Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening

Thomas A. Halgren,*,† Robert B. Murphy,† Richard A. Friesner,‡ Hege S. Beard,† Leah L. Frye,§ W. Thomas Pollard,† and Jay L. Banks†

Schrödinger, L.L.C., 120 W. 45th Street, New York, New York 10036, Department of Chemistry, Columbia University, New York, New York 10036, and Schrödinger, L.L.C., 1500 SW First Avenue, Portland, Oregon 97201

Received December 24, 2003

Glide's ability to identify active compounds in a database screen is characterized by applying Glide to a diverse set of nine protein receptors. In many cases, two, or even three, protein sites are employed to probe the sensitivity of the results to the site geometry. To make the database screens as realistic as possible, the screens use sets of "druglike" decoy ligands that have been selected to be representative of what we believe is likely to be found in the compound collection of a pharmaceutical or biotechnology company. Results are presented for releases 1.8, 2.0, and 2.5 of Glide. The comparisons show that average measures for both "early" and "global" enrichment for Glide 2.5 are 3 times higher than for Glide 1.8 and more than 2 times higher than for Glide 2.0 because of better results for the least well-handled screens. This improvement in enrichment stems largely from the better balance of the more widely parametrized GlideScore 2.5 function and the inclusion of terms that penalize ligand−protein interactions that violate established principles of physical chemistry, particularly as it concerns the exposure to solvent of charged protein and ligand groups. Comparisons to results for the thymidine kinase and estrogen receptors published by Rognan and co-workers (*J. Med. Chem.* **2000**, *43*, 4759−4767) show that Glide 2.5 performs better than GOLD 1.1, FlexX 1.8, or DOCK 4.01.

1. Introduction

The previous paper[1] introduced Glide,[2] a new method for rapidly docking ligands to protein sites and for estimating the binding affinities of the docked compounds. That paper described the underlying methodology and showed that Glide achieves smaller root-mean-square (rms) deviations in reproducing the positions and conformations of cocrystallized ligands than have been reported for GOLD[3] and FlexX.[4] Better docking accuracy is important in its own right in lead-optimization studies, where knowledge of the correctly docked position and conformation (pose) of a novel ligand can be crucial. In lead-discovery studies, however, docking accuracy is relevant mainly to the degree that it contributes to obtaining high enrichment in database screening. We believe that accurate scoring requires accurate docking, though accurate docking is not enough in itself.

This paper investigates the ability of Glide 2.5, run in "standard-precision" mode,[1] to identify known binders seeded into database screens for a wide variety of pharmaceutically relevant receptors. We present comparisons with earlier versions of Glide and show that very substantial progress has been made. Rigorous comparisons with other virtual screening methods are difficult for us to make because we generally do not have access either to the identical sets of decoy ligands or to other docking codes. However, we have been given access to the thymidine-kinase and estrogen-receptor datasets employed by Bissantz, Folkers, and Rognan[5] and offer comparisons to the results they published for GOLD, FlexX, and DOCK.[6]

The paper is organized as follows. In the section 2, we characterize our data sets and protocols for evaluating database enrichment. This section describes the receptors and ligands to be used, discusses certain issues concerning preparation of the receptor (most importantly, the use of reduced van der Waals radii, which is essential to achieve reasonable results in some cases), and defines the quantitative measures used to assess performance in database screening. Section 3 presents enrichment factors obtained using default parameters and describes the individual screens. Comparisons to published results for GOLD, FlexX, and DOCK for the thymidine kinase and estrogen receptors are then presented in the fourth section, and Glide's sensitivity to the choice of certain input factors is explored in section 5. Finally, the sixth section summarizes the results and discusses future directions.

2. Virtual Screening Protocol

Ligand Databases and Receptors Used. We have chosen the following nine receptors for our initial studies, five of which are represented by two or more alternative cocrystallized receptor sites:

1. thymidine kinase (1kim)
2. estrogen receptor (3ert, 1err)
3. CDK-2 kinase (1dm2, 1aq1)
4. p38 MAP kinase (1a9u, 1bl7, 1kv2)
5. HIV protease (1hpx)
6. thrombin (1dwc, 1ett)
7. thermolysin (1tmn)
8. Cox-2 (1cx2)
9. HIV-RT (1vrt, 1rt1)

* To whom correspondence should be addressed. Phone: 646-366-9555, extension 106. Fax: 646-366-9550. E-mail: halgren@schrodinger.com.
† Schrödinger, L.L.C., NY.
‡ Columbia University.
§ Schrödinger, L.L.C., OR.

10.1021/jm030644s CCC: $27.50
Published on Web 02/27/2004

The receptors for these screens cover a wide range of receptor types and therefore provide a proper test of a docking method. All were prepared using the procedure described in the preceding paper[1] or an earlier version of that procedure.

The known binders for the first two systems were specified by Rognan and co-workers,[6] while ligands for the CDK-2 kinase receptor screens and for p38 MAP kinase were provided by pharmaceutical and biotech collaborators. For thrombin, 12 of the 16 known binders were taken from the studies by Engh et al.[7] and by von der Saal et al.[8] Others are ligands for the same target protein taken from our docking-accuracy test set[1] or were developed from multitple sources in the literature.

As database ligands, we employed "druglike" decoys that averaged 400 in molecular weight (the "dl-400" dataset) in most cases. For thymidine kinase (1kim), which has a very small active site, however, we used a similar (but in this case more competitive) set with an average molecular weight of 360 (the "dl-360" dataset). The property distributions of these databases were characterized in the preceding paper.[1] We believe these compounds to be 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 database ligands and between 7 and 33 known actives. All compounds considered have 20 or fewer rotatable bonds and 100 or fewer atoms. Like the database ligands, the known binders were also MMFF94s-optimized, but in these cases, we used unbiased input geometries obtained via a MacroModel conformational search, as previously described for ligands taken from cocrystallized complexes.[1]

Glide's use of reduced atomic van der Waals radii to mimic minor readjustments of the protein (these should be distinguished from the more substantial induced-fit rearrangements modeled by the use of multiple receptor conformations) is an important issue in the setup of the docking runs. Glide currently supports uniform van der Waals scaling of the radii of nonpolar protein and/or ligand atoms. To characterize the performance that can be expected when Glide is run "out of the box", the principal results presented in this paper use default 1.0 protein scaling (which means that the OPLS-AA van der Waal (vdW) radii are not changed) and 0.8 ligand scaling; the same scalings were used to assess docking accuracy.[1] In the fifth section, we compare these results with results obtained using scaling factors identified in earlier docking studies with Glide as giving optimal results. The comparisons show that default scaling performs well, though optimizing the scaling factors can improve the performance in some cases.

Measures of Virtual Screening Effectiveness. To quantify Glide's ability to assign high ranks to ligands with known binding affinity, we report enrichment factors in graphical and tabular form and present accumulation curves that show how the fraction of actives recovered varies with the percent of the database screened. Following Pearlman and Charifson,[9] the enrichment factor can be written as

$$\mathrm{EF} = \frac{\mathrm{Hits}_{\mathrm{sampled}}/N_{\mathrm{sampled}}}{\mathrm{Hits}_{\mathrm{total}}/N_{\mathrm{total}}} \quad (1)$$

Equivalently, this can be written as

$$\mathrm{EF} = \{N_{\mathrm{total}}/N_{\mathrm{sampled}}\}\{\mathrm{Hits}_{\mathrm{sampled}}/\mathrm{Hits}_{\mathrm{total}}\} \quad (2)$$

Thus, if only 10% of the scored and ranked database (i.e., $N_{\mathrm{total}}/N_{\mathrm{sampled}} = 10$) needs to be assayed to recover all of the $\mathrm{Hits}_{\mathrm{total}}$ actives, the enrichment factor would be 10. But if only half of the total number of known actives are found in this first 10% (i.e., if $\mathrm{Hits}_{\mathrm{sampled}}/\mathrm{Hits}_{\mathrm{total}} = 0.5$), the effective enrichment factor would be 5.

Equation 2 is useful when sampling an initial, small fraction of a database, but to measure performance for recovering a substantial fraction of the active ligands, we prefer to modify the definition of enrichment as follows:

$$\mathrm{EF}' = \{50\%/\mathrm{APR}_{\mathrm{sampled}}\}\{\mathrm{Hits}_{\mathrm{sampled}}/\mathrm{Hits}_{\mathrm{total}}\} \quad (3)$$

In this equation, $\mathrm{APR}_{\mathrm{sampled}}$ is the average percentile rank of the $\mathrm{Hits}_{\mathrm{sampled}}$ known actives. Intuitively, this makes sense: if the actives are uniformly distributed over the entire ranked database, the average percentile rank for an active would be 50% and the enrichment factor would be 1. Unlike eqs 1 and 2, however, this formula considers the rank of each of the $\mathrm{Hits}_{\mathrm{sampled}}$ known actives, not just the rank of the last active found (which is what N_{sampled} is likely to be). As a result, the enrichment factor will be larger than the value computed from eq 1 or 2 if the actives are concentrated toward the beginning of the N_{sampled} ranked positions but will smaller if the actives are grouped toward the end of this list. This is appropriate because a key objective in database screening is to find active compounds as early as possible in the ranked database; the new definition is better at indicating when this is happening.

3. Virtual Screening Results

Overview of Glide's Performance in Database Screening. Table 1 compares the performance of Glide 1.8, 2.0, and 2.5 SP (standard-precision)[1] using as definitions of enrichment EF′(70), which measures the enrichment for recovering 70% of the known actives, and EF(2%), which measures enrichment for assaying the top 2% of the ranked database. These represent "global" and "early" enrichment, respectively. Though 70% recovery is arbitrary, we feel this is a realistic standard for docking into a rigid protein site, given that such a site is unlikely to be properly shaped to house all the known actives when the site is relatively plastic. For use in virtual screening, it can be crucial to concentrate as many active ligands as possible in the topmost portion of the ranked database. Because the present screens use only ~1000 ligands, however, 2% (20 ranked positions or so) is about the smallest percentage we can examine, given that we have roughly 10−30 known actives to place. We also report enrichment factors EF(5%) and EF(10%) for Glide 2.5. Note that the maximum attainable enrichment factors are 50, 20, and 10, respectively, for EF(2%), EF(5%), and EF(10%). Also listed are average enrichment factors computed using a generalized geometric mean that weights the smaller enrichment factors more heavily.[10] (For example, the geometric mean of 1 and 25 is 5, not 13.) This weighting

Table 1. Comparison of Enrichment Factors for Glide 1.8, Glide 2.0, and Glide 2.5[a]

		EF′ (eq 3) 70% recovery			EF (eq 2) 2% of database			EF (eq 2) 5%	EF (eq 2) 10%
screen	site	GS 1.8	GS 2.0	GS 2.5	GS 1.8	GS 2.0	GS 2.5	GS 2.5	GS 2.5
thymidine kinase (tk)	1kim	4.2	7.6	17.2	0.0	10.0	25.0	12.0	9.0
tk-pyrimidine ligands	1kim	4.5	6.7	20.4	0.0	7.1	28.6	14.3	8.6
estrogen receptor	3ert	88.4	79.8	66.9	35.0	35.0	35.0	14.0	8.0
estrogen receptor	1err	88.4	37.5	46.7	35.0	30.0	30.0	14.0	9.0
CDK-2 kinase	1dm2	3.6	3.9	6.8	5.0	10.0	15.0	10.0	6.0
CDK-2 kinase	1aq1	2.1	3.8	5.4	5.0	15.0	10.0	8.0	6.0
p38 MAP kinase	1a9u	2.0	1.8	2.5	0.0	2.9	5.9	4.7	2.4
p38 MAP kinase	1bl7	1.8	2.9	3.5	2.9	5.9	8.8	3.5	4.1
p38 MAP kinase	1kv2	4.5	2.9	4.8	8.8	11.8	8.8	8.2	4.1
HIV protease	1hpx	10.8	7.8	47.6	30.0	13.3	36.7	17.3	8.7
thrombin	1dwc	5.8	2.8	12.1	6.2	3.1	15.6	11.2	8.1
thrombin	1ett	5.2	5.6	38.0	12.5	3.1	34.4	16.2	9.4
thermolysin	1tmn	1.6	15.2	24.5	5.0	15.0	25.0	18.0	9.0
Cox-2	1cx2	3.6[b]	3.4[b]	5.0[b]	7.6	3.0	13.6	7.9	5.2
Cox-2 (site-1 ligands)	1cx2	5.7	5.7	13.9	10.9	4.3	19.6	11.3	7.4
HIV rev. transcriptase	1rt1	4.6	3.2	8.2	3.0	0.0	12.1	10.3	6.4
HIV-RT	1vrt	1.8	2.0	7.1	0.0	0.0	7.6	9.1	5.8
av enrichment factor:		5.3	6.5	14.8	6.0	7.1	18.6	11.4	7.0

[a] Enrichment factors can be at most 50 for 2% sampling, 20 for 5% sampling, and 10 for 10% sampling. [b] EF′(60) value.

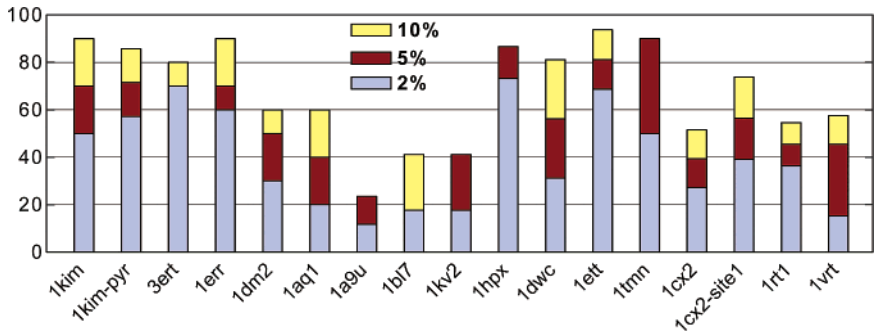

Figure 1. Percent of actives recovered with Glide 2.5 for assaying 2%, 5%, and 10% of the ranked database for the screens considered in this paper. The PDB codes are defined in Table 1.

is appropriate because what is most needed is a method that can be counted on to always perform reasonably well rather than one that does very well for some systems but is useless for others.

Table 1 shows that Glide 2.5 is much better than its predecessors at identifying active ligands. Because of improvements to the more difficult screens, both global and early enrichment have tripled since the release of Glide 1.8. The CDK-2 and p38 screens are still problematic, but thymidine kinase now does well, and thrombin, HIV protease, thermolysin, Cox-2, and HIV-RT all have improved substantially.

Figure 1 summarizes Glide's ability to rank active ligands in the first 2%, 5%, and 10% of the scored and ranked database. In many, though not all, cases, a significant fraction of the actives are found in the top 2% of the database.

Finally, Figure 2 displays the percent of known actives recovered as a function of the percent of the ranked database sampled for Glide 1.8, 2.0, and 2.5. This complementary view also shows that Glide 2.5 performs exceedingly well for many of the targets and also highlights cases in which further improvement is particularly desirable.

Detailed Results for Database Screens. The remainder of this section describes the individual database screens and presents graphical depictions of enrichment for Glide 1.8, 2.0, and 2.5. Detailed listings of the ranks of the active ligands, their GlideScore values, their hydrogen-bonding scores, and their Coulomb−vdW interaction energies with the protein site are available in Supporting Information (Tables S1−S15).

1. Thymidine Kinase (1kim). Rognan and co-workers studied the binding of 10 known thymidine kinase ligands to the protein from the 1kim complex.[6] For database ligands, they used 990 randomly chosen compounds from a filtered version of the ACD database. Only the 1kim ligand (dT) and one other ligand (idu) are reported to be submicromolar binders. The others (five are also pyrimidine derivatives and three are purines (acv, gcv, pcv)) range in activity from 1.5 to 200 μM. Realistically speaking, computational screening of compound databases usually can only hope to discover micromolar ligands. This poses a stiff challenge because Charifson et al.[11] found that the docking methods they surveyed performed reasonably well at finding low-nanomolar binders seeded into a database screen but fell off rapidly in efficacy as the activity of the known binders decreased. The ability to identify micromolar binders in a database screen is therefore a stringent and relevant test.

Figures 2a and 3 examine the thymidine kinase screen. In this case, Figure 3a uses the seven pyrimidine and three purine-based ligands defined by Rognan and co-workers as known actives while Figure 3b uses only the seven pyrimidines. Comparison of the lowest bar segments shows that Glide 2.5 is significantly more effective than Glide 1.8 or 2.0 at concentrating known

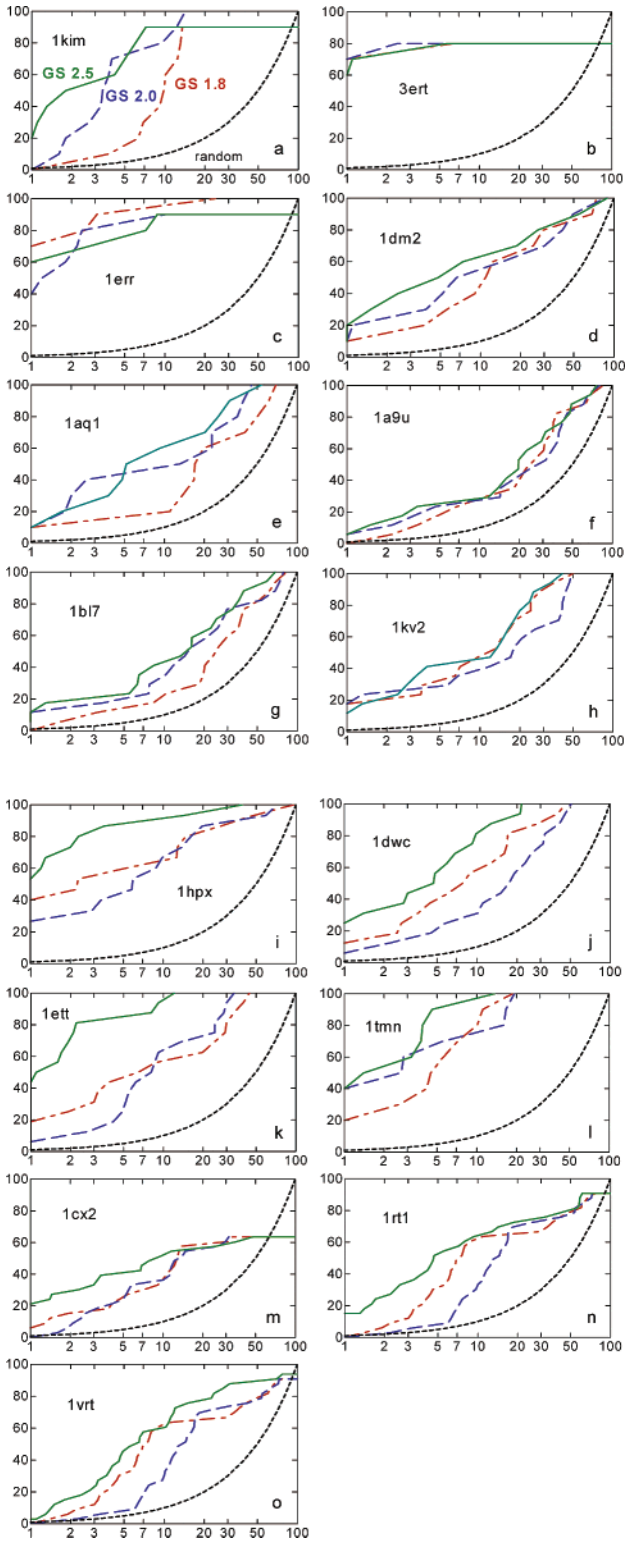

Figure 2. Percent of known actives found (*y* axis) vs percent of the ranked database screened (*x* axis) for Glide 2.5 (solid green), Glide 2.0 (blue dashed), and Glide 1.8 (red dot−dashed). Black dotted lines show results expected by chance. The listed PDB codes are defined in Table 1.

actives in the first 2% of the ranked database for both the mixed and pyrimidine-only screens; better performance relative to earlier verisons of Glide is also shown in Figure 2a for the mixed screen.

The reason for separating out the pyrimidines is that a labile Gln 125 side chain undergoes a 180° rotation

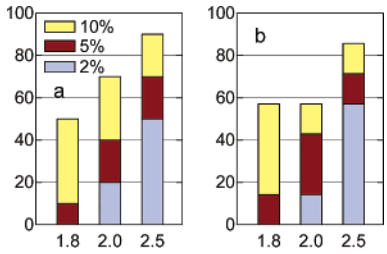

Figure 3. (a) Percent of thymidine kinase (1kim) actives recovered with Glide 1.8, 2.0, and 2.5 for assaying the first 2%, 5%, and 10% of the ranked database. (b) Percent recovered using only the seven pyrimidine-based ligands as actives.

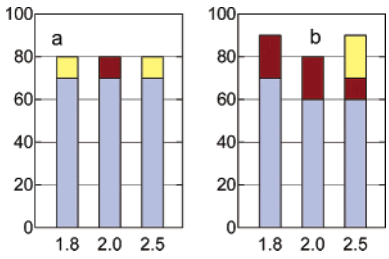

Figure 4. Percent of estrogen receptor actives recovered for assaying the top 2%, 5%, and 10% of the ranked database: (a) 3ert site; (b) 1err site.

on going from 1kim, a pyrimidine site, to one of the purine sites (2ki5, 1ki2, 1ki3). This rotation essentially exchanges the terminal NH_2 and C═O groups and means that purines cannot dock properly into a pyrimidine site, nor can pyrimidines dock properly into a purine site; in each case, the geometry that is correct for the parent site has an acceptor−acceptor and/or a donor−donor clash in the alternative site. This incompatibility results in larger rms deviations for cross-docking, as reported by Rognan and co-workers,[6] by Jain,[12] and by us.[1] Nevertheless, the misdocked purines find many favorable interactions and score well (Table S1, Supporting Information) because the terms in the standard-precision version of GlideScore 2.5 that penalize breaches of complementarity are too small to significantly penalize the misdocked purine structures. In contrast, Extra-Precision docking[13] imposes larger penalties for docking mismatches and therefore is more likely to be capable of rejecting ligands that do not dock properly into the site.

2. Estrogen Receptor (3ert, 1err). The target proteins for the estrogen receptor (ER) screen are the 3ert receptor site studied by Rognan and co-workers[6] and the 1err site used by Stahl and Rarey;[14] the native ligands are 4-hydroxytamoxifen and raloxifene, respectively. Both sites are open enough to dock antagonists as well as agonists. Our studies used the 10 low-nanomolar ERα antagonists that Rognan selected as active binders. This is one case in which the nonbonded radii need to be scaled down to allow the known binders to dock correctly. For example, five of the known binders had positive Coulomb−vdW interaction energies when no scaling was done. For Glide 1.8 and 2.0, we originally used 0.9 protein/0.8 ligand scaling, but we employ the default 1.0/0.8 scaling here.

Figures 2b,c and 4 show that both estrogen-receptor sites are treated very well by all three scoring functions. Tables S2 and S3 (Supporting Information) list the Glide 2.5 rankings.

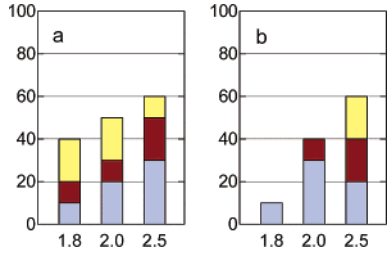

Figure 5. Percent of CDK-2 kinase actives recovered for assaying the top 2%, 5%, and 10% of the ranked database: (a) 1dm2 site; (b) 1aq1 site.

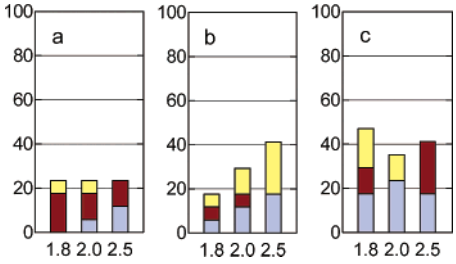

Figure 6. Percent of p38 MAP kinase actives recovered for assaying the top 2%, 5%, and 10% of the ranked database: (a) 1a9u site; (b) 1bl7 site; (c) 1kv2 site.

3. CDK-2 Kinase (1dm2, 1aq1). The CDK-2 kinase site is highly flexible. A key issue is the length of the rather narrow binding cavity, which if insufficient will prevent many active ligands from correctly docking. After examining superimposed structures for five co-crystallized PDB complexes, we chose a site from the 1dm2 complex that is more elongated than most, though not the most generous. To investigate the sensitivity of the docking to the choice of receptor site, we chose the slightly more open 1aq1 receptor as a second site. In each case, we used default 1.0/0.8 scaling.

Figures 2d,e and 5 show that Glide 2.5 outperforms its predecessors for the 1dm2 site and greatly outperforms them for 1aq1. The Glide 2.5 rankings and scorings are given in Tables S4 and S5 (Supporting Information).

4. p38 MAP kinase (1a9u, 1bl7, 1kv2). The p38 active site is particularly prone to alter its shape upon ligand binding. Therefore, we studied three different PDB receptor structures: 1a9u, 1bl7, and 1kv2. The 1kv2 site exhibits a particularly large alteration of the ligand-free structure in that a long loop undergoes a substantial change in conformation when its native ligand binds.[15] This screen employs 14 known p38 binders supplied by a colleague from the biotechnology industry in addition to the cocrystallized ligands from the 1a9u, 1bl7, and 1kv2 structures.

The plasticity of the p38 active site makes it difficult to dock a large number of active compounds properly into any single version of the receptor structure and hence leads to relatively small values of global enrichment such as EF′(70). Figures 2f−h and 6 show that the 1kv2 site is the most amenable one for the particular selection of active compounds used in this study. Glide 2.5 achieves relatively little improvement over previous versions in this case, in part because the p38 site is large and requires a very hydrophobic binding mode (in general, only one to two hydrogen bonds are made by correctly docked p38 actives). Such sites represent a severe challenge for most empirical scoring functions.

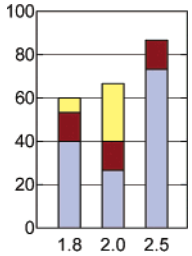

Figure 7. Percent of actives recovered for assaying the top 2%, 5%, and 10% of the ranked database for the 1hpx site of HIV protease.

We can report, however, that we have made significant progress in handling sites of this nature in the ongoing development of Extra Precision Glide.[13]

Detailed results for Glide 2.5 are shown in Tables S6−S8 (Supporting Information).

5. HIV Protease (1hpx). Our screening database contains 15 ligands from cocrystallized HIV-1 protease complexes included in our docking-accuracy test set.[1] Here, we focus on 1hpx as the target. The 1hpx complex retains the usual water under the "flaps" that enclose the active site, but we removed it so that ligands that displace this water, such as XK263 (from the 1hvr complex) and A-98881 (from 1pro), could dock. The removal of this water raises the question of whether the resultant overly generous site might recognize a large number of false positives. However, this proved not to be the case.

As with the estrogen-receptor screen, our original docking experiments used 0.9 protein/0.8 ligand scaling. However, this site is not especially "tight" and default 1.0/0.8 scaling works quite well, especially for Glide 2.5 (cf. Figures 2i and 7). Indeed, 7 ligands are found in the top 10 ranked positions and 12 are found in the top 20 (Table S9, Supporting Information). This is good performance by any standard.

6. Thrombin (1dwc, 1ett). Our docking-accuracy test set[1] contains five unique thrombin inhibitors (1dwc, 1etr, 1dwb, 1dwd = 1ets = 1ppc, and 1ett). However, only four are highly active because the 1dwb ligand, benzamidine, is too small to bind tightly (the experimental binding affinity is −5.4 kcal/mol [1]). To supplement these four binders, we included the 12 thrombin inhibitors from Engh et al.[7] and von der Saal et al.[8] that have reported binding affinities of 10 μM or better, bringing the total number of actives to 16.

To prepare for the original 1dwc screen for Glide 1.8, we found that all combinations of 0.8, 0.9, and 1.0 vdW scaling for nonpolar protein and ligand atoms gave strongly negative Coulomb−vdW energies for the known actives and afforded reasonably negative GlideScores. However, the model that used unscaled vdW radii for both the protein and the ligand gave the best overall GlideScores and yielded an unfavorable hydrogen-bonding score for only one ligand. We therefore chose to not scale the vdW radii. The present results, however, use default 1.0 protein/0.8 ligand scaling. Thus, this screen differs in the opposite sense from the estrogen-receptor and HIV-protease screens, which originally used greater scaling than the current default. As we show in the section 5, the two scaling models yield comparable results. Thus, employing larger scaling than is needed to allow the known actives to fit into the site

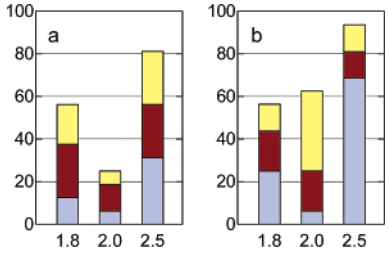

Figure 8. Percent of thrombin actives recovered for assaying the top 2%, 5%, and 10% of the ranked database: (a) human thrombin, 1dwc site; (b) bovine thrombin, 1ett site.

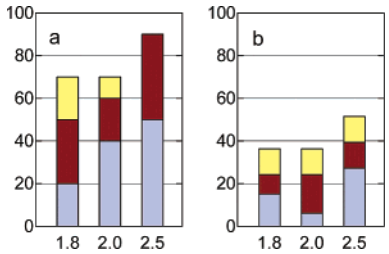

Figure 9. Percent of actives recovered for assaying the top 2%, 5%, and 10% of the ranked database: (a) thermolysin, 1tmn site; (b) Cox-2, 1cx2 site.

did not degrade Glide's ability to rank the known binders highly in this case.

Figures 2j,k and 8 show that Glide 2.5 performs better than either Glide 1.8 or 2.0. The same qualitative trend is seen in the EF′(70) and EF(2%) enrichment factors reported in Table 1. The comparisons show that the 1ett site consistently yields better results. Tables S10 and S11 (Supporting Information) list the Glide 2.5 rankings.

7. Thermolysin (1tmn). As target receptor, we chose the protein from the 1tmn complex. The 1tmn ligand, known as CLT for "carboxy-leu-trp", coordinates with the active-site Zn^{2+} ion via a carboxylate group and places its leucine side chain into a hydrophobic pocket. Including 1tmn, the docking-accuracy test set contains 13 thermolysin complexes, including one, 1lna, in which Co^{2+} replaces Zn^{2+}. However, the 2tmn, 4tln, and 1hyt ligands have only about 25 atoms (including hydrogens) and are too small to bind tightly; for example, the measured binding affinities for 2tmn and 4tln are only −5 to −7 kcal/mol.[1] In this study, we use the 10 larger, drug-sized ligands. Thermolysin is known to have a rigid active site, so the choice of target structure is probably unimportant in this case.

As for thrombin, preliminary studies suggested a preference for using unscaled radii for both the protein and the ligand (i.e., 1.0/1.0 scaling), but the results presented here use default 1.0/0.8 scaling. That unscaled radii yield good results seems clearly related to the rigid and open nature of the site. In this case, somewhat better results are obtained with 1.0/1.0 scaling (see section 5), but default scaling also does well. As Figures 2l and 9a show, Glide 2.5 performs much better than Glide 1.8 and somewhat better than Glide 2.0. For Glide 2.5, 5 of the 15 top-ranked ligands are known binders (Table S12, Supporting Information).

One key to the improved performance is that GlideScore 2.5 specifically rewards metal ligation by anionic ligand functionality. We made this change on the basis of experimental evidence that metalloproteases strongly favor anionic ligands.[16,17] A second element is that GlideScore 2.5 considers only the single strongest interaction when the ligation is bi- or multidentate. A third is that Glide 2.5 reduces the net ionic charges for most charged−charged and charged−polar interactions but leaves them unchanged for metal−ligand interactions. Thus, the Coulombic contribution to GlideScore 2.5 uses the full Zn^{2+}−ligand interaction energy, further helping to differentiate anionic ligands.

These elements were also included in GlideScore 2.0, but GlideScore 2.5 goes one step further by recognizing that neutral ligands such as imidazoles can be effective binders when Zn^{2+} and the "tripod" of protein residues on which it sits are net neutral (e.g., when Zn^{2+} is coordinated by two glutamates and a histidine, as in farnesyl protein transferase,[18] rather than by one glutamate and two histidines, as in thermolysin). In such cases, the term in GlideScore 2.5 that rewards a geometrically appropriate metal ligation by −2.0 kcal/mol when the ligand functionality is anionic is omitted when the apo site is net-neutral. While this modification seems appropriate, further studies will be needed to determine whether it gets the "balance" right for net-neutral sites.

8. Cox-2 (1cx2). We obtained structures for 33 known binders from the literature. These include the native 1cx2 ligand (SC-558, ligand 24), celecoxib (ligand 2), rofecoxib (ligand 3), indomethacin (ligand 10), deprotonated indomethacin (ligand 26), flurbiprofen (ligand 25), deprotonated flurbiprofen (ligand 33), ML-3000 (ligand 11), and deprotonated ML-3000 (ligand 31). Examination of various combinations of protein and ligand scaling factors using Glide 1.8 led us to select 1.0 protein/0.8 ligand scaling, which we also used here and for Glide 2.0.

These Glide dockings have one unusual feature, namely, that only 23 of the 33 known actives dock with negative Coulomb−vdW energies, even with relatively heavy scaling of protein and ligand nonpolar vdW radii, when a "normally sized" docking box centered around the 1cx2 ligand is used. When the docking box is made much larger, the remaining 10 ligands dock "successfully" but occupy a site that is displaced by 10−12 Å from the primary site. However, the "site 2" ligands have relatively poor Coulomb−vdW interaction energies and often make no hydrogen bonds. It thus seems unlikely that these Cox-2 ligands actually dock into this second site. The most likely explanation is that some variable element in the site geometry is responsible and that the limitations of docking to a rigid site are particularly extreme in this case.

Figures 2m and 9b show that Glide 2.5 performs much better than Glide 1.8 or 2.0. It does particularly well for early enrichment, as indicated by the lowest segments of the bar chart in the latter figure. Indeed, Glide 2.5 places 9 of the actives in the first 20 positions in the ranked database (Table 13S, Supporting Information). Figure 9b and Table 1 show that the calculated enrichment factors are relatively low when based on all 33 Cox-2 ligands. When recomputed to count only the 23 "site 1" ligands as actives, however, Glide 2.5 yields a quite decent EF′(70) value of 14.2. Thus, Glide 2.5 is effective at finding active Cox-2 ligands; it just cannot

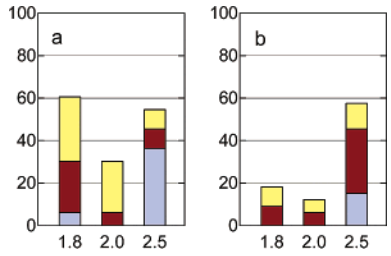

Figure 10. Percent of HIV reverse transcriptase actives recovered for assaying the top 2%, 5%, and 10% of the ranked database: (a) 1rt1 site; (b) 1vrt site.

dock and score all of them well when the 1cx2 site is used.

9. HIV-RT (1vrt, 1rt1). For HIV reverse transcriptase, we docked a set of 33 active ligands into the non-nucleoside binding site used by nevirapine, Sustiva, and other NNRTI compounds. The active ligands, taken from a variety of literature sources, include nevirapine and MKC-442. We used both the nevirapine site (1vrt) and the MKC-442 site (1rt1) as targets. On the basis of the dockings of the active ligands, we chose 0.9 protein/0.8 ligand scaling for 1rt1 and 1.0 protein/0.8 ligand scaling for 1vrt when testing Glide 2.0. The present results, however, use default 1.0/0.8 scaling for both sites.

Table 1 and Figures 2n,o and 10 show that Glide 2.5 greatly outperforms the earlier releases for these two HIV-RT sites. This site is also very hydrophobic and offers few hydrogen-bonding opportunities (cf. the hydrogen-bonding scores in Tables S14 and S15, Supporting Information). It therefore qualifies as a difficult site for an empirical scoring function such as GlideScore 2.5. In such a site it is crucial to recognize and penalize mismatches in complementarity in the docked poses. The improved performance relative to the earlier versions of Glide reflects the better balance of the more widely parametrized GlideScore 2.5 function as well as the inclusion of desolvation penalties; these terms play an even larger role in the Extra-Precision Glide 2.5 scoring function.[13]

4. Comparison to Other Methods

Comparisons usually are difficult for us to make because we do not have access to other docking codes and because published comparisons often use proprietary datasets.[11,14,19] In this section, however, we present comparisons to published results for GOLD 1.1,[3] FlexX 1.8,[4] and DOCK 4.01[20,21] for the thymidine kinase and estrogen receptors[6] using datasets provided to us by D. Rognan.[5] We caution that the earlier versions of GOLD, FlexX, and DOCK used by Rognan and co-workers may not be representative of the current capabilities of these methods. However, the same comparisons were also used by Jain in his recent paper introducing the Surflex method.[12]

Their study used GOLD, FlexX, and DOCK as docking engines and employed ChemScore,[22] FlexX, the DOCK energy score,[23] GOLD, PMF,[24] Fresno,[25] and Score[26] to rank the docked poses. The best result, obtained by a few of the 21 combinations of 3 docking engines and 7 scoring functions, found 8 of 10 actives in the first 8−10% of the ranked database. Given that only 0.8−1.0 actives would be found by chance, this performance corresponds to an enrichment factor of roughly 10. The best single models were DOCK with PMF scoring, FlexX with PMF scoring, and GOLD with GOLD scoring. Many models performed poorly, however. For example, when GOLD was used as the docking engine, ChemScore, the DOCK energy score, and Fresno all found just one active in the first 60% of the database, whereas six would be found by chance. Rognan and co-workers also found that some of the docking-method/scoring-functions combinations did poorly for the estrogen receptor, though others did well.

Table 2. Comparison of Enrichment Factors EF′(70)[a] for Glide, GOLD, FlexX, and DOCK for the Thymidine Kinase Receptor (1kim) and the Estrogen Receptor (3ert), Using Data Sets of Rognan and Co-workers[5,6]

	EF′(70), 70% recovery of known actives					
screen	Glide 1.8	Glide 2.0	Glide 2.5	GOLD[b] 1.1	FlexX[b] 1.8	DOCK[b] 4.01
thymidine kinase (1kim)	4.2	11.7	19.3	8.2	11.1	3.0
estrogen receptor (3ert)	70.0	72.1	47.1	28.5	8.9	6.7
av (geometric mean)	17.1	29.0	30.2	15.3	9.9	4.5

[a] Equation 3. [b] Reference 6.

Our calculations used Rognan's receptor and ligand preparations and employed the 0.9 protein/0.8 ligand scaling of nonpolar vdW radii we recommend for use when the protein site has not been relaxed to remove possible steric clashes (see section 5). We also used docking and scoring grids of the same size employed in section 3. As previously noted, Rognan and co-workers randomly selected the 990 database ligands from a filtered version of the ACD database. They employed various scoring functions in conjunction with each of the docking methods, but we focus here on the native GOLD-docking/GOLD-scoring, FlexX-docking/FlexX scoring, and DOCK-docking/DOCK-scoring combinations. These are the ones most likely to be used in a pharmaceutical setting, where project needs may not permit extensive explorations of alternative docking/scoring combinations to be carried out.

Comparisons of docking results for Glide, GOLD, FlexX, and DOCK are presented in Table 2 and in Figures 11 and 12. These comparisons show that DOCK docking followed by DOCK-energy scoring is the worst model. Glide 2.5 appears to be the best model overall by virtue of its superior performance for thymidine kinase, though Surflex does even better for this receptor[12] and though Glide 1.8 and 2.0 do better for the estrogen receptor (cf. Figure 12 and Table 2). The latter may reflect the better balance of the Glide 2.5 scoring function, which often yields better results for poorly treated screens at the cost of some degradation in performance for well-handled screens. FlexX and GOLD show decent enrichment, but neither is as effective as Glide.

5. Sensitivity to vdW Scaling Parameters

As noted previously, Glide by default leaves the protein radii unchanged but scales the nonpolar ligand radii by 0.8; we refer to this as 1.0/0.8 scaling. Scale factors smaller than 1.0 make the protein site "roomier"

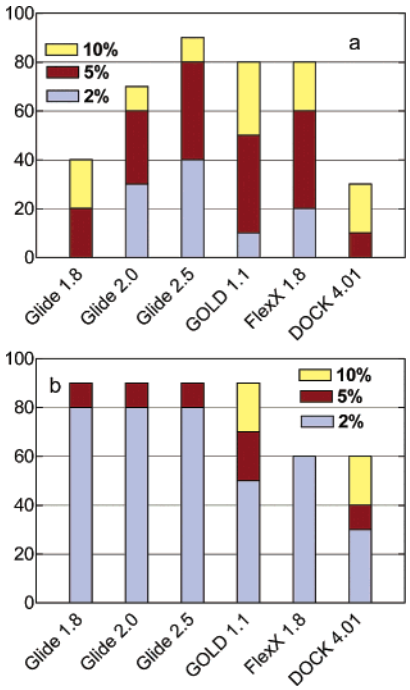

Figure 11. Percent of actives recovered by Glide 1.8, 2.0, and 2.5 and by GOLD, FlexX, and DOCK for assaying the first 2%, 5%, and 10% of the ranked database, using the datasets of Rognan and co-workers:[5,6] (a) thymidine kinase (1kim); (b) estrogen receptor (3ert).

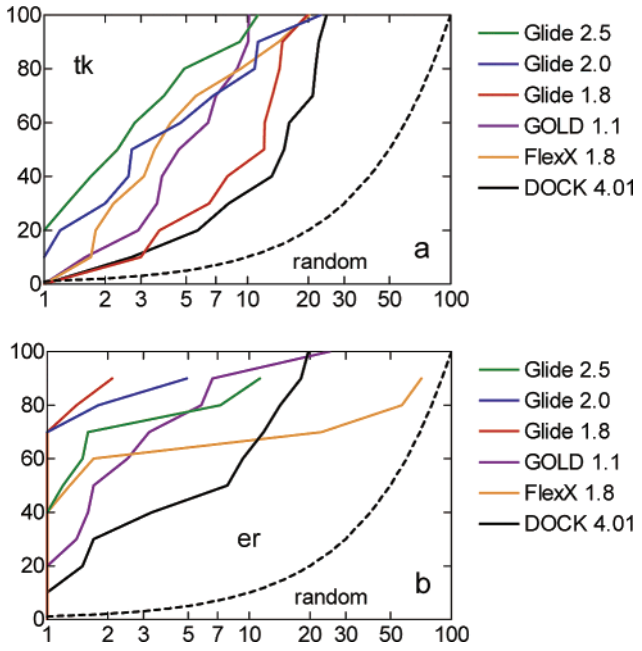

Figure 12. Percent of known actives found (*y* axis) vs percent of the ranked database screened (*x* axis) for Glide 2.5 (green), Glide 2.0 (blue), and Glide 1.8 (red) and for GOLD (purple), FlexX (orange), and DOCK (black) using the datasets of Rognan and co-workers:[5,6] (a) thymidine kinase (1kim); (b) estrogen receptor (3ert).

by "moving back" the surface of nonpolar regions of the protein and/or ligand. These adjustments emulate to some extent the effect of "breathing" motions a protein site might make to accommodate a tight-binding ligand that is slightly larger than the native, cocrystallized ligand. Cross-docking tests have consistently shown that it is important to modify the final vdW surface in this manner. Too much scaling, however, is detrimental because active ligands may no longer make suitably specific interactions with the receptor if the cavity is too large. Moreover, ligands that are too large to bind to the physical receptor may begin to dock and score well computationally, swelling the ranks of the false positives. The object is to find a "happy medium" between too little scaling and too much.

Table 3 shows how the choice of scaling factors effects the enrichment obtained in the database screens described in section 3. Thermolysin is an example of a rigid, open site, while the p38 MAP kinase site is highly mobile and the estrogen receptor contains a tightly enclosed hydrophobic channel. Alternative scalings are shown for cases in which the preferred scaling previously found for Glide 1.8 or 2.0 differs from the current default scaling. The table shows that the new default scaling works as well as or better than the previously identified preferential scaling in six of the eight cases. The original scaling gives substantially better enrichment factors for 1err and performs somewhat better for 1tmn, but the enrichment factors are high in these cases and the default enrichments are also good.

Table 4 shows that 0.9/0.8 scaling occasionally allows one or two additional actives to dock. Moreover, the more generous scaling almost always produces a significantly lower rank for the last "common" active found (e.g., the 8th for 3ert, the 9th for 1err, or the 21st for 1cx2). This, too, indicates that 0.9/0.8 scaling produces a better physical model when the fit is tight. The 1rt1 screen is an exception because 0.9/0.8 is not in fact the optimal scaling for Glide 2.5.

The conclusion we draw is that use of optimal scaling factors should be considered for "high-performance" screens. When active ligands are unavailable or will not be used to determine the scaling factors, the current default should normally be used. However, if the protein heavy atom coordinates are taken directly from the X-ray structure, it may be better to use 0.9/0.8 scaling to reduce the effect of unresolved steric clashes. This more generous scaling should also be used in cases in which it is known that the active-site region is tight and enclosed (an example being the hydrophobic channel of the estrogen receptor) because it will be difficult in such cases for certain active ligands to avoid serious steric clashes with the rigid site. Conversely, a lesser degree of scaling might be tried if the site is open and is known to be relatively rigid.

6. Discussion and Conclusions

This paper has presented results for 15 database screens covering 9 widely varying receptor types. Using recovery of 70% of the known actives as a benchmark, Glide 2.5 yields enrichment factors of at least 10 for all but CDK-2, p38, Cox-2, and HIV-RT and of less than 5 for only the 1a9u, 1bl7, and 1kv2 sites for p38. For Cox-2, the modest EF′(60) value found when all 33 actives are considered is mitigated by the finding that Glide 2.5 places 9 of the 33 known binders in the first 20 ranked positions. HIV-RT has long been problematic, but Glide 2.5 treats it considerably better than did its predecessors. Two of the most troublesome remaining screens are p38 and CDK-2, but progress is observed here, too, when XP Glide is employed.[13]

Table 3. Sensitivity of Calculated Enrichment Factors to vdW Scaling Sactors[a]

		vdW scaling		enrichment factor			
screen	site	protein	ligand	EF(2%)	EF(5%)	EF(10%)	EF′(70)
thymidine kinase (tk)	1kim	1.0	0.9	20.0	12.0	9.0	17.9
		1.0	0.8	25.0	12.0	9.0	17.9
tk-pyrimidine[a]		1.0	0.9	21.4	14.3	8.6	21.9
		1.0	0.8	28.6	14.3	8.6	22.5
estrogen receptor	3ert	0.9	0.8	40.0	16.0	8.0	70.7
		1.0	0.8	35.0	14.0	8.0	75.0
estrogen receptor	1err	0.9	0.8	35.0	18.0	9.0	60.4
		1.0	0.8	30.0	14.0	9.0	41.2
thrombin	1dwc	1.0	1.0	12.5	10.0	6.9	10.3
		1.0	0.8	15.6	11.2	7.5	11.6
HIV protease	1hpx	0.9	0.8	36.7	18.7	9.3	40.1
		1.0	0.8	40.0	17.3	8.7	46.0
thermolysin	1tmn	1.0	1.0	25.5	20.0	10.0	30.5
		1.0	0.8	25.0	18.0	9.0	24.5
HIV-RT	1rt1	0.9	0.8	6.1	7.9	6.4	6.4
		1.0	0.8	13.6	10.9	6.4	9.1

[a] In each case, the preferential scaling model used with Glide 2.0 is listed first.

Table 4. Number of Known Actives Docked with Negative Coulomb−vdW Interaction Energies as a Function of the Protein and Ligand vdW Scale Factors for Nonpolar Atoms

			no. docked		rank of last common active	
screen	site	no. of actives	1.0/0.8	0.9/0.8	1.0/0.8	0.9/0.8
estrogen receptor	3ert	10	8	9	58	17
estrogen receptor	1err	10	9	9	87	43
HIV protease	1hpx	15	15	15	408	204
Cox-2	1cx2	33	21	23	490	355
HIV-RT	1rt1	33	30	30	632	735

Comparison to results obtained for Glide 1.8 and 2.0 shows that average measures for both early and global enrichment are 2−3 times higher for Glide 2.5. Most importantly, Glide 2.5 performs significantly better for many of the more difficult screens; this qualitative improvement should be borne in mind when assessing comparitive studies based on Glide 1.8 or 2.0, which have begun to appear.[19] The improved enrichment stems partly from the inclusion of scoring-function terms that penalize ligand−protein interactions that violate established principles of physical chemistry, particularly as it concerns the exposure to solvent of charged protein and ligand groups. Given reports we have received from users that earlier versions of Glide were at least competitive in database enrichment to other commercially available methods, these results suggest that Glide 2.5 may represent a qualitative advance in scoring accuracy and virtual screening efficiency. Comparisons made to GOLD 1.1, FlexX 1.8, and DOCK 4.01 for the thymidine kinase and estrogen receptors using datasets prepared by Rognan and coworkers support this view, though we again caution that these comparisons may not be representative of the current capabilities of these methods.

Glide 2.5 has a number of advantages relative to previous versions. One is that generally good results are obtained with the new default 1.0 protein/0.8 ligand scaling. Calibrating the scale factors can lead to improved performance, but this may be less critical than with earlier versions of Glide, which employed less well-balanced scoring functions. A second advantage is that hydrogen-bond filters (i.e., imposition of a cutoff on the hydrogen-bond energy) and/or metal-ligation filters are no longer necessary. These elements broaden the range of applicability of Glide and simplify its use.

One theme that runs consistently through the results is that Glide does best when the active ligands make multiple hydrogen bonds to the receptor and does worst when the site is hydrophobic and offers few such opportunities. From what we have seen in the literature, this behavior is not unique to Glide. One of the key problems in database screening (one on which we have made considerable progress in ongoing work with XP Glide[13]) is how to properly model binding when it is mainly hydrophobic in character. These new developments will be described in a subsequent paper.

The most challenging problem in the use of docking methods in pharmaceutical applications is dealing with protein flexibility. When the protein structures differ by discrete, localized changes (protonation state modifications, alterations of side chain rotamer states), it should be possible to examine the variations in protein structure directly in a single docking run with suitable algorithms, thus saving considerable computational effort. When there is a larger perturbation of protein structure, however, as in cases such as 1kv2 in which there is a significant induced-fit component of ligand binding that results in substantial backbone or loop movement, other approaches are needed. The simplest approach is to carry out flexible docking into multiple rigid protein structures and then to combine the screening results. When knowledge of the appropriate ensemble of protein structures is available, this strategy is likely to succeed. Examples of this approach will be described in a subsequent paper in the context of Extra-Precision Glide.

Ultimately, accurate molecular mechanics modeling of the protein structure will be needed to enumerate the variations in active-site geometry that can be accessed at relatively low energies. Calculations along these lines, if successful, would obviate the need for cocrystallized examples. While modeling of this type is clearly quite difficult at present, methods using continuum solvation models such as those developed by Schrödinger in principle can address this problem effectively. If this can be accomplished, it would greatly enhance the effectiveness of any docking methodology in a wide range of practical applications. Efforts along these lines are

currently underway at Schrödinger and have yielded promising early results.

Acknowledgment. This work was supported in part by grants to R.A.F. from the NIH (Grants P41 RR06892 and GM 52018). We thank Dr. Didier Rognan for providing electronic copies of receptor, ligand, and decoy data sets for the thymidine kinase and estrogen receptors and of the rank orders found for GOLD, FlexX, and DOCK.

Supporting Information Available: Tables S1−S15, listing ranks of the active ligands, their GlideScore values, their hydrogen-bonding scores, and their Coluomb−vdW interaction energies with the protein site for the 15 database screens considered in this paper. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) 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.

(2) Schrödinger, L.L.C., New York.

(3) Jones, G.; Wilett, P.; Glem, R. C.; Leach, A. R.; Taylor, R. Development and validation of a generic algorithm and an empirical binding free energy function. *J. Mol. Biol*. **1997**, *267*, 727−748.

(4) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A. A fast flexible docking method using an incremental construction algorithm. *Chem. Biol*. **1996**, *261*, 470−489.

(5) Rognan, D. Bioinformatic Group, UMR CNRS, France. Personal communication to T.A.H.

(6) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem*. **2000**, *43*, 4759−4767.

(7) Engh, R. A.; Brandstetter, H.; Sucher, G.; Eichinger, A.; Bauman, U.; Bode, W.; Huber, R.; Poll, T.; Rudolph, R.; van der Saal, W. Enzyme flexibility, solvent, and "weak" interactions characterize thrombin−ligand interactions: implications for drug design. *Structure* **1996**, *4*, 1353−1362.

(8) von der Saal, W.; Kucznierz, R.; Leinart, H.; Engh, R. A. Derivatives of 4-amino-pyridine as selective thrombin inhibitors. *Bioorg. Med. Chem. Lett*. **1997**, *7*, 1283−1288.

(9) Pearlman, D. A.; Charifson, P. S. Improved scoring of ligand−protein interactions using OWFEG free energy grids. *J. Med. Chem*. **2001**, *44*, 502−511.

(10) The geometric mean defines the average enrichment factor as the *n*th root of the product of the *n* individual enrichment factors. This definition was further modified by replacing any enrichment factor of less than 1 by 1 and by weighting the contributions such that the composite weight is the same for each receptor type, even when two or three receptor site geometries are used.

(11) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method of obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem*. **1999**, *42*, 5100−5109.

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

(13) Murphy, R. B.; Friesner, R. A.; Halgren, T. A. Unpublished results.

(14) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem*. **2001**, *44*, 1035−1042.

(15) Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P. F.; Gilmore, T.; Graham, A. G.; Grob, P. M.; Hickey, E. R.; Moss, N.; Pav, S.; Regan, J. Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. *Nat. Struct. Biol*. **2002**, *9*, 268.

(16) Whittaker, M.; Floyd, C. D.; Brown, P.; Gearing, A. J. H. Design and therapeutic application of matrix metalloproteinase inhibitors. *Chem. Rev*. **1999**, *99*, 2735−2776.

(17) Babine, R. E.; Bender, S. L. Molecular recognition of protein−ligand complexes: Applications to drug design. *Chem. Rev*. **1997**, *97*, 1359−1472.

(18) Bell, I. M.; Gallicchio, S. N.; Abrams, M.; Beese, L. S.; Beshore, D. C.; Bhimnathwala, H.; Bogusky, M. J.; Buser, C. A.; Culberson, J. C.; Davide, J.; Ellis-Hutchings, M.; Fernandes, C.; Gibbs, J. B.; Graham, S. L.; Hamilton, K. A.; Hartman, G. D.; Heimbrook, D. C.; Homnick, C. F.; Huber, H. E.; Huff, J. R.; Kassahun, K.; et al. 3-Aminopyrrolidinone Farnesyltransferase inhibitors: design of macrocyclic compounds with improved pharmacokinetics and excellent cell potency. *J. Med. Chem*. **2002**, *45*, 2388−2409.

(19) Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based virtual screening. Evaluation of current docking tools. *J. Mol. Model*. **2003**, *9*, 47−57.

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

(21) Lorber, D. M.; Shoichet, B. K. Flexible ligand docking using conformational ensembles. *Protein Sci*. **1998**, *7*, 938−950.

(22) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions. I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des*. **1997**, *11*, 425−445.

(23) Shoichet, B. K.; Kuntz, I. D. Matching chemistry and shape in molecular docking. *Protein Eng*. **1993**, *6*, 723−732.

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

(25) Rognan, D.; Laumoeller, S. L.; Holm, A.; Buus, S.; Tschinke, V. Predicting binding affinities of protein ligands from three-dimensional coordinates: Application to peptide binding to class I major histocompatibility proteins. *J. Med. Chem*. **1999**, *42*, 6450−4658.

(26) Wang, R.; Liu, L.; Tang, Y. SCORE: a new empirical method for estimating the binding affinity of a protein−ligand complex. *J. Mol. Model*. **1998**, *4*, 379−384.

JM030644S