Cross-Docking of Inhibitors into CDK2 Structures. 1

José S. Duca,*,† Vincent S. Madison, and Johannes H. Voigt*,†

Department of Drug Design, Schering-Plough Research Institute, 2015 Galloping Hill Road, K15-1-1800, Kenilworth, New Jersey 07033

Received November 21, 2007

Predicting protein/ligand binding affinity is one of the most challenging computational chemistry tasks. Numerous methods have been developed to address this challenge, but they all have limitations. Failure to account for protein flexibility has been a shortcoming of many methods. In this cross-docking study the data set comprised 150 inhibitor complexes of the protein kinase CDK2. Gold and Glide performed well in terms of docking accuracy. The chance of cross-docking a ligand within a 2 Å RMSD of its experimental pose was found to be 50%. Relative binding potency was not properly predicted from scoring functions, even though cross-docking of each inhibitor into each protein structure was performed and only scores of correctly docked ligands were considered. An accompanying paper (Voigt, J. H.; Elkin, C.; Madison, V. S. Duca, J. S. *J. Chem. Inf. Model.* **2008**, *48*, 669−678) covers cross-docking and docking accuracy from the perspective of using multiple protein structures.

INTRODUCTION

Docking methods are a valuable tool in the computational chemist's toolkit and, as such, are widely used in different modeling scenarios: calibrating the methods using native binding modes from crystallographic protein/ligand complexes, predicting plausible novel binding modes in the absence of experimental structures, and performing virtual screening experiments. A number of publications cover these methods in detail, particularly a recent issue of the *Journal of Medicinal Chemistry* dedicated to assessing the current status of docking methods and scoring functions.[1] A comparison of docking methods and scores for pharmaceutical relevant targets was published. In this paper the distribution of RMSD values between the top-ranked docking poses and the corresponding crystal structures was examined.[2,3] A pioneering paper carefully examined cross-docking from the point of view of protein flexibility and virtual screening.[4] Due to the increasing number of crystallographic structures in the PDB (approximately 7 novel kinase structural motifs are published every year[5]), application of docking methods to kinases has been an important part of methods development and scoring function calibration.

Theoretically, cross-docking experiments can shed light on protein flexibility and how it affects docking experiments per se. Using a single rigid protein conformation could give significant docking errors for a series of analogs that induce even small conformational changes (induced-fit) in a flexible protein such as CDK2. It has been noted that CDKs regulate a series of processes, which work at the molecular level through conformational changes. The magnitude of such flexibility—in and around the catalytic cleft—can be large, and the structural changes have been described.[6] Recent computational and experimental studies have shown that conformational changes of CDKs play a critical role in ligand binding.[7]

The goal of this study is to assess the influence of induced-fit on cross-docking and to explore cross-docking schemes utilizing a large data set of crystallographic structures of ligand-CDK2 complexes to perform the docking experiments. A number of questions were addressed in this study: Does docking accuracy improve when a number of related complexes are analyzed instead of one, by increasing the sampling of experimental protein conformations? Is it possible to predict approximate relative binding affinities based upon docking scores? Is it feasible to identify the predominant factors that would permit docking to be used in a lead optimization program a priori? What are the advantages and disadvantages of using cross-docking in a modeling protocol? Can cross-docking reproduce effects of induced-fit and protein flexibility?

A large uniform data set of structures and affinities for protein/inhibitor complexes spanning a complete lead optimization program offers advantages in testing modeling techniques and in indicating needed improvements.

METHODOLOGY

Protein Preparation. For this study 150 high-resolution in-house crystal structures ranging from 1.28 to 2.08 Å (average 1.63 Å) were used.[8,9] All structures were aligned to the CDK2/magnesium ATP complex structure (PDB code 1b38) as a common reference by superimposing the backbone atoms of the residues within 5 Å of the Mg-ATP ligand.

All water molecules were deleted, and hydrogen atoms were added using Maestro 7.0.[10] All ligands were individually inspected, and the correct protonation state and tautomer were chosen to reproduce the hydrogen-bonding pattern to the hinge and other active site residues inferred from heavy atom distances. No unusual protonation or tautomeric states were required. Hydrogen atoms were minimized with all heavy

* Corresponding author e-mail: jose.duca@spcorp.com (J.S.D.), johannes.voigt@spcorp.com (J.H.V.).

† These authors contributed equally to this work; they should be regarded as joint first authors.

10.1021/ci7004274 CCC: $40.75 © 2008 American Chemical Society
Published on Web 03/07/2008

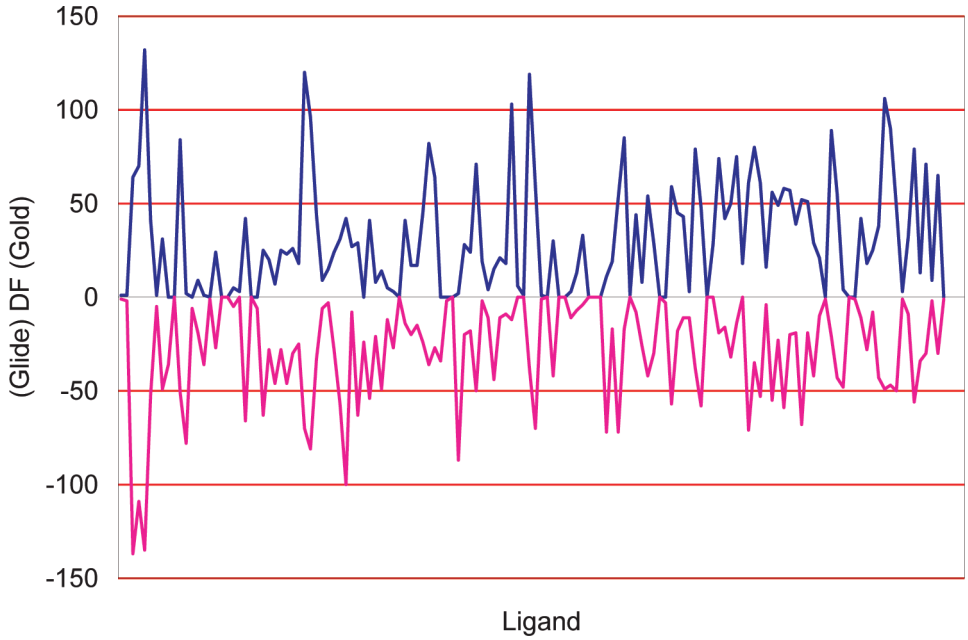

Figure 1. Gold and Glide docking frequency spectrum at 1 Å. For simplicity the Glide DF spectrum is shown as negative values.

atoms fixed (Macromodel[10] v9.0, MMFF94s force field, GB/SA implicit solvation model, PRCG, convergence threshold 0.05).

After deleting the ligand, a Glide[10] grid was computed for each of the 150 protein structures (Impact v3.5[10]) with an inner box region (the ligand midpoint remains within this box during docking) of 10 Å and an outer box region of 24 Å. The grid midpoint was defined by the centroid of one of the ligands in the data set.

3D Structural Diversity of 150 Protein Structures. In order to determine the 3D diversity of the 150 crystallographic structures, the ATP-site volume and the RMSD values of five key active site residues—Tyr-15, Val-18, Lys-33, Phe-80, Asp-145—in comparison to the published CDK2 crystal structure (PDB code 2r3i[9]) were calculated. Details about the calculations and the figures and tables can be found in the Supporting Information. The calculated ATP-site cavity volume ranges from 350 to 660 Å^3 with a majority of the structures displaying a volume of approximately 550 Å^3. This larger degree of flexibility is also reflected by variability of the position of Tyr-15 (Gly rich loop; 0−2.25 Å movement; one structure moves up to 5 Å) and Lys-33/Asp-145. The two latter residues display a bimodal distribution, between two predominant side-chain conformations. The residue positions of Val-18 and Phe-80 in contrast do not fluctuate significantly.

Ligands and Ligand Preparation. The ligand molecules were subjected to a conformational search (Macromodel, 1000 MCMM steps, 500 minimization steps, same force field as above), and the lowest energy conformation was used for the docking experiments conducted with both Glide and Gold. The conformational search was performed in order to remove all bias from the starting conformation and orientation.

For 140 of the 150 ligand molecules CDK2/cyclin A inhibition data (IC50) were available.[9] These ligands cover 21 distinct ring systems with the major class having 109 members. 13 ring classes have only one member, and the remaining seven cover two to four ligands. Ring systems were determined using the Pipeline Pilot[11] "Generate Fragments" component, with parameters: ring assemblies including exocyclic double bonds. The IC50 values from in-house measurements range from 30 pm to 60 μM (average 0.85 μM). The molecular weight of the ligand molecules spans from 188 to 559 Dalton (average MW is 392 Da). The average number of rotatable bonds is 4.8 (0 minimum − 10 maximum; Figure 2). The average length of the ligands used in this study in their bound conformation is 10.7 ± 1.8 Å (ranging from 6.5 to 16.8 Å), while in the unbound conformation the average extent is 10.1 ± 1.7 Å (ranging from 6.5 to 17.5 Å).

Docking and Result Analysis. Every ligand was docked to all 150 crystal structures.

For Glide, extra precision docking (XP, Impact v. 4.0217, Linux RH ES 3.0) default settings were used (Maestro v. 7.5), and one pose (5 poses for the validation set, best scoring pose was considered) was reported. Glide uses two "boxes" during the generation of its grids. The grids are computed within the space defined by the "outer box". This box encompasses all the ligand atoms. The "inner box" is defined to contain all acceptable positions for the ligand center upon docking. In the case of Gold the active site cavity is defined by a sphere.

Five GA runs without early termination for each ligand in Gold v2.2[12,13] were conducted. Standard default settings except for internal_ligand_h_bonds = 1, n_ligand_bumps = 0, flip_free_corners = 1, flip_amide_bonds = 1, flip_planar_n = 1, and flip_pyramidal_n = 1 were chosen. The active site was defined by a radius of 17 Å around the Leu83 backbone nitrogen. The ligand pose with the best fitness value was considered in our analysis.

Gold and Glide treat the proteins according to the rigid receptor approximation. Gold optimizes the hydroxyl and lysine ammonium torsions.

In Glide the ligands are treated flexibly as conformational ensembles during the docking process. In Gold the ligand torsion angles are optimized, and planar and pyramidal nitrogen atoms and free ring corners are "flipped".

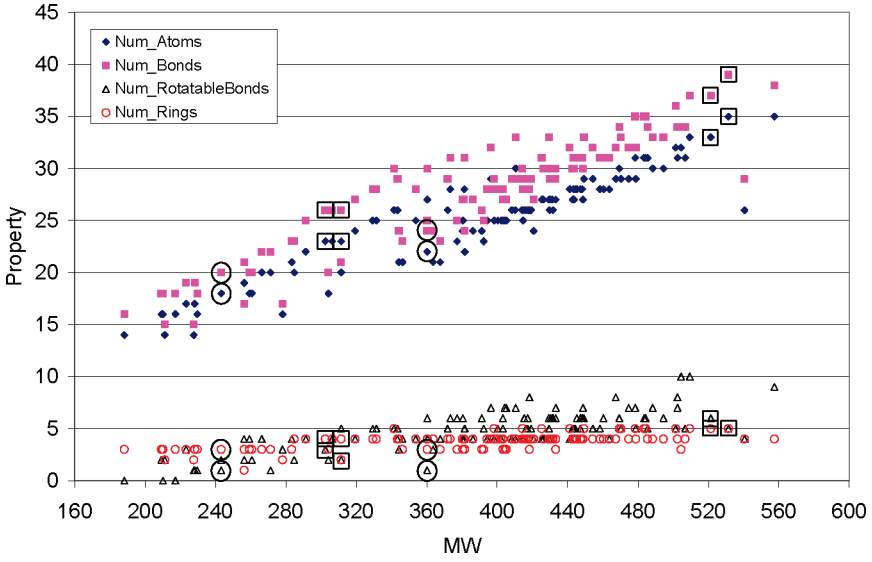

Figure 2. Property description of the entire compound data set.[11] Two promiscuous compounds are indicated with circles, and four difficult compounds are indicated by squares.

The docking accuracy is the degree of agreement between the orientation and conformation of a ligand observed in the crystal structure and the pose derived from docking experiments. Since all complex structures and thereby all ligands were aligned to a common reference, it is possible to compare the coordinates of a ligand pose docked to a "non-native" structure, to its coordinates of the crystal structure. We use as a measure of accuracy the root-mean-square deviation between the docked pose and the one observed in the crystal structure [RMSD = sqrt(Σ[distance $atom_{ref\ i}$ − $atom_{docked\ i}$]^2/*N*), where *N* is the number of atoms].

The necessity of symmetry considerations (e.g., phenyl ring rotations) makes correct calculation of RMSD nontrivial. The CACTVS system (v3.223)[14] calculated all mappings of the reference ligand substructure from the crystallographic structure onto the docked pose of this ligand. The RMSD based on the corresponding atoms was calculated for all mappings; the lowest RMSD was used.

Each ligand pose obtained from docking the 140 ligands of the crystallographic set to the 150 CDK2 protein conformations was characterized by both its accuracy (RMSD) and docking score (more negative Glide scores and higher Gold fitness values both correspond to better binding).

RESULTS AND DISCUSSION

In the first part of this section the docking results for the two docking methods of 140 ligands docked to 150 protein structures are analyzed from a "*ligand centric*" viewpoint—how a given ligand is docked across all protein conformations.

In the second part of this section, the results are analyzed from a "*protein-centric*" viewpoint. The docking accuracy was assessed for all protein conformations by considering one protein conformation at a time. Also, the log(IC50)/score correlation was compared for Glide versus Gold for single protein conformations.

Ligand-Centric Analysis. To investigate how well a given ligand is docked across all protein conformations, each of the 140 ligands was docked into each of the 150 protein structures using Glide and Gold. **D**ocking **F**requency, DF, was defined as the number of protein structures in which a given ligand is docked within the RMSD criterion. Ideally this value should reflect the ability of a given ligand to fit in binding cavities of a structure that differs from its cognate structure as well as the ability of the docking method to identify the correct binding pose as the top pose. The plot of *docking frequencies* (or docking frequency spectrum, DFS) at a given RMSD cutoff is a good measurement of docking performance on a *per ligand* basis (see Figure 1).

Table 1. Comparison of Docking Performance in Terms of Docking Frequencies, DF, at the RMSD Cutoff Values of 1 and 2 Å to the Experimental Ligand Location[a]

	1 Å			2 Å		
	Gold	Glide	Gold and Glide	Gold	Glide	Gold and Glide
0 < DF ≤ 10	51	48	35	10	9	4
10 < DF ≤ 70	71	82	52	29	29	16
70 < DF ≤ 140	18	10	2	94	95	80
<DF>	31	28		93	90	

[a] Three DF levels are defined: low (0 < DF ≤ 10), medium (10 < DF ≤ 70), and high (70 < DF ≤ 140).

Table 1 indicates that at 1 Å RMSD most of the compounds docked well and therefore had medium and high docking frequencies. In the case of Gold, 63% of the compounds (71 + 18) met the most stringent RMSD criterion. Using Glide, 66% of the compounds (82 + 10) had medium and high DF values. Gold and Glide docked the same 52 compounds (37%) at medium DF values, while only 2 compounds (1%) showed high docking frequencies with both methods.

Some of the compounds exhibited low docking frequencies. 35 compounds (25%) had low Gold and Glide docking frequencies (DF ≤ 10). This means that only 25% of the data set compounds cannot be docked in up to 10 of the CDK2 structures at 1 Å RMSD. The DF values at 1 Å indicate that the docking methods succeeded in terms of docking accuracy and that Glide and Gold can achieve

medium and high docking frequencies for most of the compounds of the data set.

From these DFS graphs and the average DF values, <DF>, is concluded that Gold performs marginally better. Although the number of compounds with low and medium docking frequencies is similar (51 and 71 compounds for Gold versus 48 and 82 compounds in the case of Glide), Gold performs at high DF values for 18 compounds versus 10 for Glide.

When the RMSD cutoff is allowed to be less stringent, both programs perform similarly. The performance of Gold and Glide was better at 2 Å. 123 ligands were docked with medium and high DFs by Gold, whereas in the case of Glide it was 124 ligands. This represents 88% of the data set. As expected, a better docking performance at 2 Å translates into a shift from medium to high DF values. These DF values along with the <DF> value indicate that while overall performance is satisfactory at 2 Å, the differences between Gold and Glide are minimal. On average one can expect that 3 times more ligands will meet the docking criterion at 2 Å than at 1 Å RMSD.

Table 1 also indicates that, at 2 Å, only 4 compounds cannot be docked in more than 10 of the CDK2 structures. This number decreased significantly from 35 compounds at 1 Å, the most stringent RMSD cutoff value. The opposite is observed at high DF values; at 1 Å only 2 compounds docked correctly in both methods, but the number increases to 80 compounds (57% of the data set) at a less stringent RMSD value of 2 Å.

In order to understand the reasons behind very low or very high docking frequencies, we took a closer look at the compounds that did not dock well at 2 Å (4 "difficult compounds") and the compounds that docked in most of the structures at 1 Å (2 "promiscuous compounds"). Is it a problem with the size and flexibility of these difficult compounds or is it a unique binding mode that is hard to reproduce?

Figure 2 shows several characteristics of these 6 compounds (MW, number of atoms, number of bonds, number of rings, and number of rotatable bonds) and a comparison to the entire data set.

Figure 2 indicates that one of the promiscuous compounds is small and compact, while two of the difficult ones are rather large and flexible. However these compounds are neither the smallest nor largest, respectively, of the data set. No clear distinction could be found for the other three compounds in terms of their molecular properties.

Analysis of the binding modes of these compounds indicates that out of the four difficult compounds, only one had a bad contact in the cognate X-ray structure. This could explain the difficulties of the docking methods in finding the correct binding pose for this compound. No other bad contacts were found for the rest of the compounds. In general, their experimental conformations could not be reproduced due to lack of sampling of linker groups or generation of the wrong initial geometry by the force field used.

To better address comparisons among DFSs, we computed the difference of docking frequencies, ΔDF, as defined by eq 1

$$\Delta DF = DF(Gold) - DF(Glide) \quad (1)$$

If the docking frequency distributions from Gold and Glide are similar at a given RMSD cutoff, then the absolute value of ΔDF should be low (defined as $|\Delta DF| < 10$). $\Delta DF > 40$ means Gold clearly outperforms Glide (Gold is better in at least 40 CDK2 structures). If Glide outperforms Gold, then ΔDF will be negative.

The difference of docking frequencies between Gold and Glide was computed at the RMSD cutoff values of 1 Å and 2 Å, as indicated in Figure 3.

These ΔDF distribution plots can be utilized as docking fingerprints to analyze the docking quality versus the RMSD cutoffs. Ideally, if all compounds were well docked by both methods, the spectra would have mean docking frequency differences near zero with small deviations for the individual ligands. Large deviations from zero are observed at RMSD cutoffs of 1 Å, 1.5 Å (data not shown), and 2 Å, indicating that for some of the ligands one of the methods performs better than the other.

Analysis of ΔDF plots shows that Gold outperforms Glide at 0.5 Å (80/140 = 57%), 1.5 Å (81/140 = 58%), and 2 Å (77/140 = 55%). At an RMSD of 1 Å, Glide performs better for 79 out of 140 compounds (56%). The plots at 0.5 and 1.5 Å are not shown.

Since Figure 3a,b is sorted by compound number, the plots can be interpreted as data set fingerprints for a given cutoff value. These docking fingerprints show that for some of the ligands the two methods considered behave quite differently. From the ΔDF plots at 0.5, 1, 1.5, and 2 Å, a subset of eight compounds (from now on, compounds 1−8) for which one of the methods clearly outperformed the other one was selected. The selection criteria will be described below.

Figure 4a represents the distribution of docking frequencies for this subset of compounds at 4 RMSD cutoff values (0.5, 1, 1.5, and 2 Å).

In the case of compound 1, Glide provided a docking pose within the cutoff criteria in almost all of the cases for cutoff values greater than 1 Å. On the contrary, Gold only docked this compound properly in $^2/_3$ of the cases for RMSD values greater than 1.5 Å.

For compounds 2−4, Glide also clearly outperformed Gold. Glide found consistently accurate poses in approximately 80−100 out of 150 grids. Gold's poor performance is reflected in small docking frequency values (DF < 55) even at 2 Å RMSD cutoff. Alternatively, in the case of compounds 5, 6, and 7 Gold outperforms Glide in terms of docking frequency. For these ligands, Gold reports accurate docking solutions at 1.5 and 2 Å cutoff values, while Glide can only identify the correct poses in 50% of the cases for compound 6. The case of compound 8 is slightly different from the rest, since it shows a solid Gold performance for RMSD values greater than 1 Å and a gradual improvement in the case of Glide for the same range of RMSDs. For this compound if it is only considered the 2 Å cutoff, it could be concluded that both methods perform similarly, when in reality Gold performs better than Glide at RMSD smaller than 2 Å.

The properties of compounds 1−8 were analyzed using the same set of descriptors as above (Figures 2 and 4b).

The first question to address is the following: Are these compounds very different from the rest of the data set? The properties profile of these compounds indicates that the 8

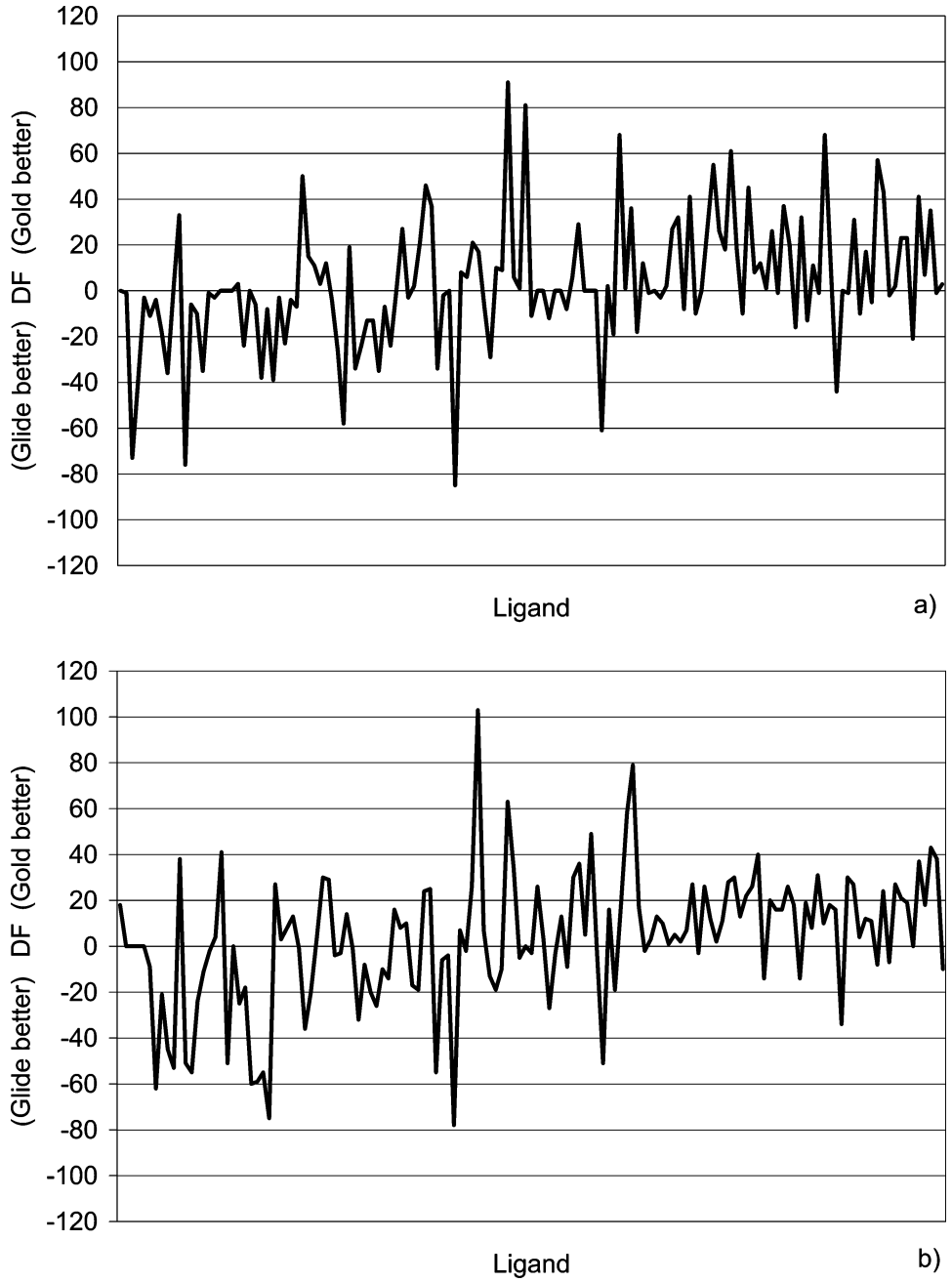

Figure 3. a, b: ΔDF spectra at 1 and 2 Å RMSD cutoff. Based on eq 1 positive ΔDF values indicate that Gold performs better, while negative ΔDF values mean that Glide performs better.

compounds do not share any particular anomalous property that differentiates them from the rest. Notably, compounds 1 and 2 are fairly small and rigid with no bad contacts in their X-ray complexes. One might expect these compounds to dock well with any docking protocol, as happened with one of the promiscuous compounds described previously in this section. After visual inspection of the 5 top reported poses it is clear that Gold finds the correct solution but does not reward it with the best score. Conversely, the two larger compounds (compounds 6 and 7; MW > 475) are not the largest compounds of the data set and yet are treated very differently by Gold and Glide, in contrast to their treatment of the two largest nonpromiscuous compounds, discussed above.

In the case of the other four compounds similar trends were found. None of these properties could explain why the docking scores were not properly assigned to reward the correct pose and place it at the top of the list, or why the correct pose was never found.

The inconsistencies found for the "difficult" compounds and this small set of 8 compounds are not representative of the entire data set. They are probably reflecting minor irregularities of the scoring functions and force fields used in this paper. Such inconsistent behavior has been described previously in docking validation setups.[15,16]

However, this situation is worrisome, since it suggests that small differences in the ligands switch the relative performance of the two methods, hinting at an under-

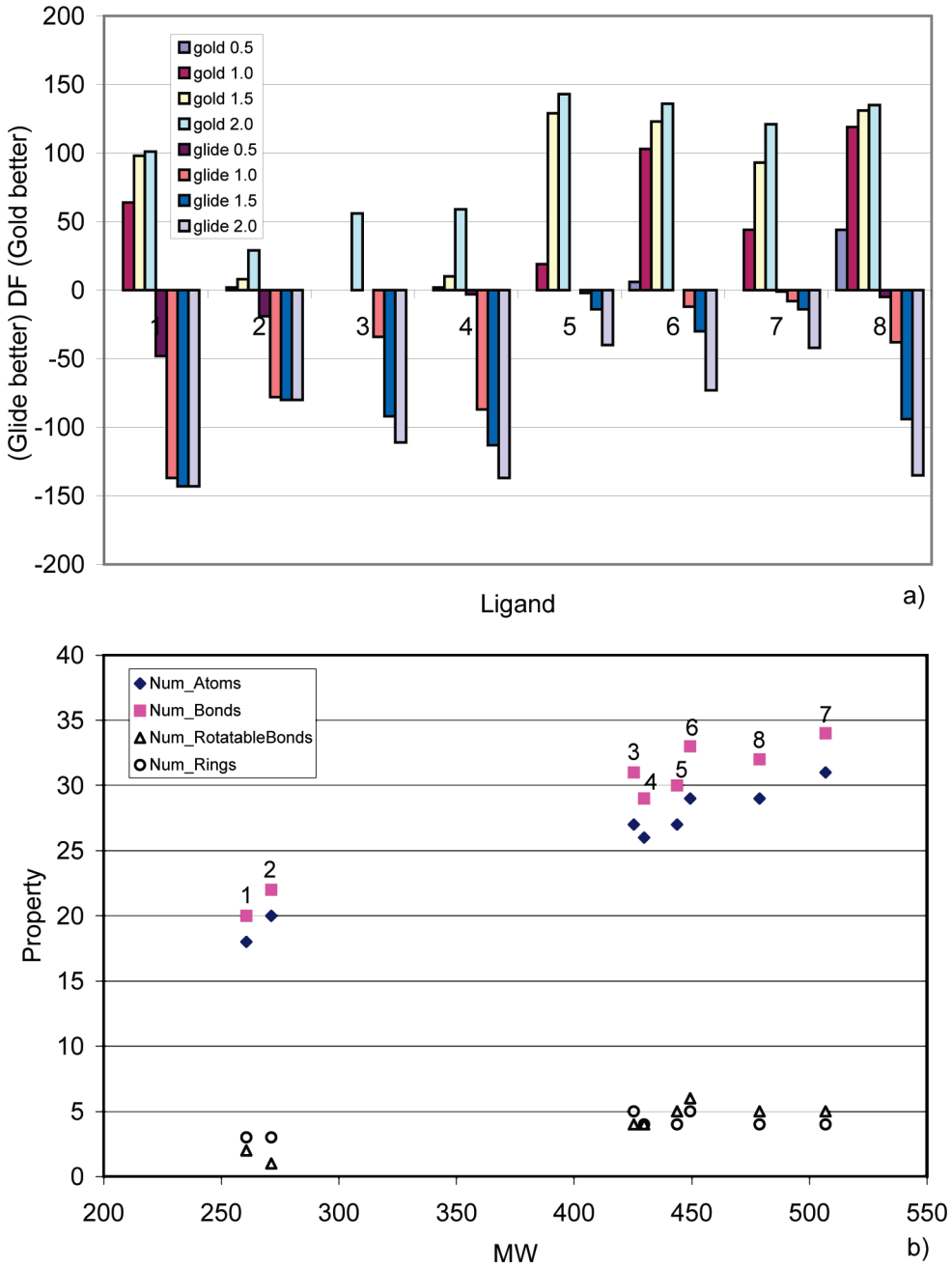

Figure 4. a: DF distribution plot at the RMSD cutoffs 0.5, 1, 1.5, and 2 Å and b: property description of compounds 1−8 at the RMSD cutoffs 0.5, 1, 1.5, and 2 Å.

lying lack of robustness that can negatively affect docking results.

Protein-Centric Analysis. *Single Grid Experiments.* This section compares the performance of Glide and Gold in terms of docking accuracy and score predictivity. RMSD is used as the principal accuracy metric. Several drawbacks have been mentioned regarding the use of RMSD to account for accuracy;[17] however, in the context of this study RMSD is a simple metric that can be used to compare with previously published work.

For scoring function predictivity, the correlation between the computed scores and the log(IC50) values will be considered.

Docking Accuracy. For Glide, the most stringent RMSD cutoff strongly decreases the number of reported poses that docked "correctly" (i.e., within the 1 Å cutoff) for the 150 crystallographic structures. Figure 5 shows that in approximately 60% of the grids 15−25% of the compounds were docked "correctly". Furthermore, in only 23 (15%) protein structures 25−35% of the ligands were docked within an RMSD of 1 Å. That means that only a small fraction of the protein conformations allow to dock just a third of the ligands correctly at this stringent RMSD cutoff.

In the case of Gold the results are slightly better, and there is a small fraction of the protein structures in which more than 35% of the compounds were docked within the 1 Å cutoff.

The lack of high docking accuracy can be due to two reasons. It is known that the active site of CDK2 is flexible and induced-fit effects can play a role. Moreover, some of

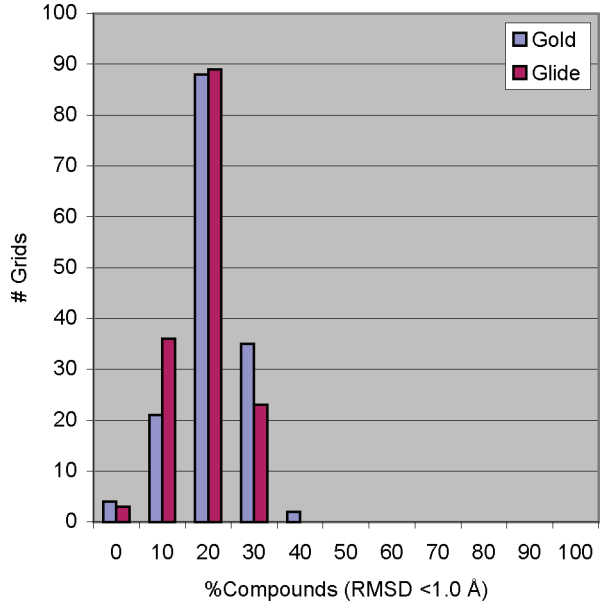

Figure 5. Centered histogram[18] of a number of grids having a given percentage of compounds docked with an RMSD of 1 Å or better.

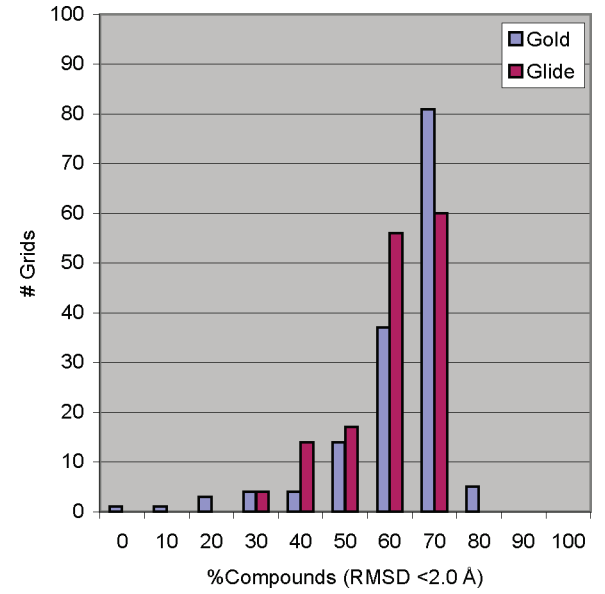

Figure 6. Centered histogram[18] of a number of grids having a given percentage of compounds docked with an RMSD of 2 Å or better.

the CDK2 inhibitors can extend beyond the enclosed binding site into regions of the protein that are mostly solvent exposed or may bind via water-bridged interactions. Since explicit water molecules are not included in the present docking protocol, the latter factor would negatively impact the docking performance. Both facts may arguably increase the RMSD beyond 1 Å for compounds with a small number of atoms; therefore, larger cutoff values were considered.

If the RMSD cutoff is increased to 2 Å (see Figure 6), both docking methods perform better, i.e., ligands are docked within the required RMSD criterion for a larger number of grids than at 1 Å RMSD.

Glide accuracy for more than 75% (56 + 60/150 = 77%) of the grids ranges between 55 and 75% of the ligands, which is similar to the values previously reported by Cummings et al.[19] and Perola et al.[2] A recent comprehensive evaluation of docking for pose prediction reports comparable docking accuracy values.[19,20]

Gold once again performs better in terms of accuracy, peaking between 65 and 75% of the compounds and showing a small fraction (5/150 = 3%) of the protein structures with an accuracy greater than 75%.

In Figure 7 the average accuracy of Glide and Gold are compared. For the 150 structures a plot binned every 0.25 RMSD unit is generated. From the plot we deduce that Gold slightly outperforms Glide for the structures with the most accurately docked ligands. 53% (26 + 54 = 80/150) of the structures allowed ligand docking within bins with an average accuracy of 2−2.25 Å,versus 48% (26 + 46 = 72/150) in the case of Glide. The worst performing protein structures (bins greater than 4.25 Å RMSD) comprised 3% of the total for Glide and 10% for Gold (4 and 12 protein structures, respectively). The difference among the worst-docked compounds may be understood noting that Gold explores a larger volume of a 17 Å diameter sphere, while Glide restricts the centroid of the ligand to a 10 × 10 × 10 Å cube.

Figure 8 compares the percentage of the ligands which dock correctly in Glide and in Gold for a given protein structure. There is a large cluster where for both methods the protein structures perform well. Notably, the worst performing protein conformation is the same for both methods. But there are several cases where Gold performs well for a given grid, while Glide's docking performance is less accurate and vice versa. This type of performance was observed and described in the Ligand-Centric section above.

The five structures, for which Gold docks at least 75% of the compounds within a 2 Å RMSD (Figure 6), are the obvious choice for docking analogous compounds in a typical lead-optimization effort. This type of analysis is usually performed as a lead-optimization program advances, and additional crystallographic structures become available.

The best and worst performing protein structures for Glide and Gold are not necessarily the same. We could not find any correlation between the accessible active site volume and the date the best docking structures were solved. We could not decide a priori what structures would yield better docking accuracy. This situation exemplifies the typical difficulties of using docking in a lead-optimization program. Since the best set of crystallographic structures cannot be anticipated, we have to assume that the chances of docking a compound correctly to a given crystallographic structure are approximately 50% with 80% of the structures predicting 70% of the compounds (Figures 5 and 6).

Single Protein Structure: log(IC50)/Docking Score Correlation. There have been several reports[16,17,21] of correlations between a docking method's scoring functions and experimental binding affinity data such as K_d or IC_{50} values. In the case of this study, such analysis only makes sense in terms of the Glide scoring function, which is derived by fitting to experimental values and is intended to reproduce the free energy of binding.[22] The Gold scoring function, on the contrary, was calibrated to identify the best docking pose out of a conformational ensemble. Herein a comparison of both scoring functions was generated, but the discussion will focus mainly on the Glide scores.

This analysis differs from previous work in that correlations between log(IC50) and score are only considered after applying an RMSD cutoff. That is, we try to couple docking

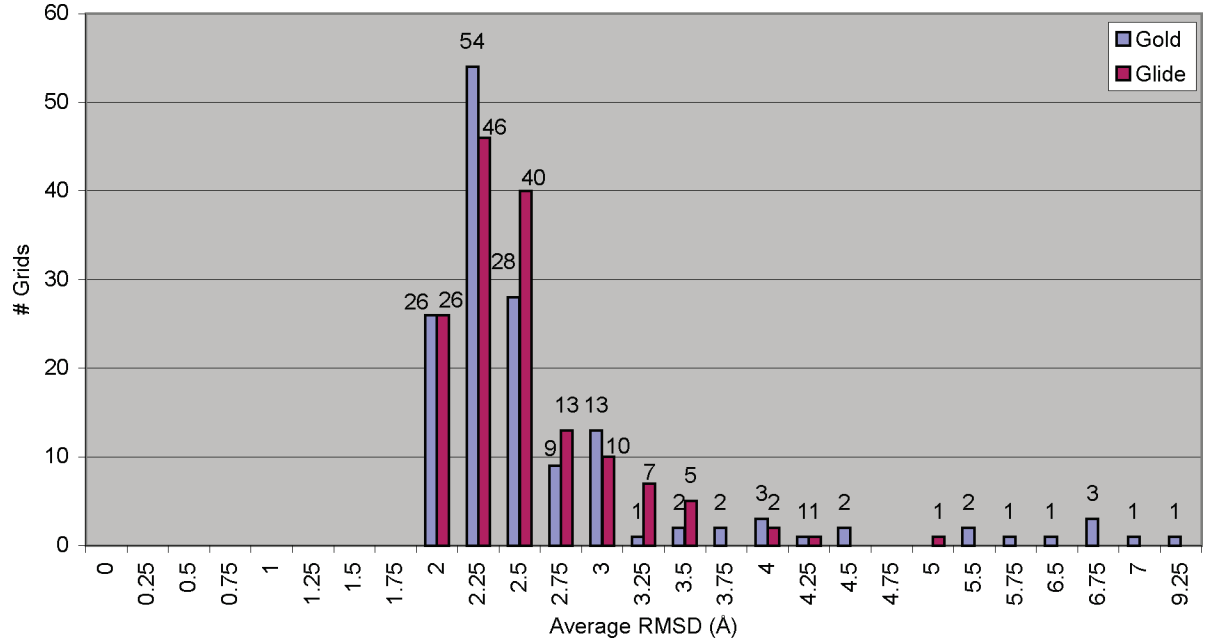

Figure 7. Centered histogram[18] of a number of grids having a given average RMSD across all docked ligands.

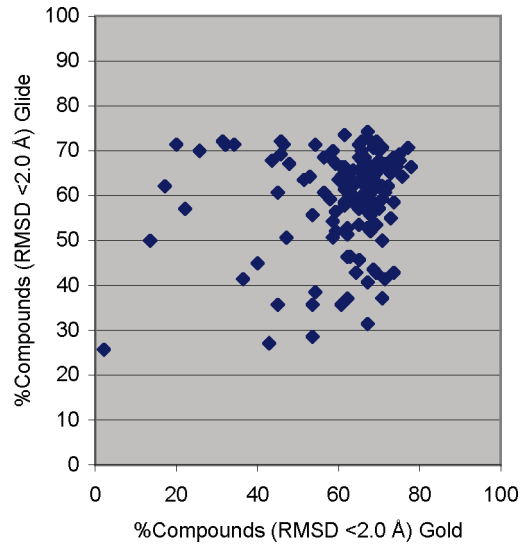

Figure 8. Percentage of compounds docked with a RMSD of 2 Å or better per protein structure. Comparison of Glide vs GOLD.

accuracy to scoring performance to assess if there is a relationship between the two variables. Furthermore, this decouples sampling—obtaining the correct docking pose—from scoring. The log(IC50)/score correlations were assessed using both R and R^2 to understand the distribution and nature of the correlation.

The Glide score was fitted to correspond to the free energy of binding; therefore, the ideal log(IC50)/score correlation should give an R value of 1 as shown in Figure 9a). In contrast, the Gold score is a fitness function which is designed to be larger for better-docked and/or better-bound ligands. Consequently for Gold the ideal log(IC50)/score correlation would be −1 (Figure 9a). When comparing with Glide (Figures 9b,c,d) the negative R for Gold is plotted.

The RMSD cutoffs allowed us to assess if better-docked compounds are rewarded with more accurate scores, which, in turn, should translate into better correlations. Given uncertainties in the scoring functions and experimental values, it is unreasonable to expect a perfect correlation, but the authors rather anticipated an approximate correlation of $R^2 \cong 0.6$ in the best cases. A recent report[23] indicated that current force fields can yield errors of up to 5 kcal/mol just during the conformational generation for a free energy prediction cycle. That being the case, the current status of force fields should not allow more than rough correlations between log(IC50) and docking scores.

The log(IC50)/score correlation value could be used to guide the selection of the best (or group of best) crystallographic structure(s) to be utilized during the lead optimization process. Ideally a set of crystallographic structures could be used to dock several non-native compounds accurately while predicting their approximate binding affinity.

Glide: Analysis of log(IC50)/Score Correlation. Figure 9b shows the correlation coefficient (R) distribution for compounds that dock to each CDK2 crystal structure with an RMSD value below 1 Å. Very high correlations with a negative sign, which is the wrong trend, can be found for a small number of grids. If R^2 were to be used for some of these grids, the correlations would be greater than 0.7. The reason why negative correlations are found is the stringent RMSD cutoff used in this case which limits the number of compounds to pass the criterion; the chance of obtaining a high correlation increases when the number of compounds decreases. The peak of log(IC50)/score correlation is about $R = 0.5$ for 22% of the grids (33/150). 1% of the grids have an R of 0.7−0.8, but this is again due to the small number of compounds that pass the RMSD cutoff.

If the cutoff is increased from an RMSD of 1 Å to 2 Å, then the number of compounds correctly docked increases considerably from ∼20% to 60−70%. Figure 9c shows that the border effects of very few structures with very high or very low R values have disappeared, and the distribution is more compact. For 42 of the 150 grids (28%) the correlation ranges from 0.25 to 0.35.

When all the compounds are considered (Figure 9d), with a cutoff of 100 Å, the correlation distribution looks similar

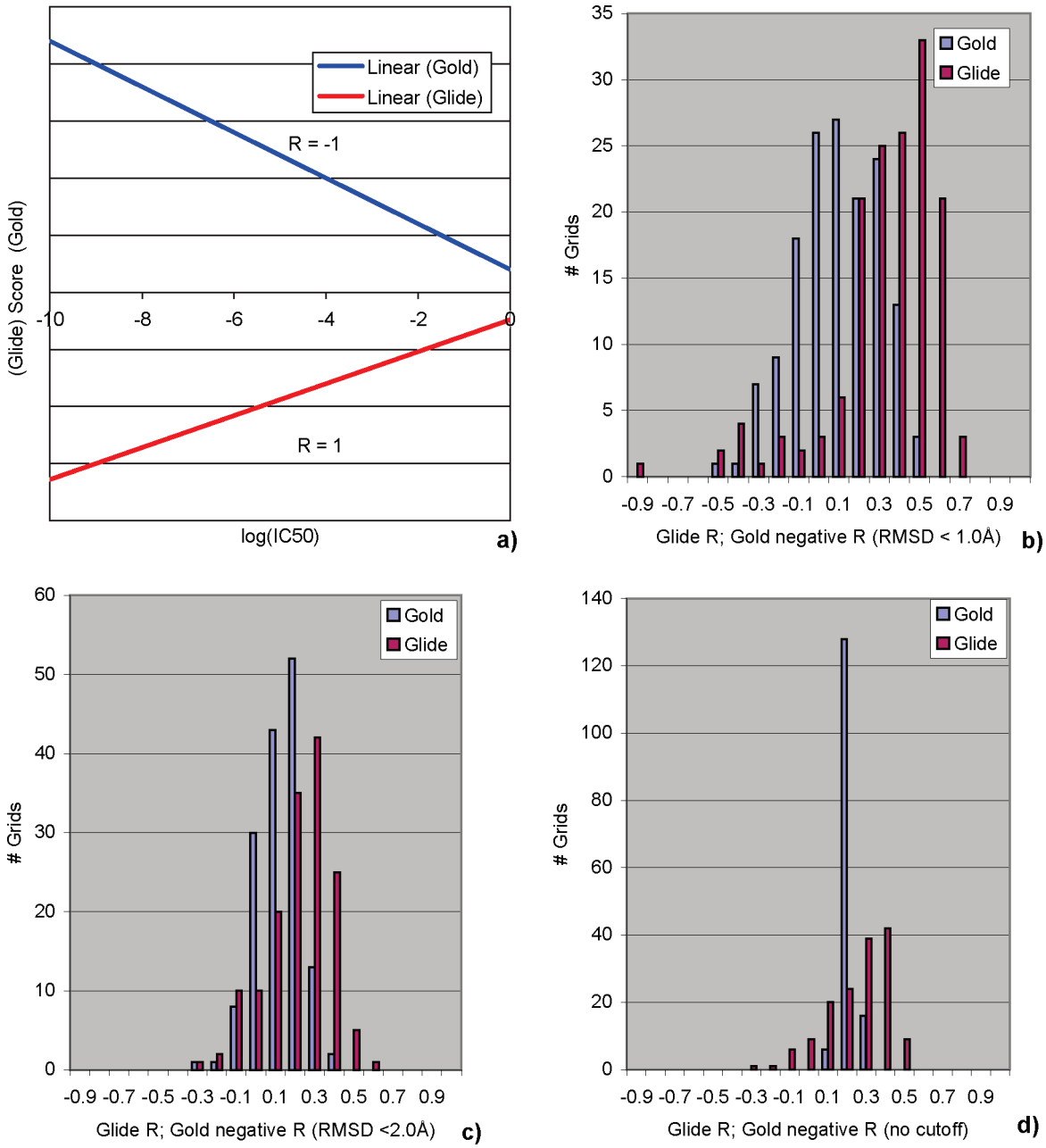

Figure 9. (a) Ideal correlations for Glide score and Gold fitness with log(IC50); centered histogram of a number of grids having a given score/log(IC50) correlation for ligands docked with a RMSD of (b) <1 Å, (c) <2 Å, (d) >100 Å, or no cutoff. For Gold negative R is plotted, since higher Gold scores correspond to better pose ranks.

to that for the 2 Å cutoff. Interestingly, 90 (39 + 42 + 9) grids now have a correlation higher than 0.25, compared to 73 grids (42 + 25 + 5 + 1) at a 2 Å RMSD.

This means that the log(IC50)/score correlation improves for compounds that are docked with an accuracy less than 2 Å for two possible reasons: a) the RMSD cutoff of 2 Å is too stringent or b) misdocked compounds that are getting high scores.[16,19]

Gold: Analysis of log(IC50)/Score Correlation. "The Gold fitness function was designed to discriminate between different binding modes of the same molecule. Extra terms are probably required to compare different molecules."[24] Figure 9b−d shows the comparison of Gold and Glide scores at the RSMD cutoffs of 1, 2, and 100 Å. In all cases the correlation distribution for Gold is quite compact and inferior to Glide.

At the most relevant RMSD cutoff of 2 Å, Gold correlation's distribution peaks at the 0.1 to 0.2 bins for 52 grids (37%), while Glide reaches the maximum correlation (0.3 to 0.4 bins) for 43 grids out of 150 (29%).

When all the compounds are considered, in 90% of the grids or more Gold has a correlation distribution between 0.1 and 0.2.

SUMMARY

The present study explores some fundamental questions that scientists face every time they run a docking protocol of a noncognate compound/protein complex: a) Do the docking scores predict potency? b) Are they accurate enough to guide improvement of binding affinity? c) How can extensive experimental structural information best be used

to help a lead-optimization program? d) Should some/all of the structures be used, and if so which?

Clearly when it comes to docking accuracy, the current methodologies examined in this paper perform quite well. Gold appears to outperform Glide slightly, but at the same time Gold does not display docking accuracy as consistently as Glide. Gold and Glide achieve medium and high docking frequencies for most of the ligands of the data set. We concluded that 3 times more compounds can be docked by Gold and Glide when the RMSD cutoff is relaxed from 1 to 2 Å. In addition, the chances of docking a compound within a 2 Å RMSD were estimated to be 50%.

This study put particular emphasis on decoupling sampling from scoring. The log(IC50)/score correlation was analyzed for given RMSD cutoffs, since only a correctly docked pose should be considered for log(IC50)/score correlations.

With respect to predictivity of scores, there seems to be a stronger signal in the case of Glide vs Gold, which is expected due to the nature of Gold's scoring function. Still, the ability of either scoring function (as it has been reported previously with other programs—see ref 16) to predict binding affinity is quite poor. The reasons behind this could be related to the inherent errors in the force fields,[23] insufficient treatment of entropy,[25] or perhaps the fact that fitting is being used to fine-tune the functions, which could introduce overfitting-related errors. In fact, it is quite troublesome to find that for Glide almost 10% of the protein structures produced negative correlations for correctly docked ligands (RMSD < 1.5 Å).

The relationship between random choice and cross-docking as well as the effect of using multiple protein structures on docking accuracy are discussed in an accompanying paper.[26]

ACKNOWLEDGMENT

We would like to thank Dr. Charles Lesburg for his careful review of this manuscript and in-depth comments. We also thank Alan Hruza and Dr. Thierry Fischmann for making available the crystallographic data, the CDK2 team for providing inhibitors and inhibition data, and our colleagues in the Drug Design group for helpful discussions and continued support. Furthermore, we are grateful to Drs. Richard Friesner, Ramy Farid, and Woody Sherman from Schrödinger, Inc. for valuable scientific input.

Note Added after ASAP Publication. This article was released ASAP on March 7, 2008 with an incorrect Figure 2. The correct version was posted on March 10, 2008.

Supporting Information Available: Methods of 3D structural diversity of 150 protein structures, histograms of the CDK2 ATP-site cavity volume (Figure 1) and of Tyr-15, Val-18, Lys-33, Phe-80, and Asp-145 RMSD from 2r3i reference (Figure 2), and RMSD from 2r3i reference crystal structure for CDK2 active site residues (Table 1). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of Protein-Ligand Interactions. Docking and Scoring: Successes and Gaps. *J. Med. Chem.* **2006**, *49*, 5851−5855, and references therein.

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

(3) Perola, E.; Walters, W. P.; Charifson, P. Comments on the Article On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors. *J. Chem. Inf. Model.* **2007**, *47*, 251−253.

(4) Cavasotto, C. N.; Abagyan, R. A. Protein Flexibility in Ligand Docking and Virtual Screening to Protein Kinases. *J. Mol. Biol.* **2004**, *337*, 209−225.

(5) Fedorov, O.; Sundström, M.; Marsden, B.; Knapp, S. Insights for the development of specific kinase inhibitors by targeted structural genomics. *Drug Discovery Today* **2007**, *12*, 365−372.

(6) (a) Pavletich, N. P. Mechanisms of cyclin-dependent kinase regulation: structures of cdks, their cyclin activators, and cip and INK4 inhibitors. *J. Mol. Biol.* **1999**, *287*, 821−828. (b) Russo, A. A.; Tong, L.; Lee, J. O.; Jeffrey, P. D.; Pavletich, N. P. Structural basis for inhibition of the cyclin-dependent kinase Cdk6 by the tumour suppressor p16 INK4a. *Nature* **1998**, *395*, 237−243.

(7) (a) Park, H.; Yeom, M. S.; Lee, S. Loop Flexibility and Solvent Dynamics as Determinants for the Selective Inhibition of Cyclin-Dependent Kinase 4: Comparative Molecular Dynamics Simulation Studies of CDK 2 and CDK 4. *Chem. Bio. Chem.* **2004**, *5*, 1662−1672. (b) Kriz, Z.; Otyepka, M.; Bártová, I.; Koca, J. Analysis of CDK2 Active-Site Hydration: A Method to Design New Inhibitors. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 258−274. (c) Villacanas, O.; Perez, J. J.; Rubio-Martinez, J. Structural analysis of the inhibition of Cdk4 and Cdk6 by p16 (INK4a) through molecular dynamics simulations. *J. Biomol. Struct. Dyn.* **2002**, *20*, 347−358.

(8) Dwyer, M. P.; Paruch, K.; Alvarez, C.; Doll, R. J.; Keertikar, K. et al. Versatile templates for the development of novel kinase inhibitors: Discovery of novel CDK inhibitors. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 6216−6219.

(9) Fischmann, T. O.; Hruza, A.; Duca, J. S.; Ramanathan, L.; Mayhood, T. et al. Structure-guided discovery of cyclin-dependent kinase inhibitors. *Biopolymers* (*Pept. Sci.*) **2008**, *89*, 372−379.

(10) Schrödinger, LLC, New York, NY, 2007.

(11) *Pipeline Pilot v. 6.0.3.0*; Accelrys, Inc.: San Diego, CA.

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

(13) 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.

(14) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and management of chemical properties in CACTVS: An extensible networked approach toward modularity and compatibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109−116.

(15) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74−82.

(16) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J. et al. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.

(17) Shoichet, B. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862−865.

(18) All histograms are centered, which means e.g. a bin labeled with 10% contains items from 5% to 15% and a 1 Å RMSD or average RMSD bin with a bin size of 0.5 Å contains items ranging from 0.75 to 1.25 Å.

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

(20) Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On evaluating molecular-docking methods for pose prediction and enrichment factors. *J. Chem. Inf. Model.* **2006**, *46*, 401−415.

(21) Lovell, T.; Chen, H.; Lyne, P. D.; Giordanetto, F.; Li, J. On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors. *J. Chem. Inf. Model*. **2006**, *46*, 401−415.

(22) (a) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J. et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749. (b) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R. et al. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **2006**, *49*, 6177−6196.

(23) Tirado-Rives, J.; Jorgensen, W. L. Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. *J. Med. Chem.* **2006**, *49*, 5880−5884.

(24) CCDC. *Gold v2.2*; Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, England.

(25) Chang, C. A.; Chen, W.; Gilson, M. K. Ligand configurational entropy and protein binding. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1534.

(26) Voigt, J. H.; Elkin, C.; Madison, V. S.; Duca, J. S. Cross docking of inhibitors into CDK2 structures. 2. *J. Chem. Inf. Model.* **2008**, *48*, 669−678.

CI7004274