6cpa	A-ZAA^p(O)F	115	11.52
4tmn	ZF^{p}LA	72	10.19
4dfr	methotrexate	115	9.7
4phv	L700,417	207	9.15
1dwd	NAPAP	92	8.52
5tmn	ZG^{p}LL	100	8.04
2gbp	galactose	1	7.6
1etr	MQPA	66	7.4
1tlp	phosphoramidon	87	7.33
1tmn	CLT	78	7.3
1rbp	retinol	115	6.72
1ppc	NAPAP	27	6.46
5tln	HONH-BAGN	96	6.37
1pph	3-TAPAP	34	6.22
1ett	TAPAP	121	6.19
1phf	4-Phe-imidazole	141	6.07
4dfr*	2,4-diaminopteridine		6
5cpp	adamantone	3	5.88
2xis	xylose	0	5.82
2ifb	C_{15}COOH	173	5.43
1ulb	guanine	67	5.3
2ypi	phoshoglycilic acid	45	4.82
3ptb	benzamidine	28	4.74
2phh	*p*-hydroxybenzoate	58	4.68
2tmn	PLN	101	4.67
3ptb*	phenylguanidine		4.14
1dwd*	amidinopiperidine		3.82
4tln	Leu-NHOH	98	3.72
3ptb*	benzylamine		3.42
4cha	indole	0	3.1
1dwb	benzamidine	110	2.92
3ptb[a]	butylamine		2.82

[a] Indicates that the respective ligand was docked in or generated by a direct modification of the native ligand in the complex.

Test Data Sets. Four sources were used to generate 29 test cases for screening utility (see Figure 2). The two data sets from the comparative paper of Bissantz et al.[16] were used, as in our previous report.[19] The original data sets included protein structures for HSV-1 thymidine kinase (TK, PDB code 1KIM) and estrogen receptor α (ERα, PDB code 3ERT): 10 known ligands of TK in arbitrary initial poses and 10 known ligands of ERα in arbitrary initial poses. The data sets were used with the modification (as above) that ligands were subjected to an automatic protonation procedure and energy minimization in order to eliminate differential bias between positive and negative ligands.

One limitation of the foregoing two cases is that the ligands either are drugs or are druglike in their potency and physicochemical properties. Therefore, they are likely to form "easy" cases for docking tools. To address this issue, we took molecular structures from two papers that reported the results of combinations of both virtual screening and high-throughput screening to form two new cases where the true positives were reflective of the type of hits that can be found in library-based screening. The protein PARP (poly(ADP-ribose) polymerase), PDB code 2PAX, along with 15 true ligands formed one case, based on Perkins et al.[27] The protein PTP1b (protein tyrosine phosphatase 1b), PDB code 1PTY, along with 11 true ligands formed the second, based on Doman et al.[28]

We used the PDBbind database[25] to generate a large number of additional cases for testing screening utility. From the full 800-complex set, we identified all proteins that were represented

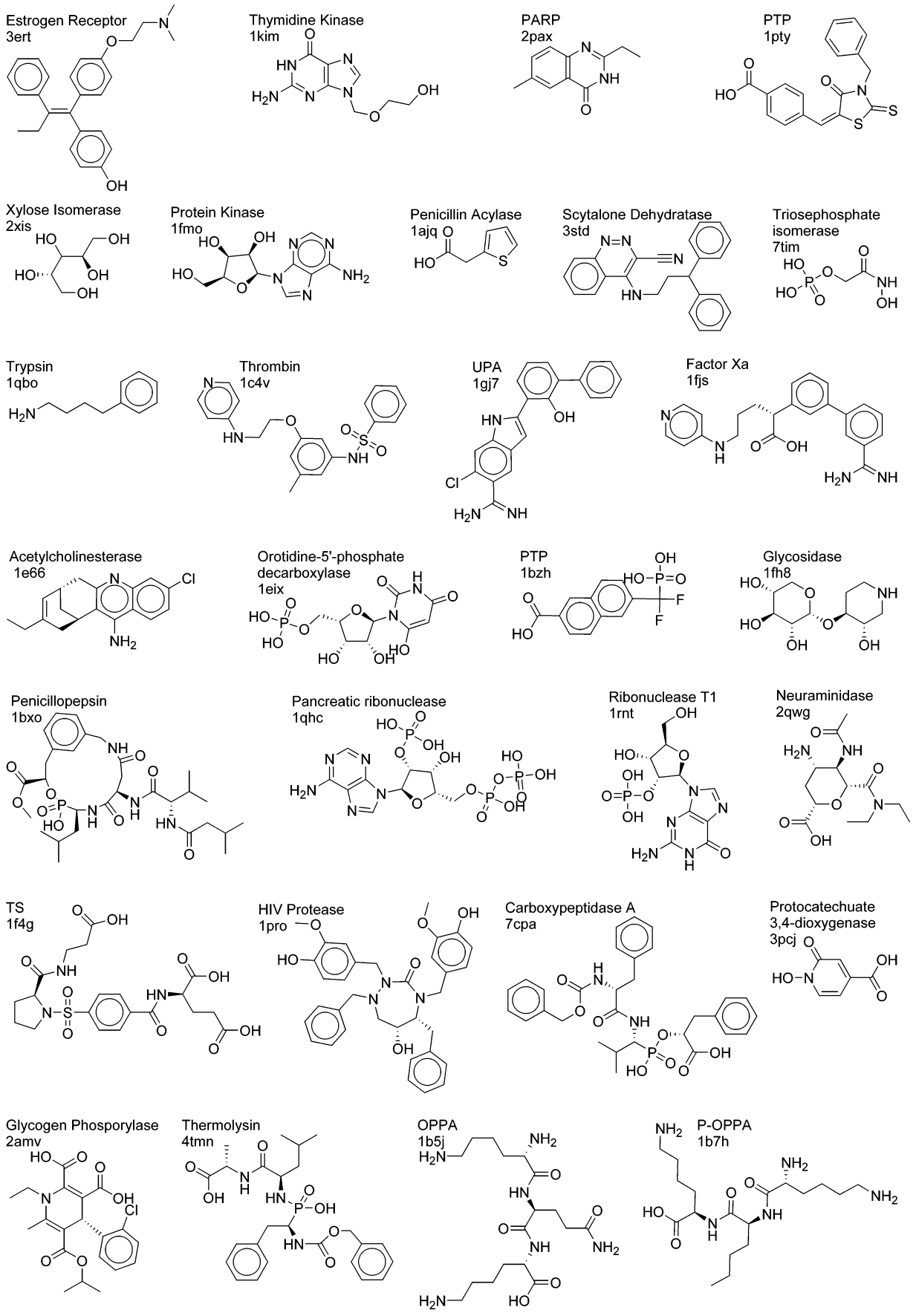

Figure 2. Example structures for the 29 screening enrichment test cases. The first row contains the ER and TK test cases from previous work as well as two new cases consisting of true ligands that were found through combinations of virtual and high-throughput screening. The remaining 25 cases come from the PDBbind database.[25]

by at least five different ligands. For each of these proteins, we arbitrarily selected one of the PDB structures to serve as the screening target, and we generated a Sybyl mol2 format protein (prepared as above for the positive training data). The cognate ligands of the proteins were subjected to the same automatic protonation and minimization above. In cases where more than 20 ligands existed for a protein, we selected the 20 most diverse, based on molecular similarity, in an procedure analogous to the IcePick method.[29] The overall procedure yielded 25 proteins, with a total of 226 ligands. We believe that this represents the largest set of screening test cases currently available. As shown in Figure 2, the functional diversity of proteins and the structural diversity of ligands were large. The set included four serine proteases (row 3 of the figure), kinases, phosphatases, isomerases, aspartyl proteases, metalloproteases, and a number of other protein types. Importantly, the range of ligand binding affinities was large, with a substantial number of lower affinity ligands. Half of the ligands had pK_d less than 6.0 (micromolar or worse K_i or K_d), with just one-fifth having pK_d greater than 9.0 (subnanomolar or better).

Optimization Procedure. To demonstrate the feasibility of the approach of using negative data, we chose to optimize two parameters: the weight of the term for noncomplementary (same charge) polar contacts and the weight of the term for protein−ligand and ligand−ligand clashes. The former term was parametrized exactly as in the original scoring function, and it will be referred to in what follows as sf_pr (*Surflex polar repulsion*). Owing to the relative success of our ad hoc approach to modeling interpenetration, employing the "pen" value, we chose to add a new term to the scoring function by including an analogous quadratic penalty. However, rather than scaling the term by the number of ligand atoms, which has no theoretical basis, we added the following new term to the original scoring function: sf_hrd$(d_{ij} + \delta)(d_{ij} + \delta)$, with variables defined as above. This formulation can be thought of in terms of another additive energy term. The parameters sf_hrd (*Surflex hard* penetration) and sf_pr, when optimized, would be expected to be both significant and negative.

In the search for an optimum parametrization, given some objective function, there is a complexity that is somewhat unique to docking and shared with 3D QSAR. Because the function being optimized changes, the optimal poses for the ligands within the training set change as well. As in our previous work,[2,30−32] we addressed this problem by interleaving parameter optimization with ligand pose optimization. In this approach, each time a ligand is optimized, the resulting pose is added to the *pose cache* for that ligand. In the inner loop of evaluation of ligand scores for computing the overall objective function, all cached ligand poses are evaluated, with the highest scoring one defining the score for the ligand. For this work, it was sufficient to retain only the highest scoring pose on each iteration (essentially a pose cache size of 1).

Our objective function was a straightforward generalization of the common mean squared error function. For positive ligands, their contribution was the square of the difference between their maximal score under pose optimization and their experimentally determined score (in units of pK_d). For negative ligands, if their score was greater than 4.0, their contribution to the error function was the same as for positive ligands (squared difference), but if their score was 4.0 or less, their contribution was zero. So any deviation from the correct score for a positive ligand induced a corrective pressure during optimization, but only in the case of inappropriately high scores would negative ligands contribute to the error function. Note that while the cutoff for producing pressure was 4.0 (which exceeds the scores of some of the known positives), the procedure allowed for many of the scores of the synthetic negatives to become much less than 4.0. The last complexity was that there were about 2 orders of magnitude more negative examples than positive examples. So a simple optimization of the total error would have vastly overweighted the contribution of the negative ligands. We balanced the relative contributions of the negative ligands and positive ligands to be equal in order to avoid this.

Since there were only two parameters to optimize, we used a simple approach that combined broad sampling with bounded random search with a fine-grained line search and small random parameter perturbation. In principle, since the error function and the scoring function are continuous and differentiable, more complex approaches could have been employed, but they were not necessary. A single stable solution that minimized error was reached with sf_pr $= -2.52$ and sf_hrd $= -0.945$.

Computational Assessment. The old and new scoring functions of Surflex-Dock were evaluated for screening utility using our large set of 29 cases and for docking accuracy using our previous 81-complex data set.[19] We used Surflex-Dock, version 1.24, to generate protomols in all cases, using standard parameters. In computing scores for comparison between the old version (which returns values of both score and pen) and the new version (which returns only a score), we avoided the use of arbitrary thresholds by simply adding the score and pen values of the old scoring function to yield a single scalar combination score. This approximates the newer functional treatment and provides an apples-to-apples comparison.

To differentiate effects of the new scoring function from the treatment of ligand self-penetration and ligand pose optimization, we conducted two separate comparisons. To test the effects of the new scoring function, we compared performance of Surflex-Dock, version 1.31, with and without specifying the −old_score parameter, which selects the old scoring function (but does not change any other behavior). These effects are the primary focus of the paper and are reported in detail in the Results and Discussion.

We conducted a separate comparison between the older version of Surflex-Dock (version 1.24) and the current version using the old scoring function (version 1.31 −old_score) in order to assess the effects of the changes in the ligand self-penetration computation and ligand pose optimization. In making comparisons of different methods from the perspective of screening utility, we employed receiver-operating-characteristic (ROC) plots. Figure 3 shows ROC plots for the ER and TK test cases using the old approach and new approach. For a given ranking of the ligands using a particular docking procedure, we computed the true positive and false positive rates at *every* possible score threshold, with the resulting pairs of values yielding the ROC plot (true positive rates on the *y* axis and false positive rates on the *x* axis). The ideal ROC curve goes from (0,0) to (0,1) to (1,1), reflecting a 100% true positive rate at a false positive rate of 0%. This occurs when the scores of all cognate ligands are larger than the scores of all random ligands. ROC curves resulting from different scoring functions can be quantitatively compared by computing the area under the curve, with the perfect ROC curve yielding an area of 1. While the ER case showed no real difference (ROC areas of 0.993 and 0.992), the TK case showed better enrichment for the older Surflex-Dock version.

The right-hand plot of Figure 3 illustrates the reason. While the scoring function is identical between the two versions, the distribution of negative ligand scores using the more aggressive

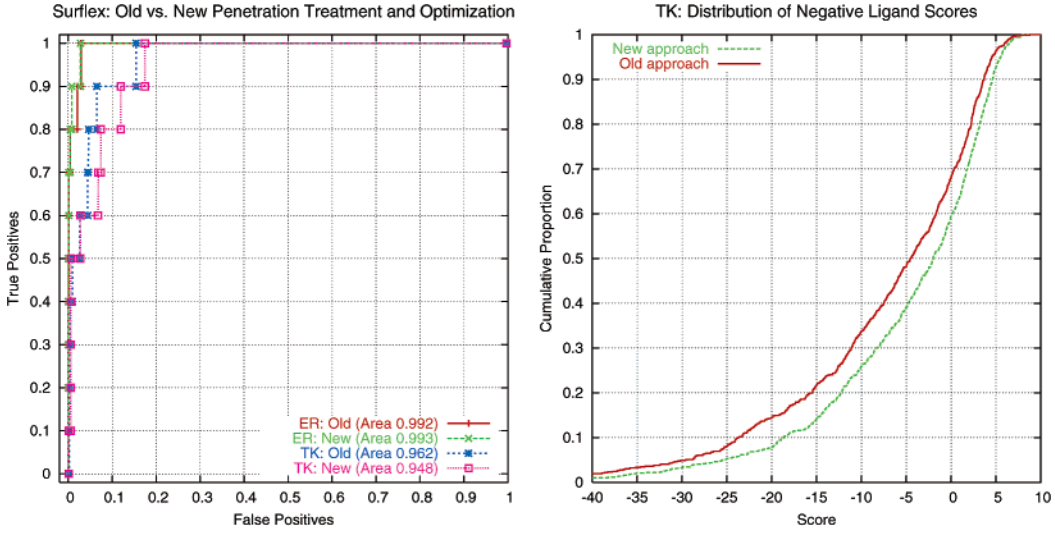

Figure 3. Left: ROC curves for the thymidine kinase and estrogen receptor screening examples using the original (Surflex, version 1.24) and new (Surflex, version 1.31) treatment of self-penetration and final ligand optimization. Right: cumulative histograms of the combined scores (score + penetration) of nonligands in the TK case. The differences are small, but the rightward shift of the scores in the new optimization procedure results in a decrease in separation between the true and false ligands for TK.

pose optimization procedure of the new version is shifted to the right of the old version. Recall that Surflex yields scored in units of pK_d, so more aggressive optimization results in *higher* scores. This reduces the separation between the positive and negative ligands (the positive ligand distribution does not change significantly). Despite poorer performance in this case, if we considered the ROC areas of the old and new versions in all 29 screening examples, we observed significantly improved performance using the new version ($p < 0.05$ by *t*-test). This is expected, given that frequently nonconvergent pose optimization would simply add a degree of noise to ligand scores. In what follows, the only difference between versions is use of the −old_score switch within version 1.31 of the Surflex-Dock software.

Results and Discussion

We focused our attention on the results attributable to the differences between the old and new scoring functions, which lay in the effects of inappropriate atomic interpenetration and noncomplementary polar contacts. Figure 4 shows the cumulative histograms of the scores for the negative ligands used in training the new scoring function. The original scores (ignoring the penetration term) were the rightmost curve, with nearly all ligands scoring greater than 4.0 (not all ligands scored greater than 4.0 because of minor changes in protein preparation between negative data set generation and final evaluation). The scores corresponding to the new function are represented by the leftmost curve. Note that following parameter optimization, approximately 70% of the negative ligands scored *less* than 4.0. It was not possible, using just the two parameters that were optimized, to simultaneously eliminate all 100% of the nominal false positives while retaining accurate scores for the positive examples. The middle curve is the cumulative histogram of the *sum* of the score and pen values for the old scoring function. While this curve was closer to that of the new function, the penalties that were learned through systematic optimization in the presence of negative training data yielded uniformly lower scores.

Figure 5 (left) shows a plot of the old and new hydrophobic terms, which reflects the negative contribution of the quadratic interpenetration penalty with its linear weight of −0.945 (sf_hrd). While this term is much more stringent than the

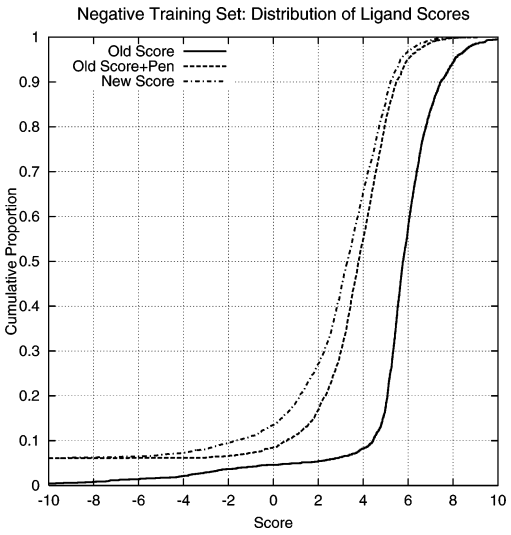

Figure 4. Cumulative histograms of ligand scores before and after scoring function modification. The rightmost curve depicts the scores of nominal false positives for all protein structures used in training. The middle curve includes the penetration values for those ligands. The left curve shows the scores of the ligands using the modified scoring function.

sigmoidal component of the original function (with a maximal penalty of 0.08 log units), it is less stiff than a standard 6-12 potential. Figure 5 (right) shows the fit to the 34 positive ligand scores, with a mean error of 1.1 log units, which is quite comparable to the original report of the scoring function as parametrized solely on positive data. So without significantly affecting the scores of known ligands, we were able to make an impact on the scores of the synthetically generated negative ligands.

Assessment of New Scoring Function in Screening Enrichment. The ER and TK cases, which have been the subject of numerous reports,[13,16,18,19] deserve special attention. Figure 6 shows the full ROC curves and underlying cumulative histograms of positive and negative ligand scores for the TK and ER test cases. Performance was excellent in both cases with both scoring functions, with all ROC areas exceeding 0.9. Maximal enrichment (ratio of true ligands found to expected number of hits at all percentages of database screening) occurred

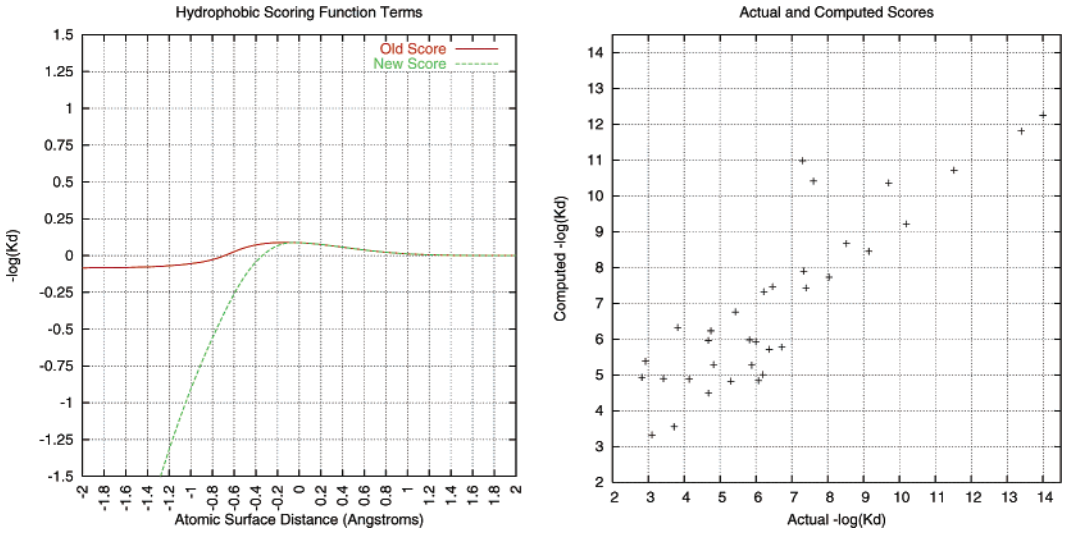

Figure 5. Left: original and modified hydrophobic scoring function term. Right: plot of computed and actual pK_d for the 34 positive training cases after parameter modification. The new interpenetration penalty is much more severe than had been learned from solely positive data, but the effect on the computed scores compared was small.

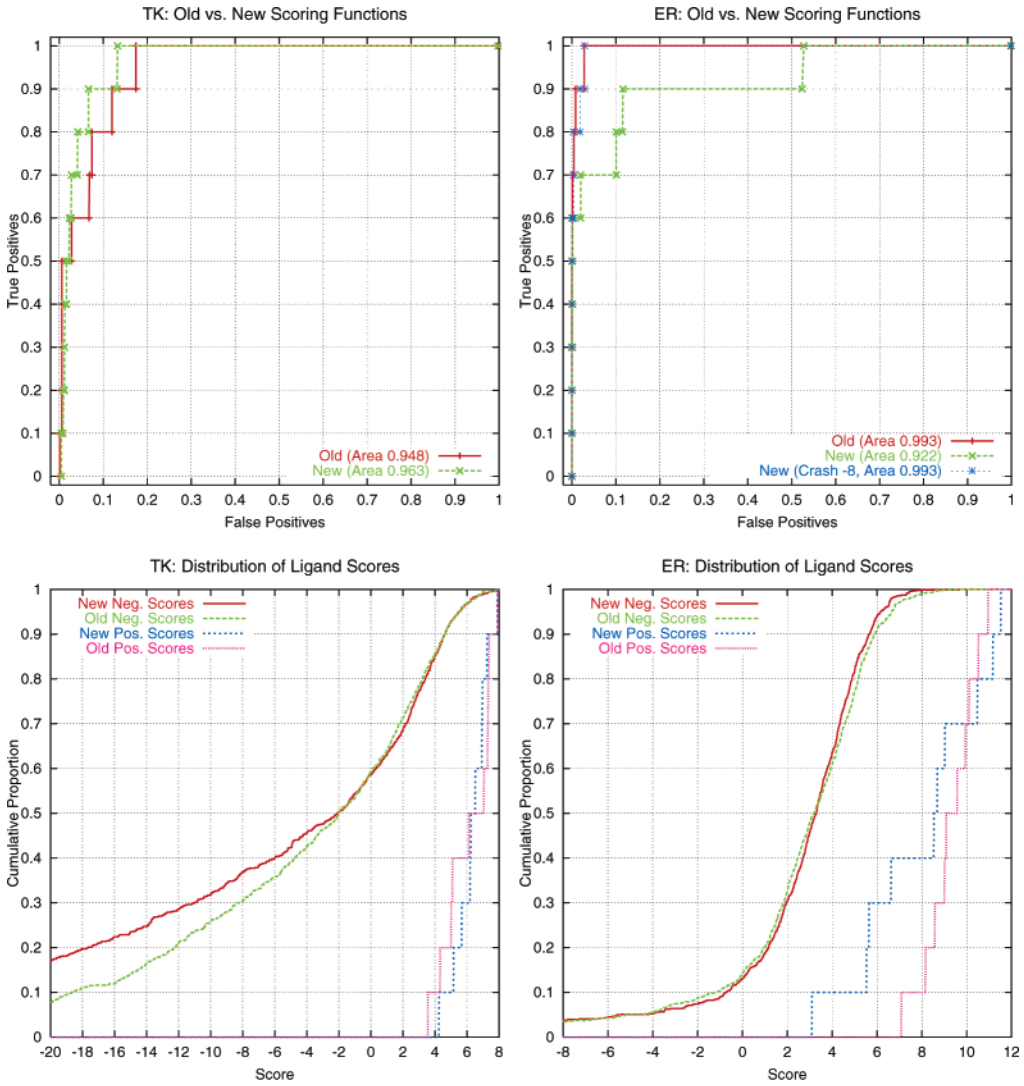

Figure 6. Top plots: ROC curves for the TK and ER cases with old and new scoring functions. In the TK case, the new scoring function performs better in terms of ROC area. In the ER case, performance at the original level was attainable by ignoring interpenetrations up to a value of −8.0. Bottom plots: cumulative histograms of positive and negative ligand scores for TK and ER. While the change to the scoring function involved increases in weight to negative terms, the changes in scores that gave rise to changes in positive/negative separation were varied. In the TK case, a rightward shift of positive ligand scores was responsible. In the ER case, a leftward shift in positive ligand scores hurt performance.

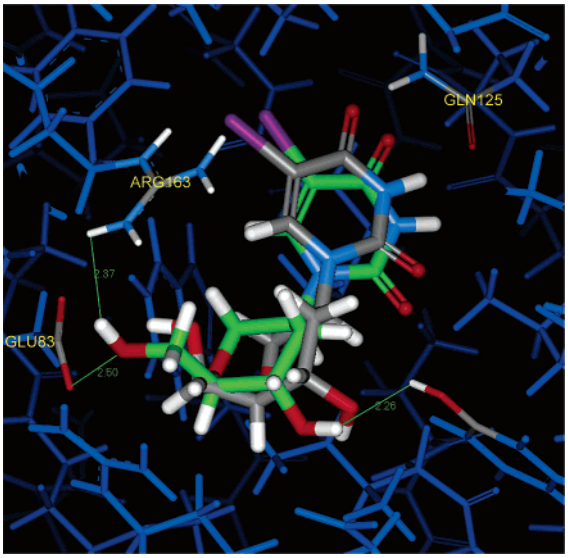

Figure 7. Thymidine kinase ligand 5-iodouracil anhydrohexitol nucleoside (AHIU) docked using the old scoring function (green carbons) and new scoring function (gray carbons). The old pose scored with the new scoring function scores more than 6 log units worse with the new scoring function because of multiple same-charge atomic interactions (indicated by green lines). The new scoring function guided the docking to a pose with a better score and better geometric relationship to the protein, in particular to GLN125, though both poses are accurate by rmsd (<1.5 Å).

at very low false positive rates for both scoring functions and exceeded 20-fold for TK and 500-fold for ER. Since ROC areas are much more stable to small changes in the scores of true and false ligands than maximal enrichment values, we will focus the quantitative comparisons between the scoring functions on ROC areas in what follows.

In the TK case, the new function leads to a reduction in the false positive rates at true positive rates of 70% and higher. In the ER case, the opposite is true. The bottom plots in Figure 6 show the cumulative histograms of positive and negative ligand scores for both cases. Surprisingly, while there were differences in the distribution of negative scores in both cases between the different scoring functions, the differences that drove the discrepancies in ROC curves were the result of changes in the *positive* ligand scores. In the ER case, some of the large positive ligands were penalized by the new scoring function's harsher treatment of interpenetration. While this is not desirable, it is an expected effect in some cases. In cases such as this, where very large ligands are desired, it is possible, as before, to treat the interpenetration portion of the score heuristically. By allowing all docked ligands a degree of penetration with no penalty, we observed performance equivalent to that of the old version (blue curve in Figure 6).

In the TK case, we saw an unexpected effect. The lowest scoring of the true positives scored *higher* using the new scoring function than with the old. This is surprising because the new function has *harsher* penalties for inappropriate contacts. However, since the scoring function of Surflex-Dock is used deep in the search process, we observed different solutions to the docking problem using the different functions. Figure 7 shows an example of this effect using the dockings observed for a single positive ligand of thymidine kinase. The solution using the new scoring function is depicted in atom color, and the solution using the old scoring function is depicted using green carbons. In the case of the old scoring function, very little weight was given to noncomplementary polar contacts, and in the pose shown, there were three very close contacts between pairs of atoms with charges of the same sign. With the new scoring function, this pose scores more than 6.0 log units worse than with the old scoring function, owing to the large negative weight given to the sf_pr term. The pose returned by the new scoring function avoids all of the improper contacts while retaining as many appropriate contacts. While both poses were very close to the experimentally determined pose (<1.5 Å rmsd), the pose returned by employing the new scoring function was clearly superior.

Table 2. Comparative True Positive Rates in Screening for Thymidine Kinase[a]

FP, %	DOCK	FlexX	Fred	Glide	GOLD	QXP	Slide	Surflex	Surflex, new
2.5	0.0	20.0	0.0	20.0	10.0	0.0	0.0	40.0	60.0
5.0	10.0	40.0	0.0	50.0	40.0	20.0	0.0	80.0	80.0

[a] Values are percentage of true positive rates for fixed false positive rates. All data but the last column are taken from Kellenberger et al.[18]

Table 3. Enrichment Factors Compared with Other Docking Programs (Large Values Are Good)[a]

	DOCK	FlexX	Glide	GOLD	Surflex-New
thymidine kinase	3.0	11.1	19.3	8.2	37.9
estrogen receptor	6.7	8.9	47.1	28.5	90.7

[a] Values are EF′(70), as defined in Halgren et al.[13] All data but the last column are taken from Halgren et al.[13]

Since these two cases have been used in a number of other studies, it is possible to make direct comparisons among different methods. Table 2 shows the true positive rates reported by Kellenberger et al.[18] for thymidine kinase for eight docking methods, amended to include the Surflex result with the new scoring function. As evidenced by the plots in Figure 6, the Surflex results did not change much, with a slight improvement at the 2.5% level of false positives. Note, however, that the new results did *not* require the choice of a threshold penetration value, which was required in the previous studies. Table 3 shows enrichment factors reported by Halgren et al.[13] amended to include the Surflex result with the new scoring function, again without any special treatment of protein interpenetration. In both the TK and ER cases, the Surflex enrichment factors were substantially better than the other methods.

While these results are encouraging, they represent a limited test, given just 2 proteins and 20 positive ligands, all of which are either drugs or druglike in potency and physicochemical properties. Figure 8 shows ROC curves and score histograms for PARP and PTP. In these cases, the positive ligands were discovered through combinations of virtual and high-throughput screening. They were all of relatively poor potency and reflect the makeup of common screening libraries. In both cases, the new scoring function improved performance, though it did so on the basis of different effects. In the case of PARP, the slight leftward shift in the distribution of negative ligand scores (lower-left plot, upper part of red curve) was responsible for the difference observed in the ROC curves. In the case of PTP, as with TK above, the effect was a substantial right shift of positive ligand scores. Again, it appears that the new scoring function guided the docking search algorithm more effectively to better solutions.

Table 4 summarizes Surflex-Dock screening performance on all 29 cases tested (ligand examples shown in Figure 2) for the old and new scoring functions, using ROC areas to characterize the separation of positive and negative ligand sets. Differences of less than 0.005 are considered negligible. The 29 cases included a diverse set of 226 ligands, with a large number having poor binding affinities (half with micromolar or worse K_d or

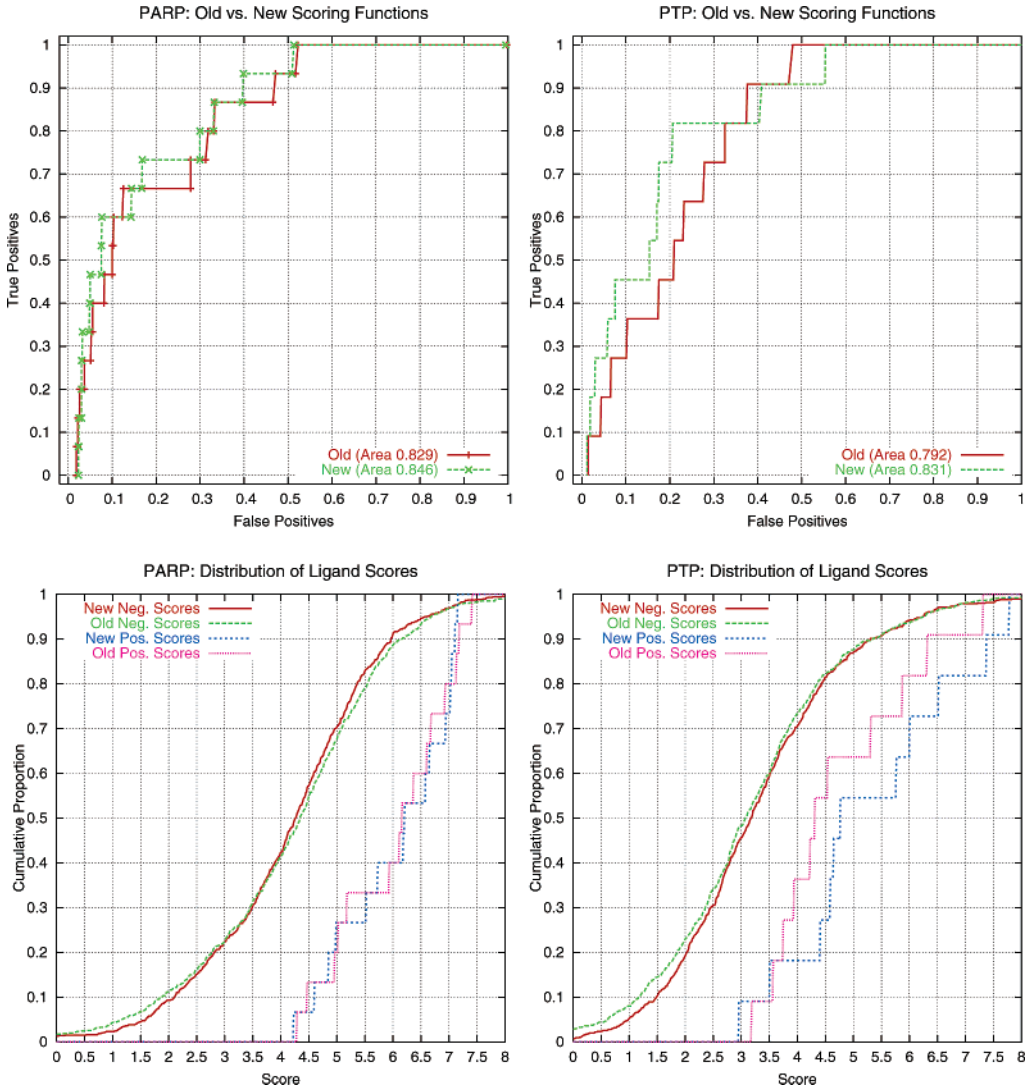

Figure 8. Top plots: ROC curves for the PARP and PTP test cases with old and new scoring functions. Bottom plots: cumulative histograms of positive and negative ligand scores for PARP and PTP.

K_i). Overall, the new scoring function performed as well as or better than the old scoring function in $^{21}/_{29}$ cases, so it is clearly not worse than the old function ($p = 0.01$ by exact binomial). The converse is not true. That is, the old scoring function performs as well as or better than the new one in $^{15}/_{29}$ cases, which allows the possibility that the old scoring function is worse. However, the number of test cases is too small to make a strong statement that the new approach is significantly better in terms of the proportion of cases where the ROC area is clearly improved. With the new scoring function, maximal enrichment of true ligands over nonligands exceeded 20-fold in over 80% of cases, with enrichment of greater than 100-fold in over 50% of cases. Given that many of these cases were clearly much more difficult, based on ligand affinities, than the widely used TK and ER cases, performance on par with those two examples in the *majority* of cases suggests that Surflex-Dock should yield strong performance in terms of screening utility in a wide variety of cases.

We further tested the performance of the system using an entirely new negative screening set of ligands, derived from the ZINC database. This was done because it was theoretically possible that the scoring function optimization procedure could have "learned" something specific about properties of the negative set derived from the Rognan benchmarks, which was used to produce the putative negative examples for the training set. However, the ROC areas derived using the ZINC negative set with the new scoring function were statistically indistinguishable from those presented above. In fact, the ROC area differences between the scoring functions using the ZINC versus Rognan negative sets were almost perfectly correlated, with a Pearson r^2 of 0.987.

Effects of Protein Conformation. In the six cases with poorest performance of the new scoring function, the old scoring function performed better only in two (1bxo and 1qhc), reflecting the possibility that these six cases may be difficult proteins in some intrinsic sense. Another possibility is that the particular conformations of the protein structures used for screening were not propitious. We arbitrarily selected different protein structures for each of these six cases and retested the performance of Surflex-Dock's new scoring function. Table 5 shows the results from the original structures, from the new structures, and from combining both screens by taking the maximal ligand score from both structures in each case. In $^{3}/_{6}$ cases, the new structures yielded much improved performance, suggesting that these cases may have been outliers. These included the two cases in which the old scoring function had outperformed the new one. In the remaining cases, there was no significant change for one (1f4g) but reduced performance for two (1fmo and 2amv). Clearly, protein conformation can have unpredictable effects. However, it appears that the simple

Table 4. Comparison of ROC Area Scores for 29 Cases with New and Old Scoring Functions

name	Nmols	new score	old score	difference
ER	10	0.922	0.993	−0.071
TK	10	0.963	0.948	0.016
PARP	15	0.846	0.829	0.017
PTP	11	0.831	0.792	0.039
2XIS	5	0.958	0.923	0.035
1FMO	8	0.764	0.722	0.041
1AJQ	6	0.922	0.897	0.025
3STD	5	0.844	0.814	0.030
7TIM	6	0.966	0.935	0.031
1QBO	20	0.990	0.978	0.011
1C4V	20	0.876	0.900	−0.023
1GJ7	12	0.953	0.948	0.005
1FJS	6	0.980	0.974	0.007
1E66	6	0.764	0.767	−0.003
1EIX	5	0.996	0.995	0.002
1BZH	12	0.917	0.916	0.002
1FH8	6	0.997	0.995	0.002
1BXO	5	0.746	0.985	−0.239
1QHC	6	0.791	0.886	−0.095
1RNT	5	0.952	0.966	−0.015
2QWG	7	0.965	0.987	−0.022
1F4G	10	0.693	0.594	0.098
1PRO	20	0.862	0.955	−0.093
7CPA	8	0.901	0.916	−0.015
3PCJ	8	0.948	0.952	−0.003
2AMV	5	0.709	0.693	0.017
4TMN	13	0.828	0.810	0.018
1B5J	16	1.000	1.000	0.000
1B7H	6	0.999	1.000	−0.001

approach of using multiple structures and reporting the maximum score of each ligand might be an appropriate safe strategy. In $^6/_6$ cases, this approach performed better than the worst of the single-protein runs ($p = 0.02$ by exact binomial), and in $^3/_6$ cases this approach performed better than either single structure alone. These results stand in some contrast to the interesting, but counterintuitive, result reported by Wei et al.[33] where they observed *worse* performance using this approach unless they corrected for cavitation energies in different protein structures.

Solvation Effects. The protein for which Surflex-Dock yielded the poorest performance was glycogen phosphorylase (2amv), and performance was not improved using multiple conformations. The active site of this protein is quite hydrophilic, with 20 protein atoms capable of making polar interactions with a ligand in the binding pocket. Figure 9 shows a positive ligand (green) and a nonligand (atom color) in the active site of the protein. The known ligand makes a network of polar interactions with three guanidine moieties. However, the nonligand makes no successful polar interactions in the binding site at all, while still scoring 5.0 pK_d. Further, the ligand effectively buries multiple polar atoms of the protein, rendering them inaccessible even to solvent. The scoring function, even in its new form, does not account for the cost of desolvating the protein (important in this case) or the ligand.

We approximated this intention in this case by requiring that the ligands of glycogen phosphorylase received a polar score

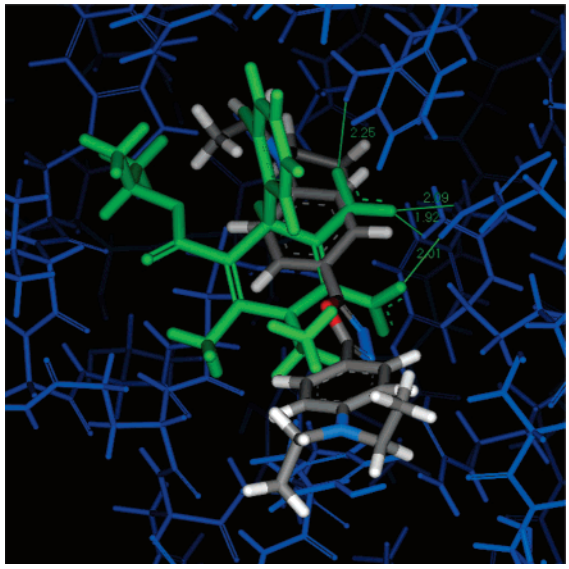

Figure 9. Native ligand (green) and docked false positive (atom color) in 2amv (blue). Three guanidines on the protein make multiple favorable polar contacts with the true ligand. The false positive makes no complementary polar contacts and buries the guanidine moieties with an aromatic ring.

of at least 3.0 in either of the two structures. If so, we recorded the maximum score of the ligand, else we recorded a zero. Employing this simple heuristic, we saw an improvement from 0.684 to 0.862 in ROC area. While this is an ad hoc procedure, it motivates the development of an effective strategy for modeling desolvation costs. Such a strategy should include some computation of the degree of buriedness of each polar atom (i.e., the degree to which it is inaccessible to solvent in the docked state), the solvated polar score of the protein and ligand, and the degree to which complementary polar contacts between the protein and ligand ameliorate loss of interactions with solvent molecules. A sufficiently refined treatment will require several parameters and will benefit from additional positive training data in addition to the negative data that have been the subject of this paper.

Docking Accuracy and Speed. We evaluated docking accuracy on the same data set used previously, consisting of 81 protein−ligand complexes.[19] Docking accuracy was not significantly different between the old and new scoring functions, with $^{58}/_{81}$ complexes (72%) in both cases having rmsd of top-ranked poses within 2.0 Å of crystallographic observation using either scoring function.

Docking speed was not significantly affected by any of the changes from the previous reported version to the modified algorithm reported here. On standard workstation hardware (Intel Xeon 2.80 GHz, 1 GB RAM, Windows XP Professional, Surflex-Dock version 1.31 with default options), the mean docking time over the 81 complexes was 17 s, with ligand flexibility ranging from 0 to 15 rotatable bonds. Docking time was roughly linear in the number of rotatable bonds, with a

Table 5. Effects of Protein Conformation on Screening Enrichment

protein		original structure		new structure		combination
name	*N*	PDB	ROC area	PDB	ROC area	ROC area
penicillopepsin	5	1bxo	0.746	1apw	0.941	0.946
pancreatic ribonuclease	6	1qhc	0.791	1afk	0.915	0.957
acetylcholinesterase	6	1e66	0.764	1gpn	0.914	0.847
thymidylate synthase	10	1f4g	0.693	1tsl	0.700	0.707
cAMP dep protein kinase	8	1fmo	0.764	1stc	0.665	0.730
glycogen phosphorylase	5	2amv	0.709	3amv	0.590	0.684

mean of 3.0 s (standard deviation of 1.54) required for a single docking per rotatable bond. Note that this is approximately 10-fold faster than the report from Kellenberger et al., which relied on older SGI hardware and for which much less efficient compiler optimization strategies were employed.

Conclusions

Our results clearly demonstrate that synthetically generated negative data can be used effectively in estimating parameters for scoring functions in molecular docking. Over a large variety of test cases, both with respect to screening utility and docking accuracy, the newly parametrized scoring function performed at least as well as the old scoring function, which relied on a less systematic, hand-tuned approach for addressing repulsive interactions.

Apart from pure performance issues, the new approach is clearly an improvement methodologically in three respects. First, the new scoring function reformulates the formerly ad hoc penetration term (which included a dependence on ligand size) into one with a theoretically more satisfying form that can be thought of as another additive energetic effect. Second, in an operational sense, the new function returns a *single* value, which has direct interpretation in ranking ligands. While heuristic methods may be layered on top of a straightforward score-based ranking, none are required. Third, both the penetration term and the term related to noncomplementary polar contacts received significant weight in the retuned function, which comports with both intuition and theory.

By incorporating negative training data, we have been able to address two of the key challenges we set out in the original Surflex report:[19] consolidation of scoring and penetration terms and inclusion of negative training data. There is still much room for improvement. On the basis of the preliminary results here regarding treatment of desolvation effects, development of a term that treats both protein and ligand desolvation symmetrically, while taking into account issues of solvent exposure, is a high priority. This will benefit from a larger training set of positive examples, which could be greatly increased by leveraging efforts such as PDBbind.[25]

The benchmark data set established in this work is publicly available and offers a large number of diverse cases for testing screening performance of docking methods. Surflex-Dock, version 1.31, incorporating the new scoring function, performed extremely well in $^{13}/_{29}$ cases, with ROC areas of 0.95 or greater, performed very well in 10 additional cases (ROC area greater than 0.80), and showed weaker performance in the remaining 6 cases. In those cases, a simple approach that made use of two protein conformations was remarkably successful in improving performance. It is our hope that other methodological researchers in the field of molecular docking will make use of (and add to) this benchmark data set.

Acknowledgment. The authors acknowledge NIH for partial funding of the work (Grants GM070481 and CA64602). The authors are grateful to Ann Cleves for comments on the manuscript. The authors also acknowledge the contribution of BioPharmics LLC, which owns and commercializes the Surflex software, for supporting academic development of improved methodologies and for allowing free academic use of the Surflex packages.

References

(1) Walters, P. W.; Stahl, M. T.; Murcko, M. A. Virtual Screening. An Overview. *Drug Discovery Today* **1998**, *3*, 160−178.
(2) 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.
(3) Bohm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein−ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.
(4) 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.
(5) Rognan, D.; Lauemoller, S. L.; Holm, A.; Buus, S.; Tschinke, V. Predicting binding affinities of protein−ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J. Med. Chem.* **1999**, *42*, 4650−4658.
(6) Wang, R.; Liu, L.; Lai, L.; Tang, Y. SCORE: A new empirical method for estimating the binding affinity of a protein−ligand complex. *J. Mol. Model.* **1998**, *4*, 379−384.
(7) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.
(8) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(9) Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.* **1996**, *3*, 449−462.
(10) Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated docking of flexible ligands: applications of AutoDock. *J. Mol. Recognit.* **1996**, *9*, 1−5.
(11) 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.
(12) 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.
(13) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.
(14) Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J. Mol. Model.* **2003**, *9*, 47−57.
(15) Zavodszky, M. I.; Sanschagrin, P. C.; Korde, R. S.; Kuhn, L. A. Distilling the essential features of a protein surface for improving protein−ligand docking, scoring, and virtual screening. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 883−902.
(16) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.
(17) Jain, A. N. Virtual screening in lead discovery and optimization. *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 396−403.
(18) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57*, 225−242.
(19) Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499−511.
(20) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med. Chem.* **2004**, *47*, 1739−1749.
(21) Nissink, J. W.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein−ligand interaction. *Proteins* **2002**, *49*, 457−471.
(22) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004**, *56*, 235−249.
(23) Murcia, M.; Ortiz, A. R. Virtual screening with flexible docking and COMBINE-based models. Application to a series of factor Xa inhibitors. *J. Med. Chem.* **2004**, *47*, 805−820.
(24) Bradley, E. K.; Miller, J. L.; Saiah, E.; Grootenhuis, P. D. Informative library design as an efficient strategy to identify and optimize leads: application to cyclin-dependent kinase 2 antagonists. *J. Med. Chem.* **2003**, *46*, 4360−4364.
(25) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: collection of binding affinities for protein−ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.
(26) Jain, A. N. Ligand-based structural hypotheses for virtual screening. *J. Med. Chem.* **2004**, *47*, 947−961.

(27) Perkins, E.; Sun, D.; Nguyen, A.; Tulac, S.; Francesco, M.; Tavana, H.; Nguyen, H.; Tugendreich, S.; Barthmaier, P.; Couto, J.; Yeh, E.; Thode, S.; Jarnagin, K.; Jain, A. N.; Morgans, D.; Melese, T. Novel inhibitors of poly(ADP-ribose) polymerase/PARP1 and PARP2 identified using a cell-based screen in yeast. *Cancer Res.* **2001**, *61*, 4175−4183.

(28) Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002**, *45*, 2213−2221.

(29) Mount, J.; Ruppert, J.; Welch, W.; Jain, A. N. IcePick: a flexible surface-based system for molecular diversity. *J. Med. Chem.* **1999**, *42*, 60−66.

(30) Jain, A. N.; Dietterich, T. G.; Lathrop, R. H.; Chapman, D.; Critchlow, R. E., Jr.; Bauer, B. E.; Webster, T. A.; Lozano-Perez, T. A shape-based machine learning tool for drug design. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 635−652.

(31) Jain, A. N.; Koile, K.; Chapman, D. Compass: predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.* **1994**, *37*, 2315−2327.

(32) Jain, A. N.; Harris, N. L.; Park, J. Y. Quantitative binding site model generation: compass applied to multiple chemotypes targeting the 5-HT1A receptor. *J. Med. Chem.* **1995**, *38*, 1295−1308.

(33) Wei, B. Q.; Weaver, L. H.; Ferrari, A. M.; Matthews, B. W.; Shoichet, B. K. Testing a flexible-receptor docking algorithm in a model binding site. *J. Mol. Biol.* **2004**, *337*, 1161−1182.

JM050040J