Comparative Performance of Several Flexible Docking Programs and Scoring Functions: Enrichment Studies for a Diverse Set of Pharmaceutically Relevant Targets

Zhiyong Zhou,[†] Anthony K. Felts,[†] Richard A. Friesner,[‡] and Ronald M. Levy*,[†]

BioMaPS Institute for Quantitative Biology, Department of Chemistry and Chemical Biology, Rutgers,
The State University of New Jersey, Piscataway, New Jersey 08854, and Department of Chemistry,
Columbia University, New York, New York 10027

Received January 26, 2007

Virtual screening by molecular docking has become a widely used approach to lead discovery in the pharmaceutical industry when a high-resolution structure of the biological target of interest is available. The performance of three widely used docking programs (Glide, GOLD, and DOCK) for virtual database screening is studied when they are applied to the same protein target and ligand set. Comparisons of the docking programs and scoring functions using a large and diverse data set of pharmaceutically interesting targets and active compounds are carried out. We focus on the problem of docking and scoring flexible compounds which are sterically capable of docking into a rigid conformation of the receptor. The Glide XP methodology is shown to consistently yield enrichments superior to the two alternative methods, while GOLD outperforms DOCK on average. The study also shows that docking into multiple receptor structures can decrease the docking error in screening a diverse set of active compounds.

I. INTRODUCTION

Virtual screening has become a widely used approach to lead discovery in the pharmaceutical industry. When a high-resolution structure of the biological target of interest is available, the most common methodology for performing virtual screening involves the use of a flexible docking algorithm, in which conformational sampling methods are used to position the ligand in the receptor, and some sort of scoring function is applied to obtain a predicted free energy of binding. A number of powerful software programs, e.g., GOLD,[1−4] FlexX,[5] DOCK,[6,7] Glide,[8,9] Surflex,[10,11] and LigandFit,[12] have been developed over the past several decades to carry out docking calculations, and good success in both binding mode and binding affinity prediction has often been achieved in selected test cases. A more challenging goal has been to improve the robustness of the methods with regard to both structural and energetic prediction; all of the above programs on occasion manifest both false negatives (active compounds which are not appropriately docked or scored by the methodology) and false positives (weakly binding compounds whose binding affinity is seriously overpredicted). A number of comparative evaluations of docking programs, conducted over the past several years, confirm this general picture.[13−17]

One of us has recently described a new approach[9] that has been implemented in the Glide program (the extra precision, or XP, Glide methodology) which incorporates novel terms into the binding free energy scoring function (as compared to standard scoring approaches,[18−23] which are similar with regard to functional form) and appears, in preliminary tests, to substantially enhance the ability of Glide to pick out known active compounds from a random ligand database. Relative to the standard precision (SP) Glide scoring function,[8] improvements in enrichment of roughly an order of magnitude are reported for XP Glide,[9] for a large and diverse set of ligands and receptors. However, this comparison is entirely internal to the Glide program and does not provide any calibration as to how well, or poorly, alternative approaches would fare with the particular data set under study. The present paper is aimed at addressing this question, employing, in addition to Glide, the DOCK and GOLD programs which are both widely used in academia and in the pharmaceutical and biotechnology industries. We experiment with a number of docking and scoring approaches available in each program in order to obtain as much comparative data as possible.

In addition to evaluating the relative abilities of the various docking methods to identify known active compounds embedded in a random database, another objective of this work is the validation of a new approach to enrichment studies described in ref 9. In the great majority of enrichment studies in the literature, no attempt is made to separate misdockings due to induced fit effects from scoring errors; ligands are simply docked into a single rigid version of the receptor, and their scores are compared with those of database ligands. However, if the goal is to fairly calibrate the accuracy of the scoring function (or to improve the parametrization by fitting theory to experimental data), it makes no sense to include ligands that do not “fit” into a particular receptor structure, due to significant steric clashes, in the enrichment test set. These ligands presumably dock in grossly incorrect poses and would require a substantial modification of the receptor active site conformation in order to dock correctly. The score associated with such a grossly incorrect pose cannot be expected to correspond to an

* Corresponding author phone: (732)445-3947; fax: (732)445-5958; e-mail: ronlevy@lutece.rutgers.edu.
[†] Rutgers, The State University of New Jersey.
[‡] Columbia University.

10.1021/ci7000346 CCC: $37.00
Published on Web 06/23/2007

experimentally measured value, and inclusion of such data in an “enrichment factor” does not provide a reasonable measure of the accuracy of the scoring function when induced fit effects are *not* an issue.

Computation of induced fit effects, in cases where the ligand does not fit into a specified receptor conformation, is quite feasible, as we have shown in a recent publication.[24] However, such an approach requires the use of a flexible receptor as well as a ligand and hence is in general much more computationally intensive than rigid receptor docking. There are also questions about how to compare the scores of ligands docked to different receptor conformations; incorporation of receptor strain energy into binding affinity prediction is an area at present in its infancy. For these reasons, we believe there is considerable value in separating the docking/scoring problem into two components: (1) calculation of induced fit effects (including strain energy estimation) when these are necessary to achieve a reasonable prediction of the binding mode and (2) docking and scoring of compounds which are sterically capable of docking more or less correctly into a specified rigid conformation of the receptor. We consider only component (2) in the present paper (as was also done in ref 9).

An approach to implement the separation suggested above is to simply eliminate the ligands from the training and test sets which do not dock in a “reasonable” pose (“nonfitting”) into the receptor conformation selected for the study. This is the approach adopted in ref 9. The assessment of docking accuracy was performed using RMSDs when a crystal structure was available, and otherwise employing visual inspection, with the correct pose inferred by analogy with known complexes in the PDB with ligands of a similar structure. The detailed criteria were documented in ref 9. A possible objection to this approach is that the ligand was misdocked specifically by Glide but would be docked successfully with alternative programs. In the present paper, we examine “misdocked” ligands from the data sets of ref 9, docking these ligands with the GOLD and DOCK programs. This investigation provides an important cross-check on our previous protocols and assumptions with regard to sorting ligands into the categories of properly or improperly docked, reducing the bias inherent in the use of a single program. While misdocking with multiple programs does not prove that the compound cannot be docked into a rigid receptor, it certainly strongly suggests that this is the case and, additionally, provides a check on whether our assessment of misdocking was unduly influenced by the score assigned to the ligand by Glide.

The paper is organized as follows. Section II discusses the data sets that we have assembled, including the initial division of ligands into “fitting” and “nonfitting” into the receptor conformation used in the study. In section III, we briefly discuss the methodologies employed in GOLD, DOCK, and Glide, with regard to both sampling and enrichment. Section IV presents results comparing enrichment factors, using fitting compounds only, for the various test cases. Results obtained with all three programs for the nonfitting compounds are then separately analyzed. Section V, the conclusion, provides a brief summary and discusses future directions.

Table 1. Data Set Used To Compare Virtual Database Screening[a]

PDB code	description	no. well-docked actives	no. poor-docked actives
	(a) Glide Training Set		
1fjs	factor Xa	9	4
1bji	neuraminidase	9	0
1hpx	HIV-1 protease	9	5
1cx2	cyclooxygenase-2	13	0
1e66	acetylcholinesterase	20	0
1ett	thrombin	15	1
1kim	thymidine kinase	4	0
1aq1	human cyclin dep. kinase	6	4
1bl7	p38 map kinase	27	9
1kv2	human p38 map kinase	10	0
1m17	EGRF tyrosine kinase	107	10
1qpe	Lck kinase	87	34
1rt1	HIV reverse transcriptase	23	6
1tmn	thermolysin	5	1
3ert	human estrogen receptor	8	2
	(b) Glide Test Set		
1m4h	BACE	34	43
1dan	factor VIIa	40	53
1fm6	PPARg (closed form)	32	61
1fm9	PPARg (open form)	25	68
1y6b	Vegfr2 (closed form)	21	90
1ywn	Vegfr2 (open form)	26	85
1aq1	human cyclin dep. kinase	143	110
1ett	thrombin	15	25

[a] All active compounds have experimental activities less than 10 μM except those for neuraminidase.

II. DATA SET

We have used the same set of receptors and ligands as were employed to evaluate the Glide XP scoring function in ref 9. This data set is divided into two components: a training set, which was used to parametrize Glide XP, and an independent test set. Fourteen targets which are of pharmaceutical interest are contained in the training set, as shown in Table 1a. One of them (p38 MAP kinase) is represented by two alternative cocrystallized receptor sites. The crystallographic resolution of all these 15 proteins is less than 3.0 Å (9 of them are less than 2.2 Å). The receptors for these screens cover a wide variety of receptor types and therefore provide a proper test of the docking methods. All protein structures were prepared using the procedure as stated in the previous Glide methodology paper.[8]

The test set, described in Table 1b, consists of four new receptors, with appropriate sets of cognate ligands, and two receptors (human cyclin dependent kinase 2, or CDK2, and thrombin) studied previously, with new sets of ligands. Among the four new receptors, two of them (PPARγ and Vegfr2) are investigated using two different conformations of the receptor—a closed form and an open form—which are appropriate for binding different classes of ligands. While a larger test set would be desirable (and will be employed in subsequent publications), development of suitable data sets is highly labor intensive; and the current test set is capable of providing an assessment as to whether there is large overfitting in the Glide XP results with the training set.

For PPARγ and Vegfr2 targets, only a small fraction (from 19% to 34%) of the active compounds can be correctly docked by Glide XP if only one form of the receptor (either the closed or open form) is used. However, if both forms of the receptor are used, 61% and 42% of all active compounds can be correctly docked for PPARγ and Vegfr2 receptors, respectively. Therefore, for these cases, docking into multiple receptor structures instead of a single structure is an effective way to decrease the misdocking due to steric clashes, which is a major error in screening a large, diverse set of active compounds.

Comparing parts a and b of Table 1, there are many more poorly docked active compounds in the test set than in the training set. When the training set was constructed, in many cases a relatively small number of active compounds were included, and even in cases where a larger number of compounds are employed, they were typically derived from a small number of literature sources. However, for the test set, we have collected a significant number of compounds from the literature, using a number of different literature sources. Consequently, the diversity of the test set is significantly larger than that of the training set, leading to a larger number of compounds being misdocked into a given structure of the receptor. In a realistic laboratory application of virtual screening, where the data set to be screened is typically a highly diverse pharmaceutical compound collection, we would thus expect the fraction of misdocked compounds for all but the most rigid receptors to be substantial. The consequences of this observation are discussed further below.

In the present study, the same set of active ligands, including both well-docked and poorly docked ones, are being docked with each program. The classification of the well-docked or poorly docked are based on Glide XP results. As database ligands [A 1K druglike ligand decoys set was created by selecting 1000 ligands from a 1 million compound library that was chosen to exhibit "druglike" properties. This decoy set is available at https://www.schrodinger.com/ProductInfo.php?mID=6&sID=6&cID=18.], we employed "druglike" decoys that averaged 400 in molecular weight (the "dl-400" data set) in all cases except thymidine kinase (1kim). For 1kim, which has a very small active site, we used a similar (but in this case more competitive) set with an average molecular weight of 360 (the "dl-360" data set). The detailed approach to creating these test databases and their property distributions was described in a previous publication.[8] We believe these compounds are representative of the chemical sample collections of pharmaceutical and biotechnology companies. As such, they should provide a fair and stringent test of the efficacy of the docking method. Each screen used 1000 decoy database ligands and between 4 and 253 known active binders as shown in Table 1. All selected known binders have experimental activities less than 10 μM except those for neuraminidase. The references for their structures and biological activities can be found in a previous publication.[9] Like the database ligands, the known binders were also MMFF94s-optimized. In these cases, we used input geometries obtained via a MacroModel[25] conformational search.

While the training set described above has been used as such for parametrization of Glide XP (and, to a small extent, Glide SP as well), it is worth pointing out that neither GOLD nor DOCK have been trained using this data set. Thus, while comparison of Glide XP to GOLD and DOCK is one (major) objective of the paper (and is best addressed by examining the test set comparisons), the "training" set provides additional data with which to evaluate the absolute level of performance of GOLD and DOCK, their performance relative to each other, and the relative performance of the various options within each program that are evaluated below. To our knowledge, an evaluation of GOLD or DOCK using exclusively "fitting" compounds has not been reported in the literature; the present study provides this information using an extensive database of such compounds, comprising all the data summarized in Table 1a,b.

III. DOCKING METHODOLOGIES

Glide (Schrodinger, Inc.). The Glide (Grid-Based Ligand Docking With Energetics, version 4.0) algorithm approximates a systematic search of positions, orientations, and conformations of the ligand in the receptor binding site using a series of hierarchical filters.[8,9,26] The shape and properties of the receptor are represented on a grid by several different sets of fields that provide progressively more accurate scoring of the ligand pose. The fields are computed prior to docking. The binding site is defined by a rectangular box confining the translations of the mass center of the ligand. A set of initial ligand conformations is generated through an exhaustive search of the torsional minima, and the conformers are clustered in a combinatorial fashion. Each cluster, characterized by a common conformation of the "core" and an exhaustive set of "rotamer group" conformations, is docked as a single object in the first stage.[8] The search begins with a rough positioning and scoring phase that significantly narrows the search space and reduces the number of poses to be further considered to a few hundred. In the following stage, the selected poses are minimized on precomputed OPLS-AA van der Waals and electrostatic grids for the receptor. In the final stage, the 5−10 lowest-energy poses obtained in this fashion are subjected to a Monte Carlo procedure in which nearby torsional minima are examined, and the orientation of peripheral groups of the ligand is refined. The minimized poses are then rescored using the GlideScore function, which is an expanded version of ChemScore[19] with force field-based components and additional terms accounting for solvation and repulsive interactions. The choice of the best pose is made using a model energy score (Emodel) that combines the energy grid score, GlideScore, and the internal strain of the ligand.[8] We investigated both Standard Precision mode (SP) and Extra Precision mode (XP) of Glide in this comparative study.

The Glide XP methodology has been described in detail in ref 9; we briefly summarize the important features here. The starting point for XP scoring is a modified version of ChemScore, as in the case of SP; however, novel terms are used to handle physical effects that are missing from ChemScore. Desolvation penalties are applied by docking explicit waters into the highest scoring docked complexes and evaluating the solvation of polar and charged ligand and protein groups by counting the number of neighboring waters and comparing these values to statistics extracted from a database of correctly docked active ligands. Molecular recognition motifs based on the concept of hydrophobic enclosure of the ligand by the protein are defined, and incremental increases in binding affinity are added to the ligand score when the appropriate motifs are recognized. In order to properly evaluate these new terms, a considerable augmentation of the sampling algorithm, which is carried out at higher resolution, is required; the algorithm itself, based on growing side chains from core positions identified by SP docking, is discussed further in ref 9. Additional terms involving special treatment of salt bridges, π-cation interactions, and various other specialized medicinal chemistry motifs are described in ref 9. The XP scoring function has been parametrized using the training set of 15 receptor structures and cognate "fitting" ligands, as is discussed further below.

GOLD (Cambridge Crystallographic Data Center). Version 2.2 of the GOLD (Genetic Optimization for Ligand Docking) docking program was evaluated in the present study. The GOLD program uses a genetic algorithm (GA) to explore the full range of ligand conformational flexibility and the rotational flexibility of selected receptor hydrogens.[1−4] The mechanism for ligand placement is based on fitting points. The program adds fitting points to hydrogen-bonding groups on the protein and ligand and maps acceptor points in the ligand on donor points in the protein and vice versa. Additionally, GOLD generates hydrophobic fitting points in the protein cavity onto which ligand CH groups are mapped. The genetic algorithm optimizes flexible ligand dihedrals, ligand ring geometries, dihedrals of protein OH and NH3 groups, and the mappings of the fitting points. The docking poses are ranked based on a molecular mechanics-like scoring function. There are two different built-in scoring functions in the GOLD program—GoldScore and ChemScore. Note that the ChemScore function implemented in GOLD[4] is an optimized version of the original chemscore function developed by Eldridge et al.[19] In parallel, the performance of two combined docking protocols was also studied. In the first combined protocol, "GoldScore-reChemScore", the dockings produced with the GoldScore function are rescored and reranked using the ChemScore function; in the second combined protocol, "ChemScore-reGoldScore", the docking produced with the ChemScore function are rescored and reranked using the GoldScore function. In both protocol, the Simplex algorithm (local optimization) is used to relax each docking in the alternative scoring function. We also compared two different speed modes—default settings (1×) and 7−8 times speedup settings (8×). In the present work, the binding site was defined as a spherical region which encompasses all protein atoms within 5.0 Å of each crystallographic ligand atom. Protein and ligand input structures were prepared as described above. Default GA settings were used for all calculations.

DOCK (UCSF). The version 5.2.0 of DOCK[6,7] was used in these studies. DOCK characterizes concavities on a protein surface using sets of spheres positioned on top of a Connolly surface generated on the binding pocket. The centers of these spheres characterize positions where ligand atoms can be found in the binding pocket. DOCK uses a graph matching algorithm to position the atoms of a ligand onto the centers of the spheres. A minimization of the ligand poses is performed allowing DOCK to refine the ligand position in the binding pocket. Flexibility of the ligands is modeled by treating the ligand as a series of fragments, where a central fragment (the anchor) is docked first followed by sequentially docking the outer fragments around the anchor. As each fragment is docked, neighbor fragments are combined.

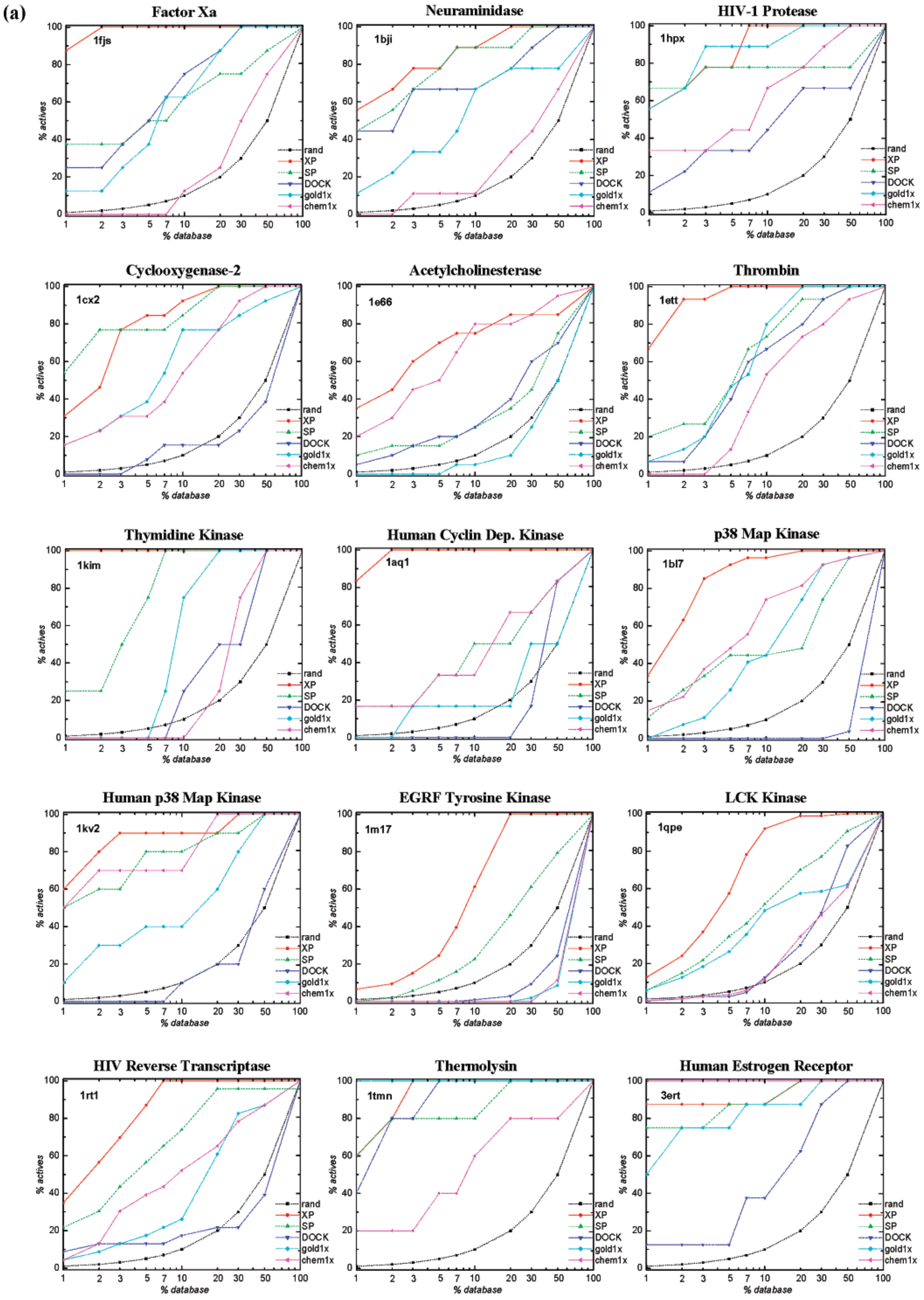
For each target, the Connolly molecular surface was calculated using a probe radius of 1.4 Å. Inside the binding pocket, spheres are generated therein with the DOCK program SPHGEN. SPHGEN outputs the spheres in clusters that overlap each other. Clusters were examined for each target, and the cluster covering the known binding site was chosen by selecting those spheres within 7.5 Å of the cocrystallized ligand. Compounds were docked allowing for ligand flexibility, using the grid-based energy scoring option for minimization after initial placement in the site. The box for the scoring grid was defined such that all spheres were enclosed with an extra 5.0 Å added in each dimension. Scoring grids for contact and energy scores were calculated with a grid spacing of 0.3 Å. The bump check was set such that compounds with atoms closer than half the sum of the van der Waals radii of the respective atoms were rejected. A 6−12 Lennard-Jones van der Waals potential was used along with a Coulomb potential using a distance-dependent dielectric constant of $4r$ to simulate solvation effects. The energy cutoff was 10.0 Å. The radii used were those in the vdw_AMBER_parm99.defn set. Ligand atoms were matched to receptor spheres using the anchor first search with the anchor size set to 10 atoms. The automatic matching option was used, and conformations were generated on the fly with the torsion drive option.

IV. RESULTS AND DISCUSSIONS

The objective of the present study is to compare how well three widely used molecular docking programs perform during virtual database screening when applied to the same protein target and ligand set. In the context of virtual screening, the measure of performance is the ability of the program to prioritize seeded active compounds for a particular target relative to the decoy compounds in the database. We have obtained enrichment data for the training set and the test set as described above.

Figure 1 displays the percent of known actives recovered as a function of the percent of the ranked database sampled for Glide XP, Glide SP, DOCK, GOLD 1xGoldScore, and GOLD 1xChemScore, for all test cases in the training set and test set. The enrichment curves show that Glide XP gives the best performance for six out of eight cases when evaluated on the test set. On the training set with 15 cases, Glide XP outperforms the other methods in database enrichment ability for 13 cases.

Figure 2a shows the average enrichment curves over all 15 training set targets for all 11 docking methods in the present study. All docking methods could identify active compounds from the ligand database since all enrichment curves are significantly better than the random selection curve. GOLD methods have similar performance as DOCK. They found between 30% and 55% of known actives in the top 10% of the ranked database on average. Glide SP found 67% of actives in the top 10% of the ranked database. Glide XP achieves better enrichment than the other methods, unsurprisingly since this training set was used to parametrize Glide XP. On average, Glide XP found 92% of the known actives in the top 10% of the ranked database. For the GOLD program, the performance of GoldScore is somewhat better than ChemScore. The known actives found in the top 10% of the ranked database are 55% and 45% for 1xGoldScore and 1xChemscore, respectively. However, the docking speed of GoldScore mode (8.5 min per ligand for 1x settings) is


(a)
Factor Xa
1fjs
Neuraminidase
1bji
HIV-1 Protease
1hpx
Cyclooxygenase-2
1cx2
Acetylcholinesterase
1e66
Thrombin
1ett
Thymidine Kinase
1kim
Human Cyclin Dep. Kinase
1aq1
p38 Map Kinase
1bl7
Human p38 Map Kinase
1kv2
EGRF Tyrosine Kinase
1m17
LCK Kinase
1qpe
HIV Reverse Transcriptase
1rt1
Thermolysin
1tmn
Human Estrogen Receptor
3ert
% actives
0
20
40
60
80
100
% database
1
2
3
5
7
10
20
30
50
100
rand
XP
SP
DOCK
gold1x
chem1x

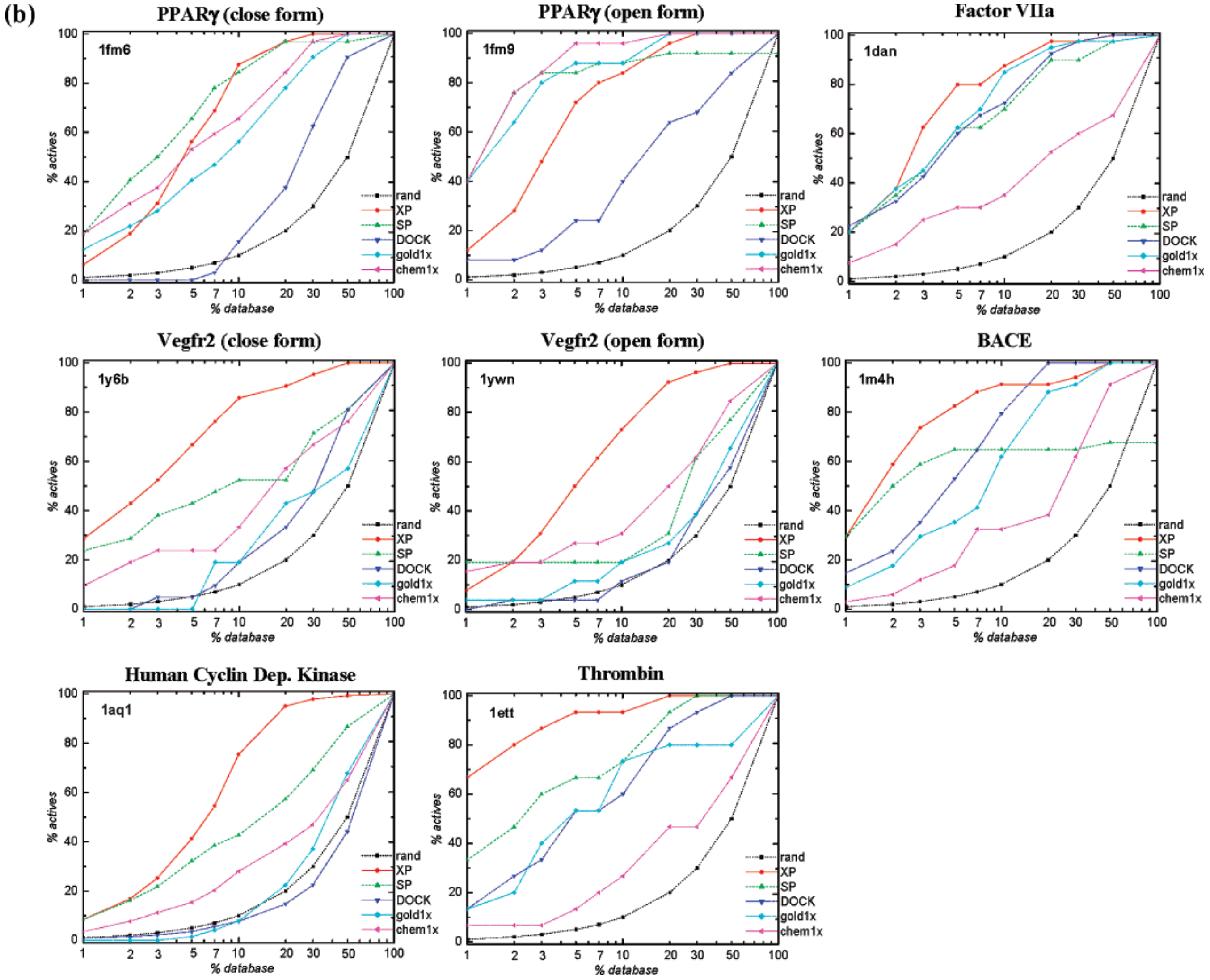


Figure 1. Percent of known actives found (*y*-axis) vs percent of the ranked database screened (*x*-axis) for Glide XP (XP, red solid), Glide SP (SP, green dash), DOCK (DOCK, blue dash dot), GOLD GoldScore1x (gold1x, cyan dash dot dot), and GOLD ChemScore1x (chem1x, magenta short dash). Black dotted lines (rand) show results expected by chance. The listed PDB codes are defined in Table 1: (a) Glide training set and (b) Glide test set.

about 3 times slower than the Chemscore mode (2.8 min per ligand for 1x settings). The combined docking protocols which correspond to docking with one scoring function and rescoring with the other do not improve the performance relative to the single scoring functions. The average docking time for each method is listed in Table 2.

Figure 2b presents the average enrichment curves for the independent test set (Table 1b) for Glide XP and SP, DOCK, and the four noncomposite GOLD methods. DOCK achieves slightly better results for the test set than the training set (38% vs 33% of the known actives in the top 10% of the ranked database. This difference likely corresponds to statistical variation based on sample size.). On average, DOCK performs similarly to the GOLD results for the test set. They found between 32% and 51% of the known actives in the top 10% of the ranked database. Glide SP found 62% of the actives in the top 10% of the ranked database. Glide XP achieves the best enrichment. On average, for the test set Glide XP found 85% of the known actives in the top 10% of the ranked database. In summary, compared with training set results, there is some quantitative degradation in the enrichment for all methods except DOCK, but no significant overfitting is found at this percentage of active compound recovery. In fact, it is possible that the slight degradation in performance of XP (seen for GOLD as well) is not because of overfitting, but because one is dealing with a more challenging set of receptors and/or active compounds. A further discussion of XP performance on the test set will be presented below.

On average, in the top 2% of the ranked database, Glide XP found 68% and 38% of known actives for the training set and test set, respectively. There are several possible explanations for the significant decrease in performance seen at this level of accuracy, which is a highly demanding one (i.e., ranking active compounds ahead of nearly all of the decoy ligands). First, at least some of the difference could be due to overfitting of Glide XP for the active ligands in the training set (thus enabling training set ligands to be ranked higher than they would be otherwise). Second, it is possible that there are novel molecular recognition motifs in one or more of the additional receptors in the test set, which have not been incorporated into the Glide XP scoring function as of yet. Glide XP is a combination of a physics based approach (to determine functional forms of the scoring

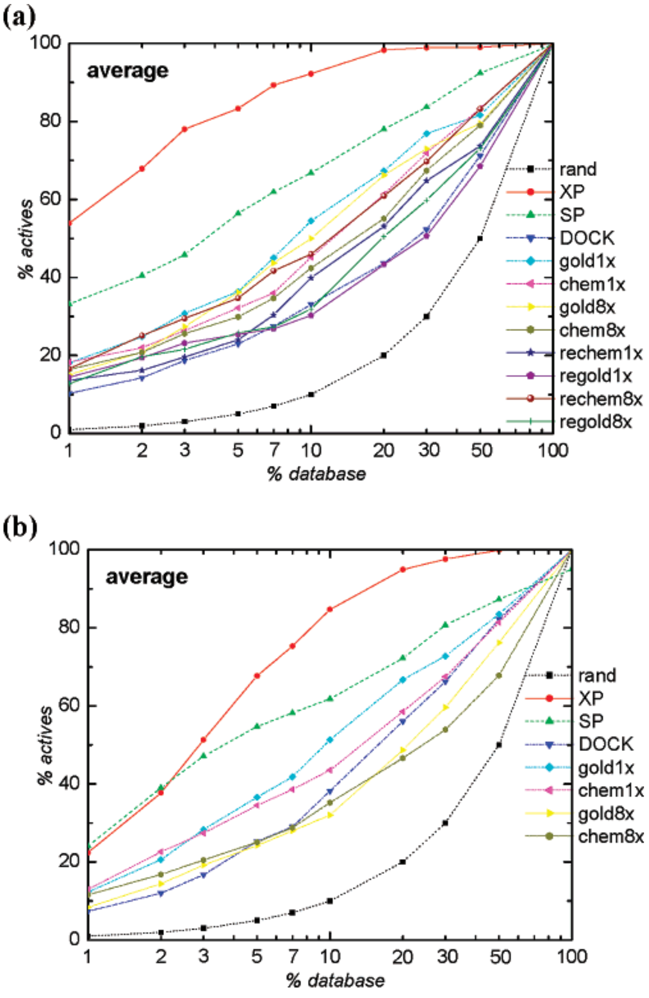


Figure 2. Average percent of known actives found over training set or test set (*y*-axis) vs percent of the ranked database screened (*x*-axis) for Glide XP (XP, red solid), Glide SP (SP, green dash), DOCK (DOCK, blue dash dot), GOLD GoldScore1x (gold1x, cyan dash dot dot), GOLD ChemScore1x (chem1x, magenta short dash), GOLD GoldScore8x (gold8x, yellow short dot), GOLD ChemScore8x (chem8x, dark yellow short dash dot), GOLD GoldScore1x-reChemScore (rechem1x, navy solid), GOLD ChemScore1x-reGoldScore (regold1x, purple solid), GOLD GoldScore8x-reChemScore (rechem8x, wine solid), GOLD ChemScore8x-reGoldScore (regold8x, olive solid). Black dotted lines (rand) show results expected by chance: (a) Glide training set and (b) Glide test set.

terms) with an expert system component (to identify particular chemistries/geometries that should be rewarded or penalized as compared to "normal" scoring); the expert system performance is dependent upon having relevant examples in the training set. The existence of novel examples missing from the training set is a problem distinct from "overfitting", which in fact would represent a very different sort of difficulty (recognition of motifs identified from previous targets which should not be rewarded or penalized with the same weights in the current targets). Analysis of which problem described above is dominant awaits further investigation.

Table 2. Average Docking Time (min) per Ligand on a 2.2 GHz AMD (Athlon MP 2800+) Single Processor[a]

method	description	min per ligand
XP	Glide XP	7.0
SP	Glide SP	0.42
DOCK	DOCK	4.0
1xgold	GOLD GoldScore 1x	8.5
1xchem	GOLD ChemScore 1x	2.8
8xgold	GOLD GoldScore 8x	1.0
8xchem	GOLD ChemScore 8x	0.35
1xrechem	GOLD GoldScore1x-reChemScore	8.5
1xregold	GOLD ChemScore1x-reGoldScore	2.8
8xrechem	GOLD GoldScore8x-reChemScore	1.0
8xregold	GOLD ChemScore8x-reGoldScore	0.35

[a] Times for combined "GoldScore-reChemScore" are identical to those for the GoldScore functions. Times for combined "ChemScore-reGoldScore" are identical to those for the ChemScore functions.

To characterize the overall performance of each docking method for database screening, Table 3 reports, for both training and test sets, a new measure of enrichment defined as the average number of database decoys outranking the active compounds in the database. Specifically, the number of database decoys with a score that is superior to each active is tabulated, these values are summed, and the result is then divided by the total number of active compounds in the data set. We believe that this metric is superior to standard definitions of enrichment, which punish active ligands when they are outranked by other active ligands; this is a particularly serious problem when the active test suite contains a large number of compounds. A "perfect" score based on this metric would thus be zero (no database ligands outranking any active compounds), and smaller numbers are better. Also, this metric can differentiate the following two circumstances with the same enrichment factor: suppose there are 10 actives in the top 50 database rankings, (a) in one situation, the ranks of 10 actives are from 1 to 10 and (b) in the other, the ranks of 10 actives are from 41 to 50. Obviously, the former is better. This new metric (0 vs 40) can clearly distinguish these two situations. The average metric for all 15 targets in Table 3a indicates the following order of performance Glide XP, Glide SP, GOLD, and DOCK. Compared with the training set results in Table 3a, Glide XP results for the test set (Table 3b) are significantly degraded (20 for the training set vs 39 for the test set), but it is still the best method in the test set by a significant margin. Surprisingly, the performance of DOCK improves on the test set (341 for the training set vs 265 for the test set).

In order to check the consistency of our new measure of enrichment with the standard definition, the average standard enrichment factors at 10% of ranked database are shown in Table 4. The standard enrichment factor (EF) is defined as

$$\mathrm{EF} = (\mathrm{HITS}_{\mathrm{sampled}}/\mathrm{HITS}_{\mathrm{total}})/(N_{\mathrm{sampled}}/N_{\mathrm{total}})$$

Here, N_{total} is the number of ligands in the docked database, N_{sampled} is the number of ligands in the docked database to be examined, $\mathrm{HITS}_{\mathrm{total}}$ is the total number of the known active ligands, and $\mathrm{HITS}_{\mathrm{sampled}}$ is the number of known active ligands found in the top N_{sampled} ligands of the docked database. Compared to the last rows in parts a and b of Table 3, the new enrichment metric and the standard enrichment factor give very similar trends with regard to the performance. The advantage of the new enrichment metric, however, is that it is not dependent on the total number of active ligands making a better comparison between different databases possible.

Besides the well-docked active set, we also compare the performance of all methods for the set that was poorly docked

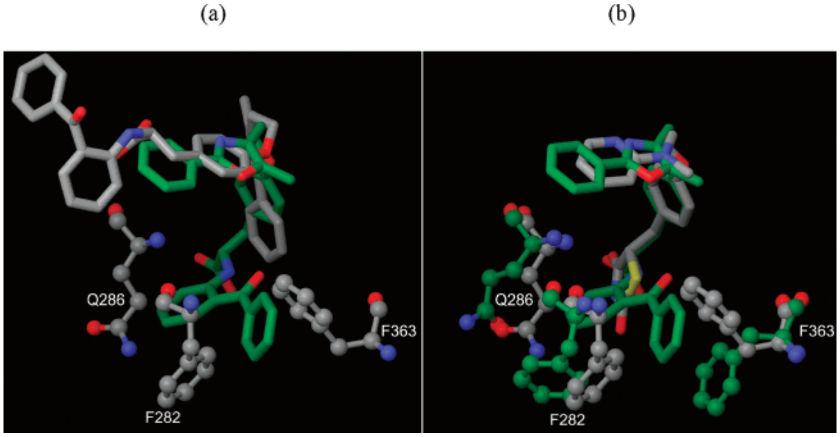


Figure 3. An example of a misdocked ligand due to steric clashes. (a) Misdocked pose (carbon atoms are colored in gray) generated by Glide XP for ligand 1293-1 docking into 1fm6 crystal structure (carbon atoms are colored in gray). The "correct pose" (carbon atoms are colored in green) is shown for comparison. (b) Crystal structure of the binding site of 1fm9 (PPARγ open form) with its native ligand 1293-1 (carbon atoms are colored in green) superimposed on the 1fm6 (PPARγ closed form) structure with its native ligand 1241-2 (carbon atoms are colored in gray). Only residues that involve steric clashes with ligand are shown. Hydrogen atoms are not shown. The molecular representations for ligands and proteins are "tube" and "ball and stick", respectively. The color schemes for elements are carbon (gray or green), oxygen (red), and nitrogen (blue).

Table 3. New Enrichment Metrics Defined as the Average Number of Outranking Decoy Ligands over Well-Docked Actives for (a) the Glide XP Training Set and (b) the Glide XP Test Set[a]

(a) Glide XP Training Set

pdb ID	XP	SP	DOCK	1x gold	1x chem	8x gold	8x chem	1xre chem	1xre gold	8xre chem	8xre gold
1fjs	1	187	87	82	428	273	427	414	554	521	465
1bji	25	37	92	248	417	224	340	263	224	132	222
1hpx	16	167	290	18	98	155	362	131	138	327	261
1cx2	24	22	511	133	115	91	89	107	316	66	200
1e66	111	344	353	514	123	435	100	235	307	108	428
1ett	2	70	98	58	180	159	313	289	406	299	377
1kim	0	29	259	100	252	68	220	208	582	148	199
1aq1	3	216	456	380	272	291	256	352	376	270	320
1bl7	7	183	806	126	95	92	79	285	96	123	66
1kv2	26	57	451	157	42	119	51	211	139	46	181
1m17	41	279	714	845	756	802	669	749	759	589	670
1qpe	13	157	316	414	426	474	435	503	573	402	620
1rt1	11	83	512	233	212	230	264	238	387	220	287
1tmn	9	32	10	2	157	50	345	356	414	207	274
3ert	14	23	165	46	0	114	0	1	5	2	221
average	20	126	341	224	238	238	263	289	352	231	319

(b) Glide XP Test Set

pdb ID	XP	SP	DOCK	1x gold	1x chem	8x gold	8x chem
1aq1	25	206	574	403	379	283	338
1dan	30	75	55	59	351	96	510
1ett	9	52	95	204	357	286	520
1fm6	45	48	276	110	77	153	113
1fm9	44	82	245	21	7	267	83
1m4h	35	315	48	99	246	634	754
1y6b	52	221	347	456	302	435	352
1ywn	69	310	477	433	285	396	273
average	39	164	265	223	250	319	368

[a] See Table 2 for the meaning of the table headings.

Table 4. Average Standard Enrichment Factors at 10% of Ranked Database for Different Methods[a]

set	XP	SP	DOCK	1x gold	1x chem	8x gold	8x chem	1xre chem	1xre gold	8xre chem	8xre gold
training	9.2	6.7	3.3	5.5	4.5	5.0	4.2	4.0	3.0	4.6	3.2
test	8.5	6.2	3.8	5.1	4.4	3.2	3.5				

[a] The standard enrichment factor is defined in the text. (See Table 2 for the meaning of the table headings).

by Glide XP. We visually inspect the binding modes of these known actives generated by docking programs and find they are quite different or missing some key interactions compared with their experimental binding modes or analogues. For example, as shown in Figure 3a, docking of ligand 1293-1 into the 1fm6 (PPARγ closed form) structure yields a ligand pose with an rmsd of 10.9 A when compared to the structure of 1293-1 in its cognate receptor 1fm9 (PPARγ open form).

Table 5. Number of Poorly Docked Active Compounds in the Top 10% of the Ranked Database[a]

(a) Glide Training Set

pdb ID	no. all misdock	XP	SP	DOCK	1x gold	1x chem	8x gold	8x chem	1xre chem	1xre gold	8xre chem	8xre gold
1fjs	5	5	4	1	2	0	1	0	1	0	1	0
1hpx	5	5	4	4	4	3	2	0	1	2	1	0
1ett	1	1	1	1	1	1	1	0	0	0	0	0
1aq1	4	2	1	0	0	1	0	1	0	0	1	0
1bl7	9	1	1	0	2	2	2	3	1	2	3	3
1m17	11	2	3	1	0	0	0	0	0	0	0	0
1qpe	34	20	15	2	8	5	5	3	4	3	7	1
1rt1	6	3	2	0	0	1	1	2	0	0	2	1
1tmn	1	1	1	1	1	0	1	0	0	1	0	1
3ert	2	1	0	1	1	2	0	2	2	1	2	2
total	78	41	32	11	19	15	13	11	9	9	17	8

(b) Glide Test Set

pdb ID	no. all misdock	XP	SP	DOCK	1x gold	1x chem	8x gold	8x chem
1aq1	110	7	8	7	19	6	13	5
1dan	53	29	30	36	31	9	26	9
1ett	25	10	15	16	15	4	14	3
1fm6	61	17	22	9	26	34	13	30
1fm9	68	10	42	6	25	34	21	28
1m4h	43	15	19	34	24	8	6	1
1y6b	90	13	5	12	7	15	6	12
1ywn	85	22	10	12	5	13	5	9
total	535	123	151	132	152	123	104	97

[a] See Table 2 for the meaning of the table headings.

The primary reason is that 1fm6 was cocrystallized with a much smaller ligand 1241-2, in which some side-chain atoms of a number of residues protrude into the binding site, thus blocking binding of the larger ligand (1293-1) in the correct pose. The most significant differences between the two structures are Phe363, Phe282, and Gln286, which in 1fm6 is rotated to a conformation that in rigid docking would block the terminal phenyl groups of 1293-1 (Figure 3b). Table 5 reports the number of poorly docked active compounds in the top 10% of the ranked database. For the training set, Glide XP prioritizes most poorly docked actives for 7 targets of 10. From the total number of poorly docked actives in the top 10% of the ranked database (as shown in the last row of Table 5a), Glide XP and SP appear to outperform other methods in ranking poorly docked actives.

For the test set, the advantage displayed by Glide XP in ranking poorly docked actives disappears, and that of Glide SP is significantly diminished. This suggests that the somewhat better ability to recognize partially correct structures may be more dependent upon the fitting data set than the performance of the scoring function for well-docked actives. The principal conclusion is that none of the programs tested performed very well when assessing the ability to rank poorly docked compounds in the top 10% of the ranked database.

V. CONCLUSION

We have carried out extensive comparisons of several docking programs and scoring functions using a large data set of pharmaceutically interesting targets and active compounds. The Glide XP methodology was shown to consistently yield enrichments superior to the alternative methods not only for the training set (Table 1a) used to develop XP but also for the independent test set (Table 1b). Glide SP scoring shows improvement as compared to the scoring in GOLD and DOCK, presumably in part because it has some component of XP scoring mixed in with the more standard terms originally derived from ChemScore (the same starting point as was used to develop GOLD). Most versions of GOLD significantly outperform DOCK on average, although results vary for individual receptors; the various scoring options in GOLD do equally well, with the exception of "rescoring" with GOLD which appears to result in degradation of performance. These conclusions apply to well docked compounds; for misdocked compounds, based on the test set results, all methods perform roughly equally poorly.

From the point of view of computational efficiency, the CPU time required on average for Glide XP calculations (7.0 min per ligand) is larger than other methods except the most accurate version of Goldscore (8.5 min per ligand). This extra cost for Glide XP is the tradeoff for the higher enrichment factors obtained. Glide SP delivers the second best overall enrichment performance while providing a considerable speedup (0.42 min per ligand) as compared to all approaches with the exception of the fast version of GOLD Chemscore setting.

While the XP scoring function can be improved, the dominant error at this point in screening a large, diverse set of active compounds with a single receptor is clearly going to be misdocking due to steric clashes which arise because the receptors are modeled as rigid structures. If virtual screening is to deliver reliable results, covering a wide range of chemotypes, this problem has to be successfully attacked. There are various approaches that are promising, including docking into multiple receptor structures (illustrated here by the PPARγ and Vegfr2 cases—note that the fractions of compounds misdocked into both closed and open forms of

PPARγ receptors and Vegfr2 receptors are quite small) and also using induced fit techniques.[24] The multiple receptor structures could be obtained from multiple X-ray crystal structures or structurally diverse high quality models generated using a torsion angle sampling tool.[27] An extensive investigation as to how well these alternative possible solutions work will be necessary in order to make significant progress.

ACKNOWLEDGMENT

This work was supported in part by a grant (GM-30580) from the National Institutes of Heath and by NIH NRSA fellowships to Zhiyong Zhou (R90 DK071502). We thank Matt Repasky from Schrodinger Inc. for providing the data sets (both training set and test set) of receptors and active ligands.

REFERENCES AND NOTES

(1) 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* (1), 43−53.

(2) 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* (3), 727−748.

(3) Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein-ligand interaction. *Proteins: Struct., Funct., Genetics* **2002**, *49* (4), 457−471.

(4) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D., Improved protein-ligand docking using GOLD. *Proteins: Struct., Funct., Genetics* **2003**, *52* (4), 609−623.

(5) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins: Struct., Funct., Genetics* **1999**, *37* (2), 228−241.

(6) Ewing, T. J. A.; Kuntz, I. D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.* **1997**, *18* (9), 1175−1189.

(7) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D., DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15* (5), 411−428.

(8) 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* (7), 1739−1749.

(9) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **2006**, *49* (21), 6177−6196.

(10) Jain, A. N., Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46* (4), 499−511.

(11) Pham, T. A.; Jain, A. N. Parameter estimation for scoring protein-ligand interactions using negative training data. *J. Med. Chem.* **2006**, *49* (20), 5856−5868.

(12) Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M., LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graphics Modell.* **2003**, *21* (4), 289−307.

(13) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of docking performance: Comparative data on docking algorithms. *J. Med. Chem.* **2004**, *47* (3), 558−565.

(14) 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* (20), 5851−5855.

(15) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins: Struct., Funct., Bioinformatics* **2004**, *56* (2), 235−249.

(16) 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* (20), 5912−5931.

(17) Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. Comparison of automated docking programs as virtual screening tools. *J. Med. Chem.* **2005**, *48* (4), 962−976.

(18) Bohm, H. J. The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant for a Protein Ligand Complex of Known 3-Dimensional Structure. *J. Comput.-Aided Mol. Des.* **1994**, *8* (3), 243−256.

(19) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P.; Empirical scoring functions .1. 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* (5), 425−445.

(20) Gohlke, H.; Hendlich, M.; Klebe, G., Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295* (2), 337−356.

(21) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42* (5), 791−804.

(22) Wang, R. X.; Lai, L. H.; Wang, S. M. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16* (1), 11−26.

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

(24) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **2006**, *49* (2), 534−553.

(25) *MacroModel*, *version 9.1*; Schrodinger, L.L.C.: New York, NY, 2005.

(26) 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* (7), 1750−1759.

(27) Knight, J. L.; Zhou, Z. Y.; Gallicchio, E.; Himmel, D. M.; Friesner, R. A.; Arnold, E.; Levy, R. M., Modeling maximal structural diversity in X-ray crystallographic refinement using Protein Local Optimization by torsion angle sampling. Manuscript submitted for publication.

CI7000346