Journal of Medicinal Chemistry

ARTICLE

pubs.acs.org/jmc

Docking Performance of Fragments and Druglike Compounds

Marcel L. Verdonk,*[,†] Ilenia Giangreco,[†,§] Richard J. Hall,[†] Oliver Korb,[‡] Paul N. Mortenson,[†] and Christopher W. Murray[†]

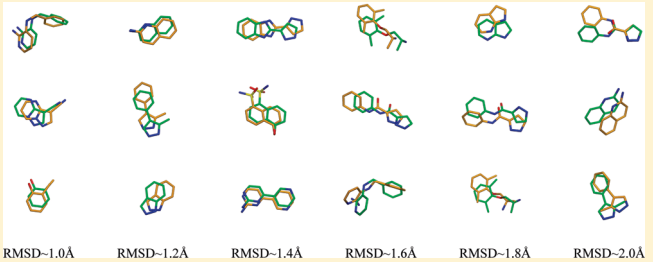
[†]Astex Therapeutics Ltd., 436 Cambridge Science Park, Milton Road, Cambridge CB4 0QA, United Kingdom

[‡]Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, United Kingdom

[§]Dipartimento Farmaco-Chimico, University of Bari, Via Orabona 4, I-70125 Bari, Italy

Supporting Information

ABSTRACT: This paper addresses two questions of key interest to researchers working with protein−ligand docking methods: (i) Why is there such a large variation in docking performance between different test sets reported in the literature? (ii) Are fragments more difficult to dock than druglike compounds? To answer these, we construct a test set of in-house X-ray structures of protein−ligand complexes from drug discovery projects, half of which contain fragment ligands, the other half druglike ligands. We find that a key factor affecting docking performance is ligand efficiency (LE). High LE compounds are significantly easier to dock than low LE compounds, which we believe could explain the differences observed between test sets reported in the literature. There is no significant difference in docking performance between fragments and druglike compounds, but the reasons why dockings fail appear to be different.



INTRODUCTION

Protein−ligand docking methodologies play a key role in structure-based drug discovery. A range of protein−ligand docking programs have been reported in the literature, and an increasing number of these tools are available to the community, e.g., DOCK,[1] GOLD,[2,3] FlexX,[4] FRED,[5] Surflex,[6] GLIDE[7,8] and ICM.[9] The performance of a docking program is normally measured by its ability to reproduce the ligand binding modes for a test set of protein−ligand complexes from the Protein Data Bank[10] (PDB). Although other measures for docking success have been proposed,[11−13] the most commonly used performance indicator is the percentage of ligands for which the top-ranked solution produced by the docking program is within a defined root-mean-square distance (RMSD) cutoff (generally 2 Å) of the experimental binding mode; we will refer to this as the "docking performance".

The most straightforward and historically most applied validation protocol is to dock each ligand against its native protein conformation; i.e., the 3D coordinates of the protein are taken from the same structure that contained the ligand; we will refer to this as "native docking". In real-life applications where the binding modes of newly designed compounds are predicted, the protein structure used to dock against will be that of a complex containing another ligand or that of the apo form of the protein; we will refer to this as "non-native docking". Finally, in ensemble docking, the ligand is docked against a number of non-native conformations of the protein. The highest scoring binding mode is then selected from the ensemble of dockings against all protein conformers.

Over the past decade, a significant number of test sets for assessing docking performance have been described in the literature, and many protein−ligand docking programs have been assessed against these sets. Some of the results of these studies are listed in Table 1. It is apparent from these data that non-native docking is a much harder problem than native docking, with docking performance approximately 20% lower for non-native docking. What is interesting, however, is that native docking performance also varies significantly between different studies (39−80%), and the same holds for non-native docking performance (26−63%). There are various reasons why this might be the case, including (i) the quality of the docking programs used, although even for the same docking program, performance varies significantly between studies; (ii) the types of targets and ligands included in the studies; (iii) the level of experience authors have with the docking software; (iv) the quality of the X-ray structures in the set (some structures may have poor electron density for the ligands, disorder, etc., in-house structures might not be fully refined); (v) the preparation of the binding sites and ligands (e.g., protonation states might be incorrect); (vi) protocol differences (e.g., site definitions can differ, some authors preoptimize complexes, etc.). One result we found particularly intriguing is that of Warren and colleagues, who obtained comparatively poor docking performance on their test set, which included only in-house GSK X-ray structures containing compounds that were synthesized to support active

Received: April 11, 2011
Published: June 21, 2011

Table 1. Recent Literature Studies into Native, Non-Native, and Ensemble Docking Performance (at 2 Å RMSD) on Test Sets of Significant Size

first author	year	structures	targets	type	docking performance (%)
Tuccinardi[14]	2010	711	≥80	native	60
Tuccinardi[14]	2010	421	22	non-native	37
Sándor[15]	2010	190	78	native	80
Sándor[15]	2010	63	8	non-native	63
Bottegoni[16]	2009	1113	99	non-native	47
Bottegoni[16]	2009	1113	99	ensemble	68
Verdonk[17]	2008	1112	65	non-native	61
Verdonk[17]	2008	1112	65	ensemble	67
Sutherland[18]	2007	246	8	native	39
Sutherland[18]	2007	246	8	non-native	26
Hartshorn[19]	2007	85	85	native	79
Warren[20]	2006	136	7	non-native	36
Friesner[7]	2004	282		native	71
Perola[21]	2004	150	63	native	61

drug discovery projects. On their test set, using GOLD, Warren only obtained a docking performance of 36%, whereas on the Astex non-native set (a carefully constructed set of drug targets containing druglike ligands from the PDB) we obtained 61%, also using GOLD. We were interested to know what causes these differences between sets and whether a similarly poor docking performance would be obtained for an Astex in-house set of X-ray structures of protein−ligand complexes, all determined to support drug-discovery projects.

Additionally, we wanted to investigate docking performance for fragment ligands vs larger, more drug-sized ligands. It has been argued that docking fragments is a particularly difficult problem. Free energy differences between different binding modes of a fragment are generally assumed to be much smaller than those of larger compounds. Given the inaccuracies inherent in current scoring functions, these smaller energy differences could make it more difficult to distinguish the correct binding mode of a fragment from proposed incorrect binding modes. Several papers discuss fragment docking, but the test sets have either been small (e.g., in the study by Kawatkar and colleagues on prostaglandin D2 synthase and DNA ligase[22] and that by Loving et al. on 12 protein-fragment complexes[23]) or the definition of what constitutes a fragment has been quite broad (like in the study by Sándor and colleagues,[15] where a relatively high molecular weight cutoff of 300 Da was used). More importantly, to the best of our knowledge, no comprehensive study has been conducted to directly compare the docking of fragments and the docking of druglike molecules.

Here, we construct a test set of in-house structures from Astex's drug discovery projects. The set comprises 206 protein−ligand complexes for 11 drug targets; 106 of these structures are complexes with fragments, and the remaining 100 structures are complexes with larger, more druglike ligands. We then use GOLD to run native, non-native and ensemble docking experiments against this test set. In addition, we apply several rescoring protocols, including minimization and scoring of each binding mode in the AMBER molecular mechanics force field.[24,25] Next, we discuss factors that may influence docking performance. Finally, we compare the docking performance of fragments to that of druglike compounds and we analyze the different reasons why dockings of fragment and dockings of druglike compounds fail.

METHODOLOGY

Test Set Construction. Protein−ligand complexes were selected from our in-house database of X-ray structures. As a surrogate for PKB we used structures of the bPKA-PKB chimera described previously.[26] For each target, only structures for which the ligand was considered either a "fragment" or "druglike" were considered. Fragment ligands were defined to contain no more than 15 heavy atoms, whereas druglike ligands had to contain at least 20 heavy atoms. In addition, fragment ligands comply with the "rule of three",[27] the only exception being that we included two β-secretase fragments that have four hydrogen bond donors. Druglike ligands comply with the "rule of five".[28] These definitions ensure that there is a clear distinction between the properties of the two sets of compounds. Next, for each set of structures (i.e., fragments and druglike compounds separately) ligands were clustered based on their Daylight fingerprint.[29] We used a single-linkage clustering algorithm, with a similarity cutoff of 0.6; i.e., after clustering, no two compounds in different clusters have a Daylight similarity greater than 0.6. Next, one complex was selected from each cluster until a maximum of 10 structures was reached or until the clusters had been exhausted. In the selection process, we ensured that there was clear electron density for the ligand and that the binding mode had been unambiguously defined. Also, we tried to include structures that we have published and deposited in the PDB. Finally, wherever possible, we selected "matching" fragments and druglike compounds, i.e., where the druglike compound was derived from the fragment hit. Using this approach, for 11 targets we could identify a sufficient number of structures for both fragments and druglike compounds; these targets are listed in Table 2. For each target, structures were carefully superimposed based on their binding site residues, using methodology we have previously described.[17]

Ligand and Protein Preparation. Ligands were prepared in the tautomer and protonation state in which we believe they bind to the target. The 3D coordinates for the ligands were generated from SMILES strings using Corina.[31] For each target, binding sites were prepared by consulting the modeler who had worked on that particular project. The modeler identified the binding site definition most commonly used by the project team, and then this template was used to copy protonation states, tautomeric states, rotamers (for asparagine, glutamine, and histidine residues), and key active site water molecules onto each of the selected protein structures in the test set. The protocol used for this was largely the same as the one used to construct the Astex non-native set.[17] Included water molecules are either highly conserved or of key importance for ligand binding (more details are given in the Supporting Information). Protonation states of protein side chains were assigned according to their pK_a values; for β-secretase, the two catalytic aspartic acid residues were both deprotonated.

Docking and Scoring. GOLD, version 5.0.1, was used for all the work described in this paper. Four different scoring functions were used to drive the dockings: Goldscore,[2,3,32] Chemscore,[33−35] ASP,[36] and ChemPLP.[37] All ligands in a target set (e.g., p38 fragments) were docked against all protein conformers in that set. The search algorithm was set to run 15 dockings, of 100 000 genetic algorithm (GA) operations each, using the "diverse

Table 2. Composition of the Fragment and Druglike Sets Used in This Work[a]

Fragment Set						
target	*N*	MW	ClogP	potency (μM)	LE	PDB code
Aurora A	9	183(23)	2.0(0.5)	110−0.86	0.39−0.72	2w1d
β-secretase	10	200(29)	1.8(0.9)	>1000−310	<0.27 to ∼0.37	2ohl, 2ohm, 2ohn
Cdk2	10	152(33)	0.9(0.7)	∼1000−62	0.36−0.59	1wcc, 2vta, 2vth, 2vtl, 2vtm
FGFR1	10	176(30)	1.2(1.0)	140−1.2	0.39−0.58	
HSP90A	10	154(17)	1.3(0.8)	∼1000−104	∼0.34−0.54	2xdk, 2xds
iNOS	10	159(23)	1.3(0.6)	∼1000−14	∼0.31−0.60	1m8d[b]
JAK2	10	150(14)	0.8(0.7)	∼1000−6.5	∼0.37−0.60	
MetAP2	10	166(16)	1.3(1.3)	∼1000−7.2	∼0.37−0.57	
p38	7	163(34)	1.7(0.7)	>3000−410	0.26−0.31	1wbo, 1w7h
PKB	10	163(35)	1.6(0.8)	∼1000−32	∼0.35−0.62	2uvx, 2uw3
urokinase	10	164(37)	1.2(1.2)	>1000−99	<0.29−0.55	2vin
Druglike Set						
target	*N*	MW	ClogP	potency (μM)	LE	PDB code
Aurora A	10	318(47)	4.0(0.7)	1.3−0.0058	0.29−0.49	2w1f
β-secretase	8	364(81)	3.7(0.9)	970−0.19	0.15−0.33	1w51, 2ohu, 2va7
Cdk2	10	362(51)	2.5(1.4)	3.2−0.0026	0.30−0.45	2vtp, 2vts
FGFR1	10	346(51)	3.0(1.2)	42−0.0013	0.30−0.40	
HSP90A	10	346(73)	3.0(0.8)	3.1−0.00040	0.31−0.57	2xab, 2xhr, 2vci[b]
iNOS	7	302(34)	2.6(1.0)	∼300−0.0084	∼0.24−0.55	
JAK2	10	342(46)	2.8(1.0)	0.55−0.0012	0.36−0.49	2w1i
MetAP2	10	344(54)	2.9(0.8)	74−0.012	0.25−0.44	
p38	8	395(82)	3.5(0.6)	∼1000−0.12	0.15−0.35	1w83
PKB	10	340(49)	3.3(0.7)	∼10−0.00050	0.30−0.49	2uw5, 2vo6, 2jdv
urokinase	7	358(53)	3.4(0.4)	260−0.020	0.23−0.44	

[a] Number of complexes (N), average molecular weight (MW) and average calculated logP (ClogP[30]) are listed. Standard deviations are given in parentheses. Also listed are potency and Ligand Efficiency (LE) ranges, and the entry codes of complexes that are available in the PDB. [b] These two structures (1m8d and 2vci) were deposited in the PDB by other researchers; all other PDB codes in this table were deposited by Astex; we used in-house structures for all the docking studies presented in this study.

solutions" option in GOLD (see below). Water molecules were allowed to spin but not toggle on and off.[38] The idea here was that the ensemble docking protocol would hopefully select the protein conformation with the best water configuration for that ligand. Also, this is the approach modelers generally would have taken, i.e., to dock against a particular configuration of water molecules (spinning but not toggling) that was considered appropriate for the chemotype of the ligand. By use of the superimposed binding sites, for each target set the binding sites were defined to include all protein atoms within 5 Å of a heavy atom in any of the ligands in that set. In previous studies we have always used a cutoff of 6 Å around a single ligand, but as we are using about 10 superimposed ligands here, a cutoff of 5 Å is more appropriate. Each docking solution was rescored in the three scoring functions not used for docking (i.e., Chemscore, ASP, and ChemPLP if Goldscore was used to drive the docking), allowing the ligand to relax in the rescore function. In addition, all solutions were minimized and rescored using AMBER (see below). All docking runs were repeated three times.

We also scored the X-ray binding mode of each ligand against each protein conformer. To do this, we used the superimposed set of protein−ligand complexes. Each ligand (in its X-ray geometry) was placed into the binding site of each protein conformer (native and non-native), and the local optimization mode[35] in GOLD was used to optimize the terminal protein and ligand OH and NH_3 groups (the Goldscore function was used for this local optimization). Each of these optimized complexes was then further relaxed in each of the four scoring functions, using the SIMPLEX algorithm in GOLD. In addition, each of the optimized complexes was minimized in AMBER using the protocol described below.

Diverse Solutions. A standard GOLD job consists of a number of docking runs of a ligand against a target. Until recently, each of these docking runs were carried out completely independently such that often the same (or a very similar) solution was generated repeatedly. To ensure a wider range of binding modes, we introduced the "Diverse Solutions" option into GOLD. When this option is switched on, GOLD keeps track of the solutions produced by each of the preceding docking runs. During each docking run, for each new solution produced by the GA, the RMSD is calculated against the solutions produced by previous docking runs. Two parameters then determine whether a solution is accepted into the pool of GA solutions. The first parameter, divsol_rmsd, defines an RMSD threshold that determines whether the new solution belongs in the same cluster as an existing solution. The second parameter, divsol_cluster_size, determines how many solutions GOLD will allow in each cluster. If the newly produced solution would belong in a cluster that is already full, it is discarded. Here we ran 15 dockings, with divsol_rmsd = 1.5 Å and divsol_cluster_size = 3, which generally

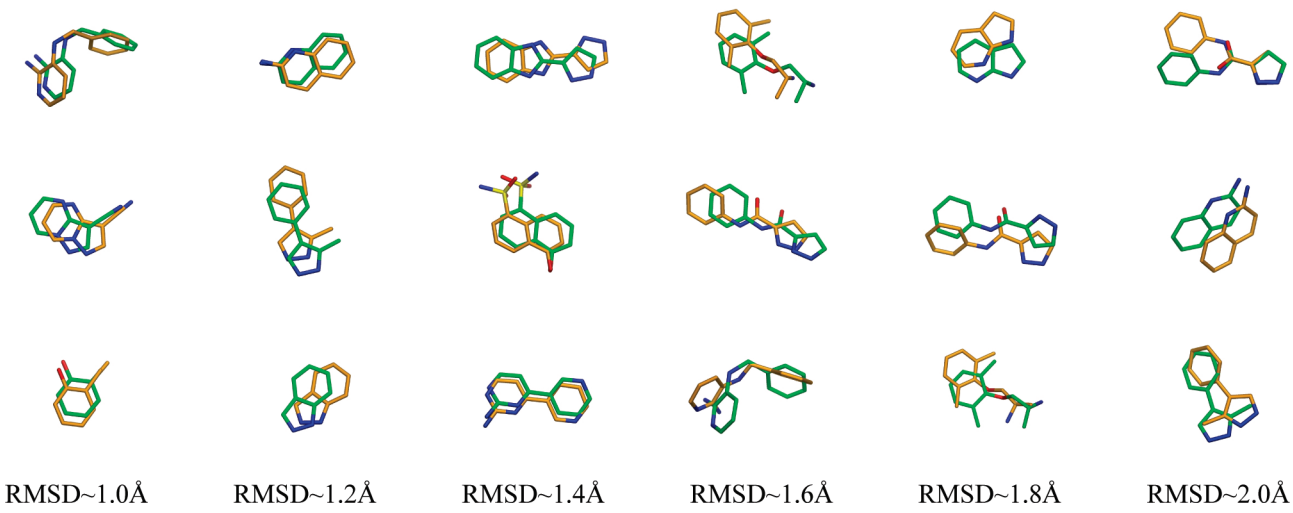


Figure 1. Examples of docking solutions of fragments with RMSD values varying between 1.0 and 2.0 Å.

results in five different clusters, each containing three solutions. The Diverse Solutions option is not designed to improve docking performance directly but is particularly useful when the solutions are passed on to a rescoring protocol. Also, in real-life docking applications, a lot of the analysis and interpretation of docking results is still done visually by project modelers and we have found it beneficial to have multiple diverse solutions to consider.

AMBER. In addition to rescoring the docked poses using the scoring functions available in GOLD, we also implemented a force field based rescoring scheme using AMBER 9. Full details of the methodology we used are provided in the Supporting Information, and so only a brief summary is presented here. For each docked complex, first the positions of hydrogen atoms were optimized using a short energy minimization with heavy atoms held fixed. A second energy minimization was then performed, with the ligand and any residues within 10 Å of a ligand heavy atom allowed to move. Both minimizations were performed using a distance-dependent dielectric constant and a cutoff of 12 Å for nonbonded interactions. Finally an interaction energy was calculated for the optimized complex, using a generalized Born solvation model[39] and no cutoff for nonbonded interactions.

Docking Performance Assessment. For native docking the RMSD between the top-ranked solution and the X-ray binding mode can be computed directly from the docked and X-ray coordinates of the ligand. For non-native and ensemble docking, the binding site of the non-native protein conformer (together with the docked ligand) needs to be superimposed on the native protein−ligand complex in order to calculate the ligand RMSD and hence the docking performance; this was done using a procedure we have described previously.[17]

Docking performance was defined as the percentage of complexes in a set (e.g., "fragment" or "druglike") for which the RMSD between the top-ranked solution and the X-ray binding mode is below a certain cutoff. In the literature, an RMSD cutoff of 2.0 Å is used widely to assess docking performance, and for druglike compounds this generally is appropriate. However, in our experience an RMSD cutoff of 2.0 Å is too high for fragments. Between 1.5 and 2.0 Å RMSD, fragment dockings frequently contain significant errors, e.g., where key interactions are missed or the overall orientation of the ligand is significantly different from that in the X-ray structure (see Figure 1). Having inspected hundreds of the fragment dockings presented in this study, we believe an RMSD cutoff of 1.5 Å is much more appropriate for fragments and is in line with a 2.0 Å cutoff for druglike compounds. Hence, in this work we have used two different RMSD cutoff values to define docking performance: 2.0 Å for druglike compounds and 1.5 Å for fragments. Results for both thresholds and both sets are presented in the Supporting Information.

For each ligand i, we also defined a combined success rate, $p(i)$, in which the results from all docking/scoring protocols and all repeats are taken into account:

$$p(i) = n_{\text{correct}}(i)/n_{\text{total}}(i)$$

where $n_{\text{total}}(i)$ is the total number of docking/scoring protocols and repeats available for ligand i and $n_{\text{correct}}(i)$ is the number of these cases for which the RMSD of the top-scoring docking solution is within the defined cutoff value (i.e., 2.0 Å for druglike compounds and 1.5 Å for fragments). $n_{\text{total}}(i)$ can be written as

$$n_{\text{total}}(i) = n_{\text{df}} n_{\text{sf}} n_{\text{repeats}} n_{\text{conf}}(i)$$

where n_{df} is the number of scoring functions used to drive the dockings (i.e., 4), n_{sf} is the number of scoring functions used to score each docked solution (i.e., 5), and n_{repeats} is the number of times each docking run was repeated (i.e., 3). $n_{\text{conf}}(i)$ is the number of protein conformers considered for ligand i. For native and ensemble docking $n_{\text{conf}}(i) = 1$ (only the native protein conformer and the best scoring non-native protein conformer are considered, respectively). For non-native docking $n_{\text{conf}}(i)$ is the number of non-native protein conformers available (9 in most cases).

The idea is that $p(i)$ reflects how straightforward or difficult it is to predict the binding mode for ligand i when a range of different docking/scoring protocols are considered. In an analogous manner, for a particular set of ligands j, we can define the combined docking performance for this set as

$$P(j) = \frac{\sum_i p(i)}{N_{\text{total}}(j)}$$

where the summation is over the ligands in set j and $N_{\text{total}}(j)$ is the total number of ligands the set.

dx.doi.org/10.1021/jm200558u |J. Med. Chem. 2011, 54, 5422–5431

RESULTS AND DISCUSSION

Test Set Composition. Table 2 lists the targets included in the test set constructed here. On average, the MW of the fragments is about 150 Da lower than that of the druglike compounds. Because a diversity criterion was used during the selection of compounds, the resulting sets contain ligands with a wide potency and ligand efficiency (LE) range. Also, the diversity of the compounds, combined with the nature of some of the targets, makes these test sets tough but realistic for structure-based drug discovery projects. For example, several of the kinase compounds bind away from the so-called hinge region. We have included ligands that induce different conformations of the "flap" in β-secretase. Both "DFG-in" and "DFG-out" conformations were included for p38, and for HSP90A examples of the "collapsed" and "un-collapsed" helix around Gly108 were included.

Interestingly, although it was outside the scope of this work (we specifically wanted a set of structures from in-house projects), we believe it would have been impossible to construct a test set like this from the PDB, particularly when the same quality standards are applied. To illustrate this, we applied the same filters (drug/fragment-likeness, structure factors deposited, ligand diversity, no clashes or symmetry contacts) on all PDB structures for the 85 targets in the Astex diverse set. For only two of these targets (Cdk2 and carbonic anhydrase II) are there at least five complexes containing fragment ligands and five complexes containing druglike ligands in the PDB.

Overall Docking Results. Docking performance for fragments and druglike compounds is given in Table 3. Overall, docking performance is relatively poor compared to the other test sets we have reported docking validations on (see below). As we and others have seen before, there is a significant drop-off from native docking to non-native docking performance of roughly 20−30% depending on the scoring functions used. Some of this performance loss (about 5−15%) can be regained by using an ensemble docking protocol. Compared to the other three scoring functions, ChemPLP performs consistently well, giving the best direct docking performance (i.e., without rescoring) for native, non-native and ensemble docking for both fragments and druglike compounds. Rescoring the docking solutions with a second scoring function gives a slight improvement in docking performance in some but not all cases.

AMBER Results. Because the docking performance on these sets is relatively poor, we decided to investigate whether a force-field based scoring protocol would improve the results. Apart from a more sophisticated description of the interactions, the energy minimization also allows the protein to relax in the presence of the ligands and should hopefully deal with small induced fit effects. Rescoring docking solutions using force-field-based methods is not a new concept. Already in 1999, Hoffmann and colleagues minimized docking solutions produced by FlexX in the CHARMM force field and observed a significant improvement in binding mode prediction.[40] Recently, and specifically for fragment docking, Gleeson and Gleeson observed that rescoring docking solutions using QM/MM force fields improved binding mode predictions significantly for a small set of structures of fragments bound to kinase targets.[41]

The results for the AMBER rescoring protocol applied on our in-house set are shown in Table 3. Overall the results are encouraging; on a naive basis, for the 24 docking protocols used (native, non-native, and ensemble docking for both fragment and druglike sets, with 4 scoring functions), AMBER is the best

Table 3. Docking Performance for the Fragment and Druglike Sets[a]

Fragment Set					
	rescoring				
native docking	Goldscore	Chemscore	ChemPLP	ASP	AMBER
Goldscore	56.3(2.9)	65.1(2.8)	66.7(1.4)	50.3(2.4)	***69.5(2.0)***
Chemscore	66.0(0.9)	61.6(0.5)	68.9(0.0)	48.7(2.4)	67.9(1.6)
ChemPLP	62.6(2.2)	60.7(1.1)	***63.5(2.0)***	51.9(0.9)	69.2(2.0)
ASP	58.5(2.5)	56.9(2.4)	63.5(1.4)	45.3(1.6)	62.6(0.5)
	rescoring				
non-native docking	Goldscore	Chemscore	ChemPLP	ASP	AMBER
Goldscore	31.6(0.4)	41.1(0.6)	39.2(0.4)	36.5(1.4)	41.0(0.3)
Chemscore	36.1(0.9)	35.2(0.9)	38.0(0.7)	34.9(0.5)	***42.6(0.3)***
ChemPLP	33.8(0.4)	37.6(0.2)	***35.6(0.2)***	33.2(0.2)	41.2(0.3)
ASP	33.9(0.4)	37.2(1.5)	38.3(1.0)	32.3(0.4)	40.2(0.1)
	rescoring				
ensemble docking	Goldscore	Chemscore	ChemPLP	ASP	AMBER
Goldscore	36.8(1.9)	47.2(0.9)	42.6(2.3)	38.4(3.7)	51.9(3.4)
Chemscore	40.7(3.4)	38.9(2.0)	40.6(0.9)	38.7(2.1)	***58.5(1.9)***
ChemPLP	37.7(3.5)	42.6(2.6)	***41.2(0.8)***	34.3(0.5)	53.1(0.5)
ASP	39.5(3.6)	40.7(2.6)	40.6(0.8)	33.7(2.1)	50.9(1.9)
Druglike Set					
	rescoring				
native docking	Goldscore	Chemscore	ChemPLP	ASP	AMBER
Goldscore	58.3(1.2)	69.0(1.0)	68.3(0.6)	66.0(1.0)	66.7(0.6)
Chemscore	61.0(3.5)	59.7(0.6)	68.0(1.7)	61.7(2.3)	69.3(4.9)
ChemPLP	65.7(0.6)	65.0(1.0)	***70.3(1.5)***	66.3(0.6)	***75.7(2.5)***
ASP	60.3(2.5)	60.0(4.6)	64.0(2.0)	54.0(1.0)	67.3(4.6)
	rescoring				
non-native docking	Goldscore	Chemscore	ChemPLP	ASP	AMBER
Goldscore	26.3(0.4)	***35.7(0.4)***	34.2(0.1)	31.2(0.3)	34.7(0.7)
Chemscore	27.6(0.5)	29.5(0.1)	29.5(0.1)	26.6(0.3)	33.1(0.5)
ChemPLP	26.8(0.6)	29.9(0.7)	***31.6(0.3)***	28.5(0.8)	33.8(1.1)
ASP	27.7(0.8)	30.7(0.6)	31.6(0.4)	27.3(0.5)	31.7(0.6)
	rescoring				
ensemble docking	Goldscore	Chemscore	ChemPLP	ASP	AMBER
Goldscore	28.8(2.5)	47.0(4.4)	***56.0(1.7)***	35.9(1.6)	47.3(2.9)
Chemscore	34.9(1.5)	43.3(0.6)	47.7(2.3)	31.7(2.9)	52.3(3.2)
ChemPLP	36.3(1.5)	42.5(3.1)	***49.0(1.7)***	34.0(3.6)	48.3(2.9)
ASP	38.3(2.9)	42.5(6.1)	50.5(4.8)	35.2(5.3)	46.7(0.6)

[a] Results are shown for native, non-native, and ensemble docking and for all combinations of docking/scoring functions. An RMSD cutoff value of 2.0 Å is used to define docking performance for druglike compounds, whereas a cutoff of 1.5 Å is used for fragments. Standard deviations over the three repeats are shown in parentheses. The best-performing docking protocol and the best performing rescoring protocol are boldface in each section.

Table 4. Combined Docking Performance for Subsets of the Astex In-House Set[a]

	N	native docking		non-native docking		ensemble docking	
fragment	106	61%	$p = 0.39$	37%	$p = 0.051$	42%	$p = 1.0$
druglike	100	65%		30%		42%	
low potency[b]	94	61%	$p = 0.27$	37%	$p = 0.33$	46%	$p = 0.84$
high potency[b]	83	66%		33%		45%	
low LE[c]	87	55%	$p = 0.00022$	26%	$p = 0.0000048$	35%	$p = 0.0000062$
high LE[c]	86	73%		45%		57%	

[a] Results are shown for native, non-native, and ensemble docking, and estimated significance levels are given for differences between subsets. [b] Low-potency compounds have $IC_{50} > 10\ \mu M$. High-potency compounds have $IC_{50} < 10\ \mu M$. [c] Low-LE compounds have LE < 0.40. High LE compounds have LE > 0.40.

performing rescoring protocol on 16 occasions. The results look better for the fragment set in isolation, with the AMBER protocol performing best for 9 out of the 12 docking experiments. Particularly encouraging is the performance for ensemble docking with the fragment set, where the AMBER results are substantially better than any other protocol. The results for the druglike set are more equivocal, and from Table 3 it is hard to argue that for this set rescoring with AMBER is the best protocol. However, even for this set its performance appears similar to that of the best performing traditional scoring function (ChemPLP).

It is not immediately apparent why the AMBER rescoring protocol should be more successful for the fragment set than the druglike set. One possibility is that (as commonly hypothesized) fragment docking performance is more dependent on the quality of the scoring function used (although this does of course presuppose that this AMBER protocol is a better scoring function than the others tested here). Another possibility is that the parametrization of the force field (and AM1-BCC charges) is better for smaller molecules than larger ones. The fragments in general will be less flexible than the druglike compounds; so perhaps the conformational energetics of the druglike ligands are being poorly modeled. Also, docking solutions for smaller compounds may provide better starting points for AMBER energy minimization because the energy landscape is likely to be less rugged than that for docking solutions for larger, more complex molecules.

Fragments vs Druglike Compounds. The hypothesis that fragments may be harder to dock than druglike compounds is not supported by the results shown in Table 3. If the appropriate cutoff values are used (i.e., 2.0 Å for druglike compounds and 1.5 Å for fragments), the docking performance is generally very similar between fragments and druglike compounds. If we apply the commonly used threshold of 2.0 Å to both sets, fragments actually appear easier to dock. However, as discussed previously, we do not believe that this is an appropriate analysis.

A much more comprehensive way to compare docking performance of fragments and druglike compounds is to calculate the combined docking performance for both sets (see Table 4). It is clear that when all docking and scoring protocols are considered, the docking performance for fragments is not lower than that for druglike compounds. For non-native docking, the difference in performance is close to being statistically significant (in favor of docking fragments), but for native docking and ensemble docking (the latter being the most relevant benchmark) there is no significant difference.

We wanted to understand to what extent docking failures for fragments and druglike compounds are due to inadequate sampling or due to incorrect scoring. Hence, in order to simulate the situation where the docking program always generates the correct binding mode as one of its solutions, we added the X-ray binding modes (see Methodology) to the list of possible docking solutions for native, non-native, and ensemble docking. In addition, for ensemble docking, we added the dockings against the native protein conformation; this simulates the situation where the docking program always has the correct protein conformer available in the ensemble of protein conformers it docks the ligands against. Figure 2 shows the effect on combined docking performance for fragments and druglike compounds, when these extra solutions are added. What is immediately clear from this plot is that the effect of adding "correct" solutions to the list of binding modes is larger for druglike compounds than it is for fragments. For example, when the X-ray binding modes of the ligands are added (horizontally striped bars), the improvement in

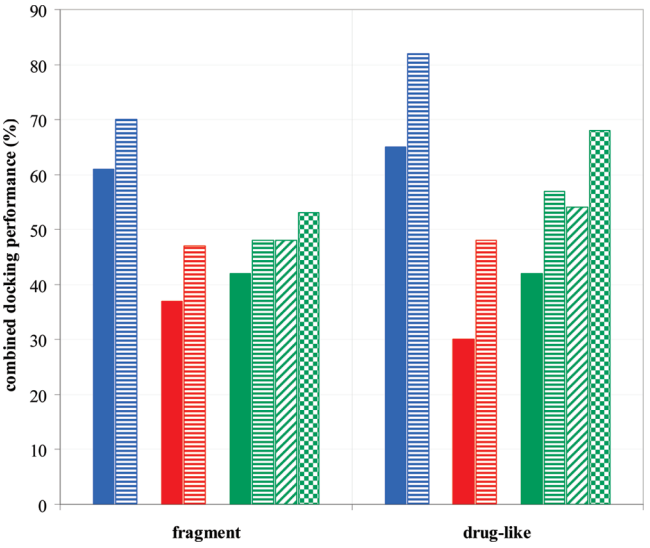


Figure 2. Effect of including information of the experimental binding modes on the combined docking performance for fragments and druglike compounds. Results are shown for native (blue), non-native (red), and ensemble docking (green). Standard combined docking performance is represented by the solid bars (these are the results shown in Table 4). Horizontally striped bars represent the experiments where the X-ray binding modes of the ligands were added to the list of docking solutions. For ensemble docking, results are also shown for which the dockings against the native protein conformer were included but the X-ray binding modes of the ligands were excluded (diagonally striped bars) and for which all X-ray binding modes and docking solutions against all protein conformers (including the native one) were included (checkered bars).

dx.doi.org/10.1021/jm200558u |J. Med. Chem. 2011, 54, 5422–5431

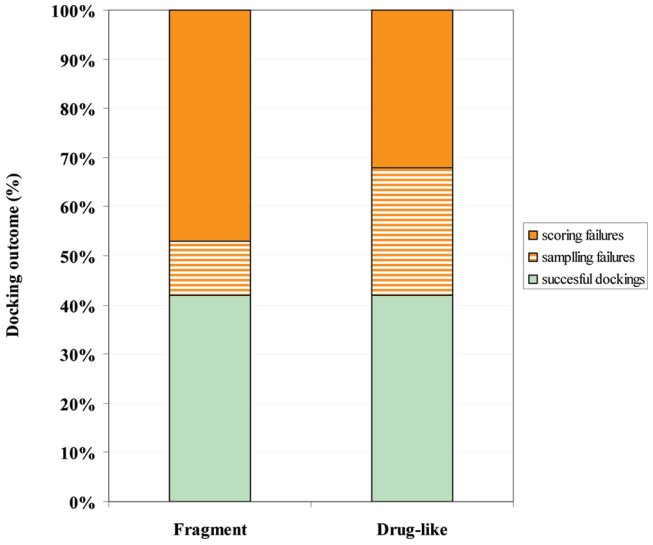


Figure 3. Docking outcome distributions for ensemble docking of fragments and druglike compounds. "Successful dockings" corresponds to the combined docking performance for standard ensemble docking. "Sampling failures" represents the percentage of cases for which a docking can be "rescued" either by adding docking solutions against the native protein conformation or by adding the X-ray binding mode of the ligand to the list of docking solutions. "Scoring failures" represents all remaining cases, i.e., cases for which all correct solutions (including the X-ray structure of the native protein−ligand complex) score worse than an incorrect solution.

combined docking performance for druglike compounds is 17%, 18%, and 15% for native, non-native, and ensemble docking, respectively. For fragments, the corresponding improvements are only 9%, 10%, and 6%, respectively. Similarly, when the native protein conformation is included for ensemble docking, the combined docking performance improvement is 12% for druglike compounds vs 6% for fragments. It is interesting to note that for ensemble docking, adding the native protein conformer does not recover docking performance to the performance obtained for native docking because of the noise introduced by the scores obtained against the other protein conformers.

In Figure 3, docking outcomes for ensemble docking are categorized as successful dockings, sampling failures, and scoring failures, based on the data presented in Figure 2 (see Verkhivker et al.[42] and Mukherjee et al.[43] for similar analyses). For fragments, the vast majority of docking failures are a result of poor scoring; i.e., even when correct solutions are generated, they are often not scored better than incorrect ones (see Figure 4 for an example). For druglike compounds, however, a significant proportion of docking failures is caused by insufficient sampling (see Figure 5 for an example). In other words, although in practice we observe no significant difference in docking performance between fragments and druglike compounds (see Table 4), in the theoretical case where the docking program always generates the correct binding mode as one of its solutions (chequered bars in Figure 2), fragment docking is significantly more difficult than docking druglike molecules. This is consistent with the assumption that energy gaps (and score differences) between different binding modes are smaller for fragments than they are for druglike compounds. The fact that accurate scoring is the key problem for fragment docking might also explain why rescoring with a more sophisticated scoring function like AMBER has a much larger impact on the docking performance of fragments than it does for druglike compounds (see above).

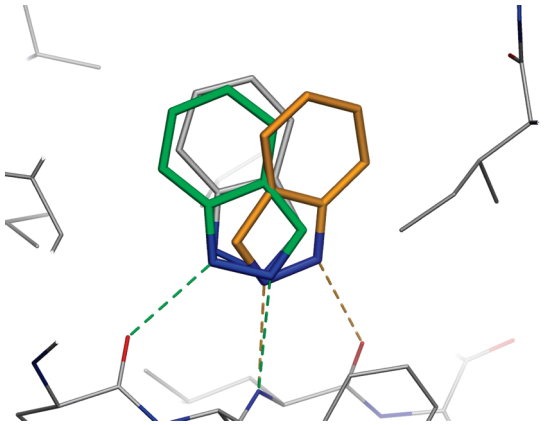


Figure 4. Example of a typical docking failure for fragments. This is a native docking example against Cdk2, where Goldscore was used to generate the binding modes and ChemPLP was used to score them. The X-ray structure is shown in gray. The top-scoring binding mode (orange) is incorrect (RMSD = 2.46 Å, Score = 36.1). A solution very close to the experimental binding mode is generated by the docking algorithm (green, RMSD = 0.60 Å), but it scores slightly worse (Score = 35.2).

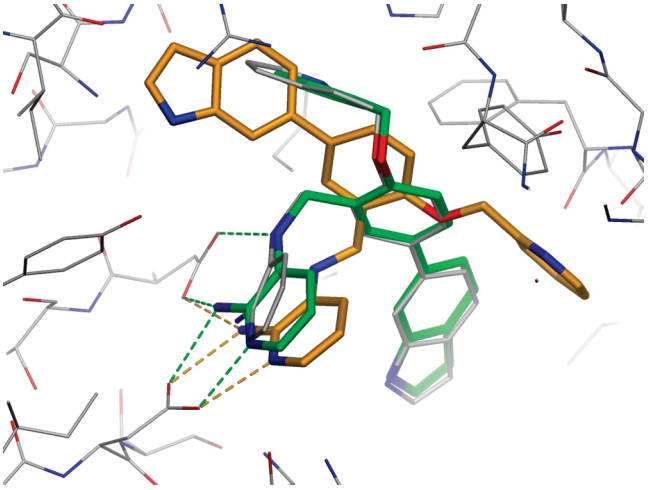


Figure 5. Example of a sampling failure for druglike molecules. This is a native docking example against β-secretase, where Goldscore was used both to generate and to score the binding modes. The X-ray structure is shown in gray. The top-scoring solution shown in orange is incorrect (RMSD = 7.29 Å, Score = 56.2). Although the placement of the 2-aminopyridine moiety is mostly correct, the rest of the molecule has not been docked correctly. The docking algorithm does not produce any solutions that are close to the experimental binding mode. However, when the X-ray binding mode is relaxed and scored in the Goldscore function (shown in green), it scores significantly better than the top-ranked docking solution (RMSD = 0.85 Å, Score = 66.7).

It is worth pointing out here that there are several factors contributing to what are considered "sampling failures" in Figure 3. The most obvious cause of a sampling failure is when the docking program fails to sample solutions close to the experimental binding mode (for example, because it does not vary a particular internal degree of freedom of the ligand). In addition, the 3D ligand geometries that were used for the docking experiments were generated using Corina, and it is possible that this could

prevent the experimental binding mode of the ligand to be generated (it is difficult to decide where ligand geometry sampling becomes the responsibility of the docking program). Finally, sampling failures can be caused by not including a protein conformation in the ensemble that is close enough to the native protein conformer. From the ensemble docking results in Figure 2, it appears that about half of the sampling failures are due to poor ligand geometry sampling (either because the docking program fails to sample the correct conformer or because the Corina geometry prevents sampling of the bioactive ligand conformer); the rest of the sampling failures are due to inadequate protein conformer sampling (i.e., because no protein conformer was included that was close enough to the native protein conformer).

The results in Figures 2 and 3 also indicate that if sampling could be improved for druglike compounds, docking performance could be improved significantly. Interestingly, in 92% of the ensemble dockings of druglike compounds, a solution with RMSD < 2.0 Å is generated by the docking program. However, this appears not to be close enough to the experimental binding mode for these solutions to be top scoring. Although outside the scope of this study, it would be interesting to investigate (for the ligand sampling failures in Figure 3) how close a solution needs to be to experiment in order for it to become the top scoring solution.

Potency and Ligand Efficiency. It is clear that the docking performance obtained for both the fragment and the druglike set is much lower than the docking performance we have previously reported on the Astex diverse set[19] and the Astex non-native set[17] (see Table 1). These differences cannot be attributed to the quality of the structures, the experience of the researchers, the preparation of ligand and target structures, the docking programs, or the docking protocols used, as all these factors are largely unchanged between these studies. Instead, the difference in docking performance has to be related to the types of targets and ligands that were included. The comparison between fragments and druglike compounds (see above) showed that molecular size is not a key parameter affecting docking performance. One possible hypothesis might be that more potent compounds are easier to dock than weaker binding compounds. Hence, we split our complete set (fragment and druglike combined) into a high-affinity set and a low affinity set and calculated the combined docking performance for both sets (see Table 4). It is clear that for native, non-native, and ensemble docking there is no significant difference in performance between high-affinity and low-affinity ligands.

However, when the complete set is split into a low ligand efficiency (LE)[44,45] and a high LE set, then there is a clear difference in combined docking performance: for native, non-native, and ensemble docking, the combined docking performance is significantly higher for high-LE complexes. Intuitively, this appears to make sense, as the high LE compounds are likely to form both a greater number and higher quality of interactions (relative to their size) and may therefore be more straightforward to dock correctly (both in terms of scoring and sampling) than low LE compounds. One might suspect that the performance difference observed between the high LE set and the low LE set is driven largely by the types of targets in the sets. However, when the three targets with the lowest average LE (β-secretase, p38, and urokinase) are removed from the analysis, the results are essentially the same (results not shown).

We also reanalyzed the docking performance for the Astex non-native set in terms of the LE of the compounds in that set. To do this, we used the potency values we extracted from the literature for the entries in the set.[19] We did not rerun all dockings using the protocols described here but instead used the native and non-native docking results obtained previously with Goldscore.[17] When the entries in the Astex non-native set are split into a high-LE set and a low-LE set, a similar difference in docking performance is observed: for native docking the docking performance is 73% for the low-LE set vs 89% for the high-LE set, and for non-native docking the docking performance is 47% for the low-LE set vs 68% for the high-LE set. Unfortunately, however, because of the small sample sizes, these performance differences are not statistically significant to the same level as we observe for the in-house set described in the present paper (p = 0.12 and p = 0.10 for native and non-native docking, respectively).

If we compare the subsets of entries in the Astex diverse set and the current Astex in-house set for which LE data are available, the average LE values of the two sets differ only marginally (0.43 for the Astex diverse set vs 0.40 for the Astex in-house set). However, the Astex diverse set contains a much higher fraction of very high LE (>0.60) complexes than the Astex in-house set (12% vs 1.5%), and the docking performance for these complexes in the Astex diverse/non-native set is high (100% for native, 86% for non-native docking). Also, for a significant fraction (29 complexes) of the Astex in-house set, no suitable potency data are available to derive a LE value (these were also left out of the calculation of the average LE value). Most of these 29 complexes will in reality have a low LE, and the combined docking performance is poor for these examples (59%, 7%, and 14% for native, non-native, and ensemble docking, respectively). Hence, although it is impossible to conclude categorically that LE differences are responsible for the difference in docking performance observed between the Astex diverse/non-native set and the Astex in-house set, the results presented here strongly suggest that LE plays an important role.

CONCLUSIONS

The construction of a test set of X-ray structures of protein−ligand complexes from Astex in-house drug discovery projects was described. Half of the complexes in this set contain fragment ligands, and the other half contains druglike ligands. The docking performances obtained in this study, although in-line with the results Warren et al. obtained for their in-house set,[20] are poor; even using the AMBER force field to optimize and rescore binding modes, the best docking performance for ensemble docking is only 59% for fragments and 52% for druglike compounds. We believe that in the setting of a drug discovery project, the docking performance is normally significantly higher because modelers will utilize any structural data available on the system they are working on in order to improve the docking. There are various ways of using structural information about target and ligands within a docking program, including hydrogen-bond constraints, pharmacophore constraints, adding ligand similarity overlap to the docking score, interaction fingerprints, etc., all of which should improve docking performance.

We have used an Astex in-house set of protein−ligand complexes to investigate which parameters affect docking performance. Interestingly, no correlation was observed between docking performance and the potency of the ligands in the set. However, docking performance for high ligand efficiency compounds was significantly higher than that for low LE compounds. We believe the reason behind this could be that high LE compounds form high-quality interactions with the target, which should make it easier for a docking program (both from a scoring

and searching perspective) to dock these compounds correctly. LE differences could also go some way toward explaining the differences in docking performance between different test sets. It seems likely that carefully constructed test sets based entirely on high-quality structures from the PDB (with the ligands often being drug molecules or close analogues) would contain more high LE complexes than test sets that consist of protein ligand complexes determined to generate structure−activity data for drug discovery projects.

Although it is outside the scope of the current study, it would be interesting to investigate whether LE has a similar effect on enrichments obtained in docking-based virtual screening. If high-LE compounds are docked better in a virtual screen, then one might also expect them to have better scores (and therefore rank closer to the top) than low-LE compounds.

Our study also revealed no significant difference in docking performance between fragments and druglike compounds. However, an analysis of the types of docking failures highlights an interesting difference between docking fragments and docking druglike compounds. For fragments the main problem is that the scoring functions used are often unable to distinguish the correct binding mode from incorrect ones. For druglike compounds, however, in a significant proportion of cases, the docking program does not generate a solution close enough to the X-ray binding mode for it to be top-scoring. Therefore, whereas for druglike molecules significant improvement in docking performance can be expected from improved sampling alone, for fragments essentially any performance improvement will need to come from improvements in the scoring functions. The most likely methods for achieving improved scoring for fragments are probably force-field-based scoring protocols and other more advanced methods for evaluating protein−ligand interaction energetics. The simple AMBER rescoring protocol we applied here shows that such methods do have the scope to improve docking performance for fragments.

■ ASSOCIATED CONTENT

Supporting Information. (1) Description of the methodology used to deal with water molecules during active site preparation, (2) description of the AMBER protocols used, and (3) table of docking performance for fragments and druglike compounds at 1.5 and 2.0 Å RMSD cutoffs. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: 44 1223 22 62 06. Fax: 44 1223 22 62 01. E-mail: m.verdonk@astex-therapeutics.com.

■ ACKNOWLEDGMENT

The authors thank the protein crystallographers and assay biologists at Astex for providing us with the X-ray structures and potency data used in this study and in particular Valerio Berdini and Gianni Chessari for providing us with the binding site definitions used in this work. The authors also thank the *Journal of Medicinal Chemistry* for granting an exception waiver for the deposition of X-ray structural data for this paper.

■ ABBREVIATIONS USED

LE, ligand efficiency; PDB, Protein Data Bank; RMSD, root-mean-square distance

■ REFERENCES

(1) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule−ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.

(2) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43–53.

(3) 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.

(4) 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.

(5) McGann, M. FRED pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2011**, *51*, 578–596.

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

(7) 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.

(8) 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*, 6177–6196.

(9) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488–506.

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

(11) Kroemer, R. T.; Vulpetti, A.; McDonald, J. J.; Rohrer, D. C.; Trosset, J. Y.; Giordanetto, F.; Cotesta, S.; Mcmartin, C.; Kihlen, M.; Stouten, P. F. Assessment of docking poses: interactions-based accuracy classification (IBAC) versus crystal structure deviations. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 871–881.

(12) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.

(13) Yusuf, D.; Davis, A. M.; Kleywegt, G. J.; Schmitt, S. An alternative method for the evaluation of docking performance: RSR vs RMSD. *J. Chem. Inf. Model.* **2008**, *48*, 1411–1422.

(14) Tuccinardi, T.; Botta, M.; Giordano, A.; Martinelli, A. Protein kinases: docking and homology modeling reliability. *J. Chem. Inf. Model.* **2010**, *50*, 1432–1441.

(15) Sandor, M.; Kiss, R.; Keseru, G. M. Virtual fragment docking by glide: a validation study on 190 protein−fragment complexes. *J. Chem. Inf. Model.* **2010**, *50*, 1165–1172.

(16) Bottegoni, G.; Kufareva, I.; Totrov, M.; Abagyan, R. Four-dimensional docking: a fast and accurate account of discrete receptor flexibility in ligand docking. *J. Med. Chem.* **2009**, *52*, 397–406.

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

(18) Sutherland, J. J.; Nandigam, R. K.; Erickson, J. A.; Vieth, M. Lessons in molecular recognition. 2. Assessing and improving cross-docking accuracy. *J. Chem. Inf. Model.* **2007**, *47*, 2293–2302.

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

(20) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.;

Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.

(21) 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.

(22) Kawatkar, S.; Wang, H.; Czerminski, R.; Joseph-McCarthy, D. Virtual fragment screening: an exploration of various docking and scoring protocols for fragments using Glide. *J. Comput.-Aided Mol. Des* **2009**, *23*, 527–539.

(23) Loving, K.; Salam, N. K.; Sherman, W. Energetic analysis of fragment docking and application to structure-based pharmacophore hypothesis generation. *J. Comput.-Aided Mol. Des* **2009**, *23*, 541–554.

(24) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(25) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merk, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*; AMBER: San Francisco, CA, 2006.

(26) Davies, T. G.; Verdonk, M. L.; Graham, B.; Saalau-Bethell, S.; Hamlett, C. C.; McHardy, T.; Collins, I.; Garrett, M. D.; Workman, P.; Woodhead, S. J.; Jhoti, H.; Barford, D. A structural comparison of inhibitor binding to PKB, PKA and PKA-PKB chimera. *J. Mol. Biol.* **2007**, *367*, 882–894.

(27) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A "rule of three" for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8*, 876–877

(28) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.

(29) *Daylight*; Daylight Chemical Information Systems Inc.: Aliso Vieja, CA, 2006; www.daylight.com.

(30) Software from BioByte Corp., 201 W. 4th Street, No. 204, Claremont, CA 91711-4707; 2006.

(31) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537–547.

(32) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Further Development of a Genetic Algorithm for Ligand Docking and Its Application to Screening Combinatorial Libraries. In *Rational Drug Design: Novel Methodology and Practical Applications*; Parrill, A. L., Reddy, M. R., Eds.; American Chemical Society: Washington, DC, 1999; pp 271−291.

(33) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins* **1998**, *33*, 367–382.

(34) 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*, 425–445.

(35) Verdonk, M. L.; Cole, J. C.; Hartshorn, M.; Murray, C. W.; Taylor, R. D. Improved protein−ligand docking using GOLD. *Proteins* **2003**, *52*, 609–623.

(36) Mooij, W. T. M.; Verdonk, M. L. General and Targeted Statistical Potentials for Protein−Ligand Interactions, 2005. Unpublished.

(37) Korb, O.; Stützle, T.; Exner, T. E. Empirical scoring functions for advanced protein−ligand docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49*, 84–96.

(38) Verdonk, M. L.; Chessari, G.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Nissink, J. W.; Taylor, R. D.; Taylor, R. Modeling water molecules in protein−ligand docking using GOLD. *J. Med. Chem.* **2005**, *48*, 6504–6515.

(39) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* **2004**, *55*, 383–394.

(40) Hoffmann, D.; Kramer, B.; Washio, T.; Steinmetzer, T.; Rarey, M.; Lengauer, T. Two-stage method for protein−ligand docking. *J. Med. Chem.* **1999**, *42*, 4422–4433.

(41) Gleeson, M. P.; Gleeson, D. QM/MM as a tool in fragment based drug discovery. A cross-docking, rescoring study of kinase inhibitors. *J. Chem. Inf. Model.* **2009**, *49*, 1437–1448.

(42) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering common failures in molecular docking of ligand−protein complexes. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 731–751.

(43) Mukherjee, S.; Balius, T. E.; Rizzo, R. C. Docking validation resources: protein family and ligand flexibility experiments. *J. Chem. Inf. Model.* **2010**, *50*, 1986–2000.

(44) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9997–10002.

(45) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004**, *9*, 430–431.