Cross-Docking of Inhibitors into CDK2 Structures. 2

Johannes H. Voigt,*[,†,§] Carl Elkin,[‡] Vincent S. Madison,[§] and José S. Duca*[,†,§]

Department of Drug Design, Schering-Plough Research Institute, 2015 Galloping Hill Road, K15-1-1800, Kenilworth, New Jersey 07033, and Schering-Plough Research Institute Cambridge, 320 Bent Street, Cambridge, Massachusetts

Received November 21, 2007

In the preceding paper (Duca, J. S.; Madison, V. S.; Voigt, J. H. *J. Chem. Inf. Model.* **2008**, *48*, 659−668), the accuracy of docking and affinity predictions of the Gold and Glide programs were investigated using single protein conformations spanning 150 CDK2/inhibitor crystallographic complexes. High docking accuracy was observed with both methods; furthermore, Glide showed modest log(IC50)/score correlations. In this part of the study, the effect of combining docking results from multiple protein conformations in a consensus fashion was probed. This approach enhanced docking accuracy only for Glide, which was attributed to the nature of its scoring function. For log(IC50)/score correlations, particular emphasis was placed on considering only scores from correctly docked poses. Using multiple instead of single protein structures showed an improvement in the correlations. Validation sets and scrambling experiments were used to examine the statistical significance and predictivity of these correlations. Rather than actual improvements in scoring accuracy, docking to multiple protein conformations produced overfitting artifacts.

INTRODUCTION

The importance of protein/ligand docking methods in structure-based drug design is indisputable. Docking is widely used spanning the range from predicting new binding modes to virtual screening experiments. Recent publications and reviews cover these methodologies in detail.[1]

Despite the advances in the field many disadvantages still exist, in particular, related to protein flexibility. The use of a single rigid protein conformation could yield significant docking errors. Cross-docking to ensembles of protein conformations is one answer to protein flexibility and induced-fit effects. A pioneering paper[2] examined cross-docking from the viewpoint of protein flexibility and virtual screening.

Part 1 of this study[3] constituted the first comprehensive study of cross-docking using a large and uniform data set of crystal structures and affinities for protein/ligand complexes. The ligands of 140 inhibitor complexes of the protein kinase CDK2 were docked into all 150 protein structures. For the docking programs Gold and Glide, the docking accuracy and the ability of docking scores to predict experimental binding affinity were thoroughly examined from a ligand-centric and single protein-centric point of view. It was found that in terms of docking accuracy, the two docking methods perform comparably well. For a majority of ligands (>50%) the experimental pose could be reproduced within 2 Å RMSD in more than 75% of the protein conformations. For Glide only, a weak correlation between the docking score and log(IC50) was observed.

In this second part of the study, an approach of combining docking results from multiple protein structure focuses on protein flexibility, induced-fit effects, and docking accuracy. The combination of docking results from multiple conformations was probed for possible improvement of the log(IC50)/score correlation for Glide. This approach was further extended by including a 136 ligand test set and scrambling experiments in order to assess the predictivity and the statistical significance of the log(IC50)/Glide score correlations for single and multiple protein structures. Furthermore, the influence of combining multiple protein structure docking results on docking accuracy was compared for Glide and Gold.

This cross-docking study addresses a series of questions: Is there an improvement in docking accuracy when a number of related complexes are analyzed instead of one? If so, does this happen by increasing the sampling of experimental protein conformations? Could it be possible to predict approximate relative binding affinities based upon docking scores? Could it be feasible to identify the predominant factors that permit docking to drive a lead optimization program? Are there any advantages and disadvantages of using cross-docking in a modeling protocol? Can cross-docking reproduce effects of induced-fit and protein flexibility?

METHODOLOGY

The full details describing the data set, ligand and protein preparation, docking parameters, and the result analysis can be found in Part 1 of this study.[3]

The 150 high-resolution in-house crystal structures were aligned to a common reference. All water molecules were deleted, all ligands were individually inspected, and the correct protonation state and tautomer were chosen to reproduce the hydrogen-bonding pattern to the hinge and

* Corresponding author e-mail: johannes.voigt@spcorp.com (J.H.V.), jose.duca@spcorp.com (J.S.D.).
† These authors contributed equally to this work; they should be regarded as joint first authors.
‡ Schering-Plough Research Institute Cambridge.
§ Department of Drug Design, Schering-Plough Research Institute.

10.1021/ci700428d CCC: $40.75
Published on Web 03/07/2008

other active site residues inferred from heavy atom distances. No unusual protonation or tautomeric states were required. Hydrogen atoms were minimized with all heavy atoms fixed (Macromodel v9.0,[4] MMFF94s force field, GB/SA implicit solvation model, PRCG, convergence threshold 0.05). After deleting the ligand, a Glide[4] grid was computed for each of the 150 protein structures (Impact v3.5[4]).

The ligand molecules were subjected to a conformational search (Macromodel,[4] 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.[5] These ligands cover 21 distinct ring systems with the major class having 109 members. Ring systems were determined using the Pipeline Pilot[6] "Generate Fragments" component, with parameters: ring assemblies including exocyclic double bonds.

The validation set contains 136 molecules of the major ligand class for which in-house inhibition data, but no crystal structures, were available. The validation set was chosen using the dissimilarity selection tool as implemented in Sybyl 7.1.[7] 2D Unity fingerprints[7] along with atom-pair fingerprints were defined to carry out the diversity selection, done by hierarchical clustering as implemented in the Selector module.[7] These molecules were prepared for docking in the same fashion as the 150 crystallographic complex ligands.

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 validation set, best scoring pose was considered) was reported.

Five GA runs without early termination for each ligand in Gold v2.2[8,9] were conducted. The active site was defined by a radius of 17 Å around backbone nitrogen Leu83. 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.[12] 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".

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 [$\mathrm{RMSD} = \mathrm{sqrt}(\Sigma[\text{distance atom}_{\text{ref i}} - \text{atom}_{\text{docked i}}]^2/N$), where N is the number of atoms]. Using the CACTVS system (v3.223)[10] the symmetry independent RMSD was calculated for the docked pose of a ligand from the coordinates of crystallographic structure.

Each ligand pose obtained from docking the 140 ligands of the crystallographic set and the 136 molecules of the validation set to the 150 CDK2 protein conformations was characterized by both its accuracy (RMSD of the common core for the validation set ligands) and docking score (more negative Glide scores and higher Gold fitness values both correspond to better binding.

Exhaustive log(IC50)/Score Correlation Evaluation of All Grid Combinations. For each of the 551 300 nonredundant three-grid combinations the ligand poses from each of the three grids with an RMSD value above the chosen cutoff were discarded. The lowest score of the remaining poses for a given ligand was then used for calculation of R, R^2, and the slope of the logIC50/docking score correlation.

Determination of Grid Combinations with Optimum log(IC50)/Score Correlation by Monte Carlo Optimization. There are more than 20 million possible four grid/protein conformation combinations for 150 protein structures. Thus, an exhaustive evaluation of all n-fold protein structure combinations (where $n > 3$) is not feasible, and Monte Carlo optimization was used instead. In the same fashion as described above, for a given protein structure all poses with an RMSD value above the chosen cutoff were discarded, and the score of the pose for the ligand with lowest score was used for the calculation of score/logIC50 correlation (R and R^2). This correlation value was used as the fitness function for the optimization using the following procedure: an initial grid set was randomly chosen and grid set R^2 was evaluated. One grid was randomly chosen to be replaced, and the grid set R^2 was re-evaluated. The change is accepted if it meets the Metropolis criterion: If the new R^2 is greater than the old R^2, then the change is accepted. Otherwise, the change is accepted with a probability of $\exp(-(\Delta R^2)/t)$ with $t = 0.001$.

The process was repeated 25 000 times. The entire Monte Carlo procedure was repeated 10 times (each time from a different starting point). The 10 best grid sets found during this entire process were saved.

Exhaustive Evaluation of All Grid Combinations for Docking Accuracy. There are 551 300 nonredundant three-grid combinations of the 150 grids. For each three-grid combination, the number of ligands whose pose fit the experimental one within a 2 Å RMSD-cutoff to the experimental ligand was determined for each of the three protein structures. **S** was defined as the maximum of these for the best performing single protein structure. Subsequently, for each ligand, the pose with the best docking score was selected from the three structures and tested against the RMSD-cutoff. The total number of ligands meeting this criterion was defined as **C**, the docking accuracy of the grid combination. The difference **C**−**S** gives the increased number of ligands which fit the combination versus the best single structure.

RESULTS AND DISCUSSION

Multiprotein Structure: log(IC50)/Docking Score Correlation. From Part 1 of this study[3] it seems clear that high docking accuracy was achieved using a number of crystallographic structures, while potency predictions were inaccurate. These results are in agreement with previous work[11] for single grid experiments. One of the goals of this paper was to assess if a poor log(IC50)/score correlation could be due to the intrinsic flexibility of CDK2 and the induced-fit effects of different ligands. Therefore, a series of detailed

Table 1. R^2 between log(IC50) and Docking Scores for Grid 1−10 Combinations and Several RMSD-Cutoff Values[a]

no. of grids	cutoff 1.0 Å	cutoff 1.5 Å	cutoff 2.0 Å	no cutoff
1	0.52 (8)	0.44 (59)	0.35 (89)	0.30 (140)
2	0.55 (35)	0.42 (79)	0.37 (110)	0.38 (140)
3	0.60 (44)	0.45 (93)	0.43 (113)	0.42 (140)
4	**0.61 (49)**	0.47 (103)	0.46 (121)	0.46 (140)
5	0.60 (50)	0.48 (104)	0.47 (120)	0.48 (140)
10	0.60 (73)	0.52 (114)	0.50 (128)	0.50 (140)

[a] The number of compounds that meet the RMSD criterion is shown in parentheses.

cross-docking experiments was designed and performed. We reason that it is possible that several structures to which a given ligand docks properly could score the ligand pose differently due to induced-fit effects. By exposing the properly docked ligand to many protein environments, the chances of identifying those interactions that are rewarded by the scoring function[12] increase.

For multiple-protein structure cross-docking experiments, each ligand was analyzed as explained in the Methods section.

Since the number of grid permutations grows quite rapidly with n (e.g., when $n = 2$ is 11 325, $n = 3$ is 551 300, and when $n = 4$ is 20 811 575), a Monte Carlo approach was designed to efficiently carry out the aforementioned multigrid docking protocol.

Monte Carlo experiments were compared and validated against the corresponding full three-grid exhaustive enumerations.

Table 1 summarizes the Monte Carlo results for multigrid cross-docking experiments carried out at different RMSD-cutoff values. The best multigrid runs are compared to the best single-grid docking experiments, listed in the first row of Table 1. R^2 increases with the increasing number of grids and decreases for larger RSMD cutoff values. The compounds that docked better, at an RMSD-cutoff value of 1 Å, showed considerably better correlations than the compounds at RMSD cut-off values of 1.5 Å or greater. Moreover, the multiprotein structure cross-docking runs yield mostly better correlation values than single-grid runs. Table 1 indicates that the combination of four grids at an RMSD-cutoff of 1 Å gives the best correlation: 49 compounds (35% of the data set) are docked within the cutoff, and the log(IC50)/score correlation explains 61% of the experimental IC50 data.

To apply these promising results to a real lead-optimization program would not be trivial, since the percentage of high correlation grid-combinations is quite small and difficult to predict in advance. Moreover, since the trends in correlation coefficients seem to agree with a multidimensional induced-fit effect, there is no guarantee that the necessary structural information would be available at a given time for a given project.

Moreover, a number of trends in Table 1 put into question the significance of high correlations observed at 1 Å. For example, the correlation drops quite rapidly over all permutations from 1 to 1.5 Å in the RMSD-cutoff, while the number of compounds that satisfy these criteria increase at least 50%. Are these correlations due to a smaller number of compounds meeting the 1 Å RMSD criteria?

To understand the source of these multiple-grid log(IC50)/score correlations, further validation was deemed necessary.

Validation Experiments. A validation set containing 136 analogs from one series present in the initial training set was extracted from our CDK2 compound database. Structural diversity was used to select these compounds with an activity spread of 3.5 log(IC50) units. As in the initial set of crystal structures, all IC50 values were determined in the same laboratory over a short period of time. While there is no crystallographic structural data on the validation set we can assume the same binding mode at least for the main core, due to the structural similarities of the common core present in several structures of the initial data set. Therefore, for the validation set, the RMSD will only refer to the common core. Since the common core is smaller than the whole inhibitors, the RMSD-cutoffs will not be as stringent as in our previous examples.

The validation experiments were carried out in the following manner: a) The top *n*-fold grid combinations at a given RMSD-cutoff (for this section we refer to this RMSD-cutoff as *cutoff a*) were identified from Monte Carlo runs; b) The validation data set ligands were docked to our initial data set of 150 structures; c) The *n*-fold grid combinations at the original *cutoff a* were applied to the validation docking poses, and the correlation between docking score and log(IC50) values was determined at the newly defined RMSD-cutoff value (*cutoff b*), which refers to the common core these ligands contain. The *cutoff b* values are 1.0, 1.5, and 2.0 Å.

Table 2 contains the results from the validation experiments. There was no correlation signal between the original and validation results. For those promising correlations found at a *cutoff a* of 1.0 Å, the best validation correlation was $R^2 = 0.25$ at a core *cutoff b* of 1.0 Å and a three-grid combination. The validation experiments strongly suggest that the correlations observed in Table 1 are not linked in a consistent way to docking score performance. In light of these results and the previous correlation found for a limited scrambled data set (R^2 near 0.22), it was decided to scrutinize further by scrambling experiments.

Data Scrambling. Traditionally, a limited number of scrambled data sets have been used to verify that QSAR/QSPR models with high predictive values are not generated randomly. Recent QSAR results were published[13] in which 1000 scrambled data sets were considered significant to validate the derived QSAR models.

Using the initial data set of 140 ligands we generated 5000 scrambled data sets and monitored the random function used to generate these data sets as well as the homogeneity of the randomization. The data sets were designed to be distinct, and the correlation between the scrambled log(IC50) values versus the real experimental data set was monitored and kept very low.

The Python implemented random() function (and the bookkeeping function choice()) was chosen for the pseudo-randomization of the assignment between the labels of the ligands and the IC50 values (Mersenne Twister's based method).[14]

Scrambling Experiments. Figure 1 shows the correlation slope distribution for 5000 scrambled data sets, after an RMSD-cutoff value of 1.5 Å was applied. To analyze the homogeneity of the scrambled data sets the distribution of positive and negative correlations for each scrambled set among all 150 grids was considered, and the percentage of

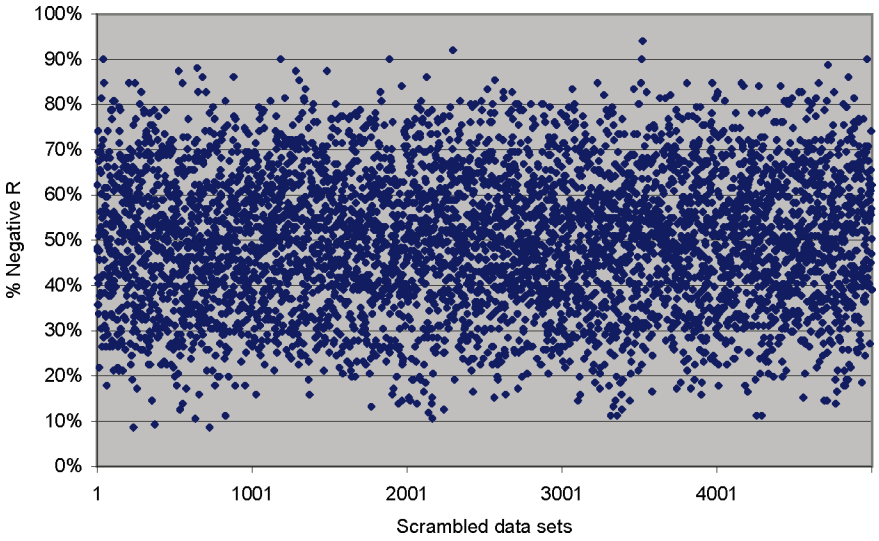

Figure 1. Percentage of negative R over 150 structures per scrambled data set.

Table 2. Cross-Docking Experiments on the Validation Data Set: Correlation between the Glide Docking Score and log(IC50) at Different RMSD-Cutoff Values[a]

	cutoff 1.0 Å[a]			cutoff 1.5 Å[a]			cutoff 2.0 Å[a]			no cutoff[a]		
grids	1.0 Å	1.5 Å	2.0 Å	1.0 Å	1.5 Å	2.0 Å	1.0 Å	1.5 Å	2.0 Å	1.0 Å	1.5 Å	2.0 Å
1		**0.52**			**0.44**			**0.35**			**0.30**	
	0.00	*0.03*	*0.06*	*0.26*	*0.19*	*0.23*	*0.26*	*0.19*	*0.22*	*0.38*	*0.29*	*0.29*
2		**0.55**			**0.42**			**0.37**			**0.38**	
	0.10	*0.18*	*0.18*	*0.17*	*0.15*	*0.19*	*0.15*	*0.17*	*0.16*	*0.38*	*0.33*	*0.33*
3		**0.60**			**0.45**			**0.43**			**0.42**	
	0.25	*0.21*	*0.21*	*0.13*	*0.16*	*0.16*	*0.19*	*0.17*	*0.17*	*0.30*	*0.35*	*0.39*
4		**0.61**			**0.47**			**0.46**			**0.46**	
	0.24	*0.23*	*0.23*	*0.25*	*0.28*	*0.28*	*0.17*	*0.23*	*0.17*	*0.35*	*0.34*	*0.34*
5		**0.60**			**0.48**			**0.47**			**0.48**	
	0.17	*0.16*	*0.11*	*0.31*	*0.33*	*0.33*	*0.17*	*0.23*	*0.23*	*0.36*	*0.36*	*0.36*
10		**0.60**			**0.52**			**0.50**			**0.50**	
	0.22	*0.22*	*0.22*	*0.26*	*0.32*	*0.32*	*0.25*	*0.23*	*0.23*	*0.30*	*0.23*	*0.18*

[a] The original correlation is kept in boldface for reference at the cutoff, *cutoff a*. The three RMSD criteria used for the validation set (*cutoff b* 1.0, 1.5, and 2.0 Å) are listed for each *n*-grid combination.

negative correlation values was plotted per data set. For the experimental data set there are 10% of the grids with a negative R or slope. If the scrambled data sets were perfectly random with continuous values, one would expect that the percentage of negative slopes should form a distribution around 50%. However, in the case of an experimentally determined data set, with finite values and not necessarily a homogeneous distribution of IC50 values, the correlation slope distribution should be random depending of the ability of the random() function to avoid periodicity.

Figure 1 presents a relatively homogeneous pattern, and no evidence of periodicity was found, which reinforces the concept that the 5000 scrambled data sets are independent of each other and randomly scrambled. Figure 1 is consistent with a homogeneous Gaussian-like distribution of negative correlations. The same is the distribution of positive correlations (data not shown).

Figure 2 is a different representation of the data shown in Figure 1. In this case the maximum, minimum, and mean correlation values are plotted, after sorting by mean correlation. Once again, homogeneous scrambling is manifested in the mean correlation, which presents a symmetric distribution with a zero intersection at the 2500th data set.

The experimental data set has the following correlation distribution over the 150 crystallographic structures: minimum R = −0.30, maximum R = 0.66, and mean R = 0.24.

An interesting observation from Figure 2 is that a small fraction of the scrambled data sets (39/5000 = 1% with |R| > 0.66) outperforms the real data set. These results support the concept of random selection being the most important factor behind the log(IC50)/score correlation values obtained for single-grid experiments. In multiprotein structures cross-docking experiments random fluctuation should have an increased influence on the log(IC50)/score correlations.

The plot in Figure 3 compares the correlations for the experimental activity data to the distribution for the 5000 scrambled data sets. The mean R is close to zero as expected. The minimum and maximum indicate the range for R observed for the scrambled data sets, while 2.58 times the standard deviation of R over the 5000 sets gives the 99% confidence interval. 55 (37%) of the 150 grids have a correlation R for the real activity data which is outside the 99% confidence interval for the R observed for the random data sets. Thus, one can conclude that the correlation for score/logIC50 is statistically significant only for 37% of the grids.

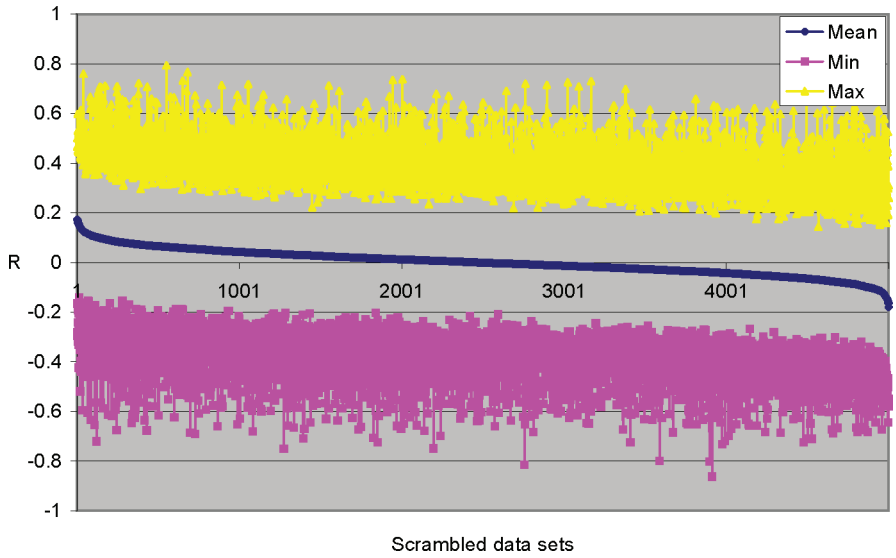

Figure 2. Minimum, maximum, and mean R for 150 structures per scrambled data set.

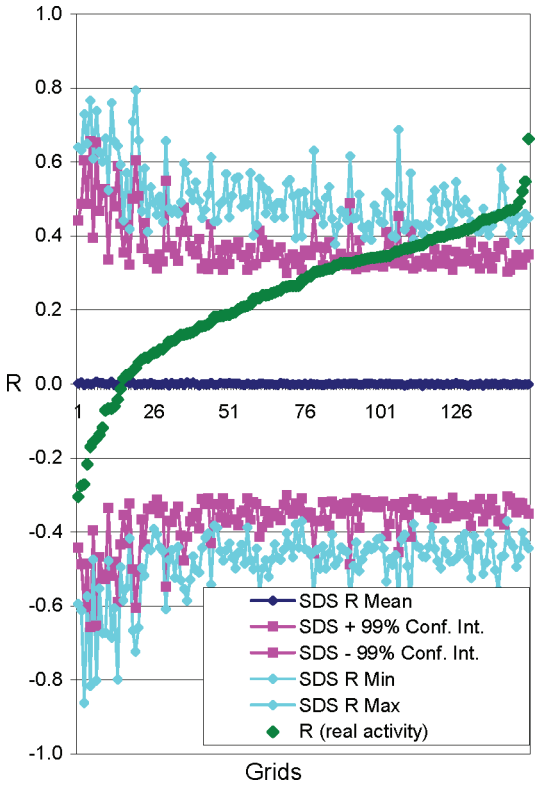

Figure 3. Plot of R for real activity data, and of minimum, maximum, mean, and upper (+) and lower (−) boundaries of the 99% confidence interval of the 5000 scrambled data sets (SDS) for each grid. The grids are sorted by R for the real data.[15]

In Figure 4, the exhaustive generation of correlation slope distributions is depicted for three-grid cross-docking at an RMSD-cutoff of 1.0 Å. The best three-grid combination for the experimental data set yielded a high correlation of $R^2 = 0.60$. Out of 5000 scrambled data sets 205 (4.1%) had three-grid combinations with better R values than the real data. Thus, applying the 99% confidence interval as above for the single grids, the R^2 of 0.60 observed for the real data is likely not statistically significant.

These scrambling experiments for single protein structures and three-structure combinations showed that the observed correlations with the real data are not statistically significant except for 37% of the single structures. This result is in accord with the lack of predictivity which was observed for the 136 ligand validation set.

Multiprotein Structure − Docking Accuracy. After it was probed how combining the docking results to multiple protein conformations affected the log(IC50)/score correlation, it was investigated if multistructure docking could improve the docking accuracy, which was demonstrated in Part 1 of this study[3] to be high. For an RMSD-cutoff of 2 Å, more than 55% of the ligands were properly docked in 75% of the protein structures. It was recently demonstrated[2] that taking the pose with the lowest docking score from poses for the same ligand derived from docking to multiple protein conformations tends to increase the number of correctly docked ligands. Seven p38 kinase and thirteen CDK2 structures were considered, and the combined docking accuracy was reported only for a few of the possible three-structure combinations. That might not be enough data to arrive at statistically significant conclusions.

In this study, with a significantly larger number of crystallographic structures available, the percentage of correctly docked ligands at an RMSD-cutoff of 2 Å was calculated for all possible three-protein structure combinations, as described in the Methodology section. The difference **C**−**S** (Δ**C**−**S**) between the docking accuracy for the best single protein structure (number of ligands **S**) and the accuracy derived by taking the pose with the best score among the three structures (number of ligands **C**) was also calculated. Δ**C**−**S** is positive if the docking accuracy of a given protein structure combination is improved over that of the best single structure.

Combinations of a maximum of three protein structures can be exhaustively analyzed as explained in the previous section. To better understand the results, an analysis of all the combinations is necessary. From the aforementioned log(IC50)/score correlations derived from multiple structures, it was found that just considering the best combination can be misleading.

Glide. Figure 5 shows for an RMSD-cutoff of 2 Å the comparison of the distributions of percentages of ligands docked correctly for single grids versus three-grid combinations. For single grids the distribution includes the 150 grids,

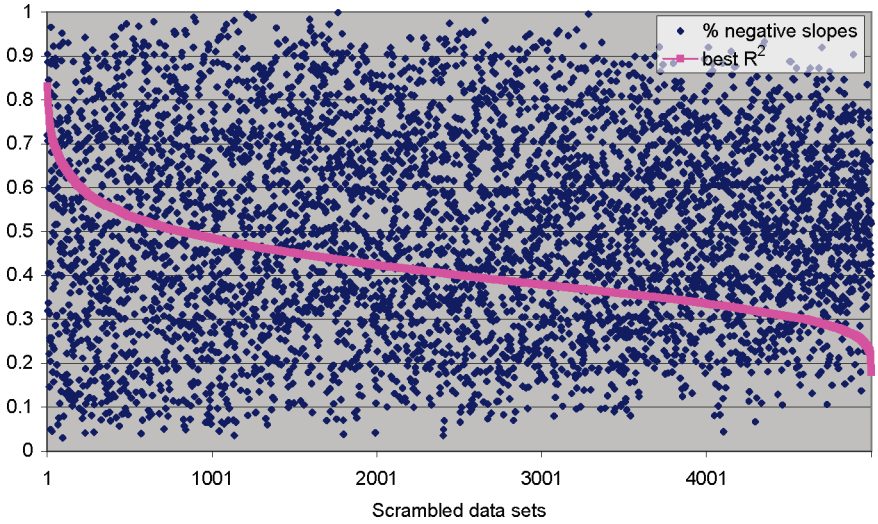

Figure 4. Percentage of grids with negative slope and best R^2 for 5000 scrambled data sets at an RMSD-cutoff of 1 Å for all three-grid combinations.

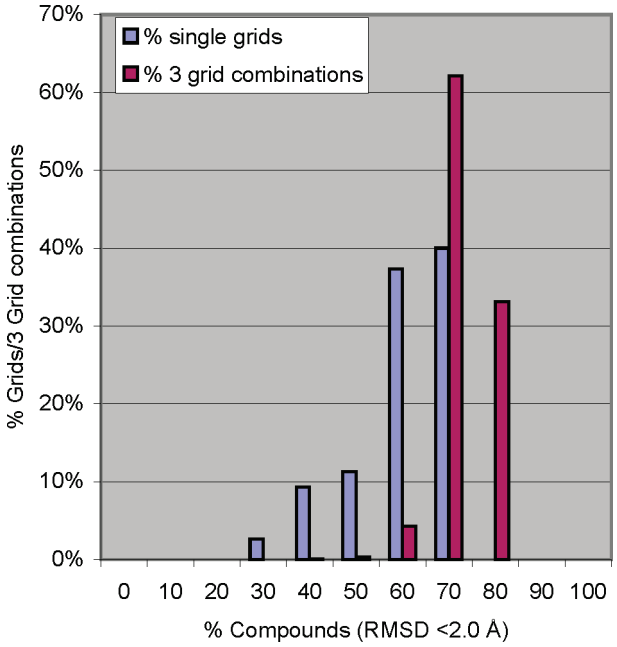

Figure 5. Glide - comparison of distributions of single-grid versus three-grid combinations having a given percentage of ligands docked correctly (RMSD-cutoff 2 Å; centered histogram[16]).

while for three-grid combinations it encompasses the 551 300 unique combinations. For the three-grid combinations the distribution is shifted toward higher percentages of ligands docked correctly. 95% of the grid combinations have 65% or more of the ligands docked correctly in comparison to 40% for single grids with 65% or more of the ligands docked correctly. For the same RMSD-cutoff, Figure 6 shows the distribution of $\Delta\mathbf{C}-\mathbf{S}$ for all possible three-grid combinations of 150 grids for Glide. Consistent with the distribution of the total percentage of correctly docked ligands in a majority of the cases (90%), combining docking results from the grids improves the accuracy by up to 36 correctly docked ligands for one case. The bar for 36 ligands is too small to be visible in Figure 5. The average improvement is 7 ligands (5%). These results might be encouraging and could be explained by the fact that scoring functions are in particular optimized to distinguish between "right" and "wrong" poses of the same ligand within the same binding site/protein conformation. Extending this trait by observing that the scoring function could distinguish between "correct" and "incorrect" poses of the same ligand across multiple conformations of the same protein might be valid. Alternatively, the reason for docking accuracy improvement could be enhanced sampling, since each ligand is being docked three times.

In order to probe the predictivity of this approach, the 140 ligands were randomly split 10 times into "training/test" sets of *x*/*y* compounds. For a given three-grid combination, C_x-S_x for the training set and C_y-S_y for the test set were calculated. If C_x-S_x and C_y-S_y are both greater than 0, then accuracy improvement by this grid combination is correctly predicted. The individual results for the 10 runs were averaged for *x*/*y* ratios of 100/40 and 75/65. In both cases, 64% for $x/y = 100/40$ and 66% for $x/y = 75/65$, the majority of grid combinations had improved docking accuracy for both the training and the test set. The change from 90% combinations with docking accuracy improvement observed for the full 140 ligand set can be explained by a shift of the $\mathbf{C}-\mathbf{S}$ distributions down to approximately 80% improved combinations for the subsets (graphs not shown). Apparently for fewer ligands the likelihood of beneficial combinations decreases. Thus the likelihood of a grid combination to have a positive $\Delta\mathbf{C}-\mathbf{S}$ for both training and test sets is $80\% \times 80\% \cong 64\%$.

In light of the fact that even for single grid docking experiments Glide docks ~100 ligands correctly (2 Å RMSD-cutoff) for many grids (60) the mean improvement of 7 (5%) ligands for the whole set is significant but not large. Furthermore, a ~65% predictivity for docking improvement by grid combinations shows that this method might not be robust enough for "real life" situations where the correct docking pose is not known.

Gold. The same comparison of distributions of the percentage of ligands docked correctly at an RMSD-cutoff of 2 Å for a single protein structure versus combinations of three-protein structures (Figure 7) shows little or no improvement of the docking accuracy with combinations of three protein structures. Figure 8 shows for Gold the $\Delta\mathbf{C}-\mathbf{S}$ distributions for all three protein structure combinations. This histogram demonstrates more drastically the difference to

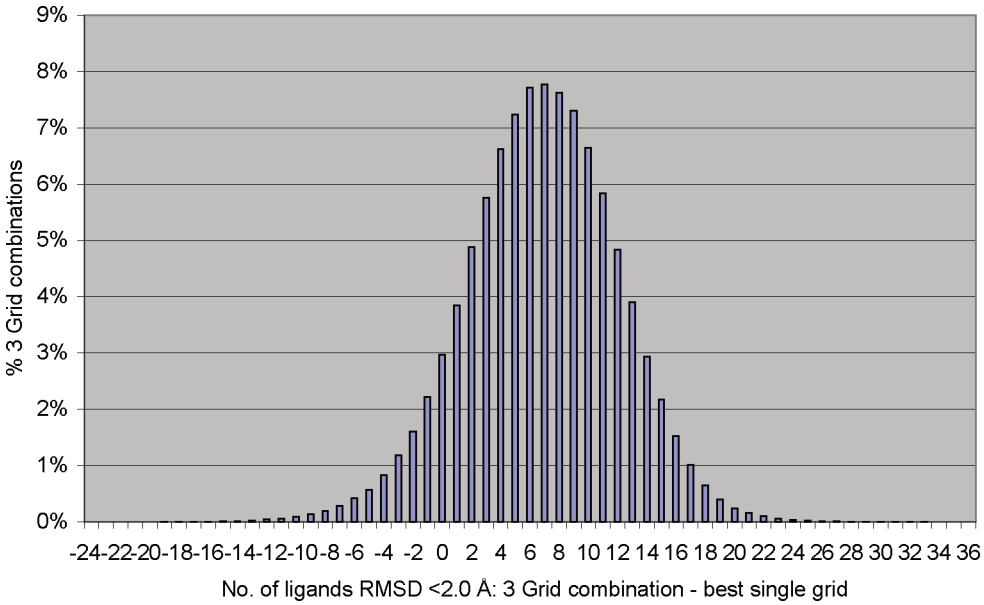

Figure 6. Glide - distribution histogram $\Delta\mathbf{C}-\mathbf{S}$ for the **C** (number of correctly docked for grid combination) − **S** (number of correctly docked ligand for best single grid) (RMSD-cutoff 2 Å).

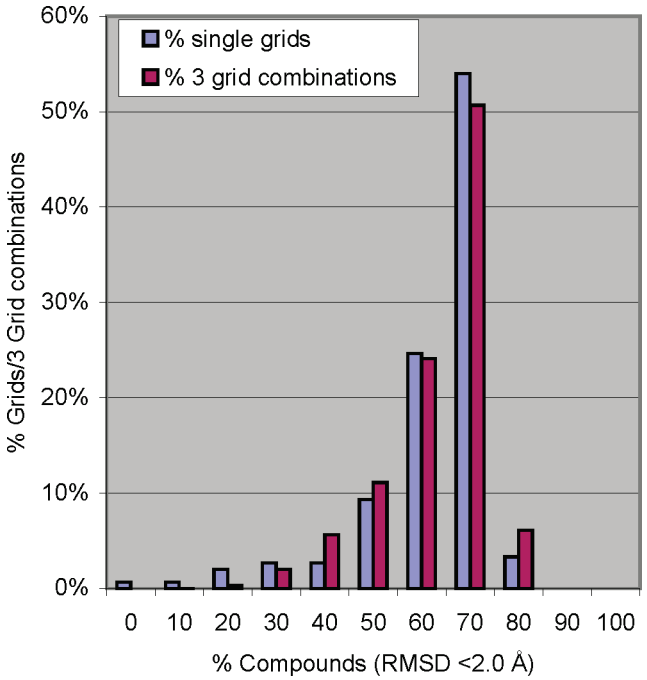

Figure 7. Gold - comparison of distributions of single- versus three-protein structure combinations having a given percentage of ligands docked correctly (RMSD-cutoff 2 Å; centered histogram[16]).

what was observed for Glide. Surprisingly only 21% of combinations display a better docking accuracy for the combination of three protein structures over the best single protein structure docking accuracy. One reason is that for Gold there are 26 more structures than for Glide with a high docking accuracy of 65% or more ligands being docked correctly at an RMSD-cutoff of 2 Å for single protein structures (see Figure 8). Thus, because the single protein structure docking accuracy is already high for Gold, it is more difficult to obtain a better performance for the combination. Also striking is the very long negative "tail" of the distribution for Gold, while the Glide distribution shown in Figure 6 is more symmetrical. Closer investigation of the combinations representing this tail revealed the effect of several protein structures with very poor docking accuracy. These had comparatively better scores for the incorrectly docked ligands than the score for the same ligand docked correctly to a different structure. To analyze this effect for Gold, the average RMSD and average score over all ligands docked to this structure was computed for all protein structures. Point **A** in Figure 9a is an example for a protein structure with an overall poor docking accuracy but a high average score, while point **B** is a protein structure yielding overall correctly docked ligands but having a low score. (The higher the Gold score/fitness the better the pose ranks.) Combinations of structures including these two would predominantly select the wrong poses from structure **A** over the mostly correct ones from structure **B**, and this combination would be in the long negative tail of the distribution shown in Figure 8. Figure 9a also shows that for Gold there is no correlation between the average RMSD and the average score. Protein structures with more ligands docked correctly are not rewarded by a better average score.

This indicates that though the Gold scoring function performs very well at distinguishing the correct from the incorrect poses for the same ligand within the same protein conformation, the score cannot be compared across multiple structures.

Figure 9b shows that for Glide there is a significant correlation of $R^2 = 0.57$ between the average RMSD and the average score as one would expect from the observed docking accuracy improvement for grid combinations. That means that grids with more ligands docked correctly are rewarded by a better average score. (The lower the Glide score is the better the pose ranks.)

The Glide scoring function was adjusted and fitted to reproduce the binding affinities for the ligands of 198 crystallographic complexes.[12] As an unintended result, the Glide scoring function appears to be "scaled" or "normalized", which allows comparison of the score for one and the same ligand across multiple grids.

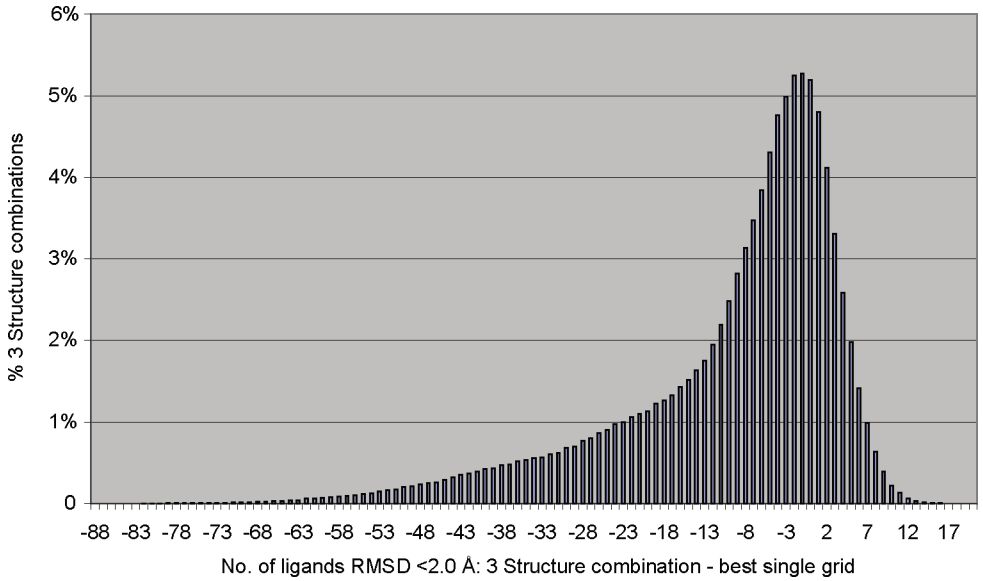

Figure 8. Gold - distribution histogram $\Delta\mathbf{C}-\mathbf{S}$ for the **C** (number of compounds correctly docked for a protein structure combination) − **S** (number of ligands correctly docked for best single protein structure), at an RMSD-cutoff of 2 Å.

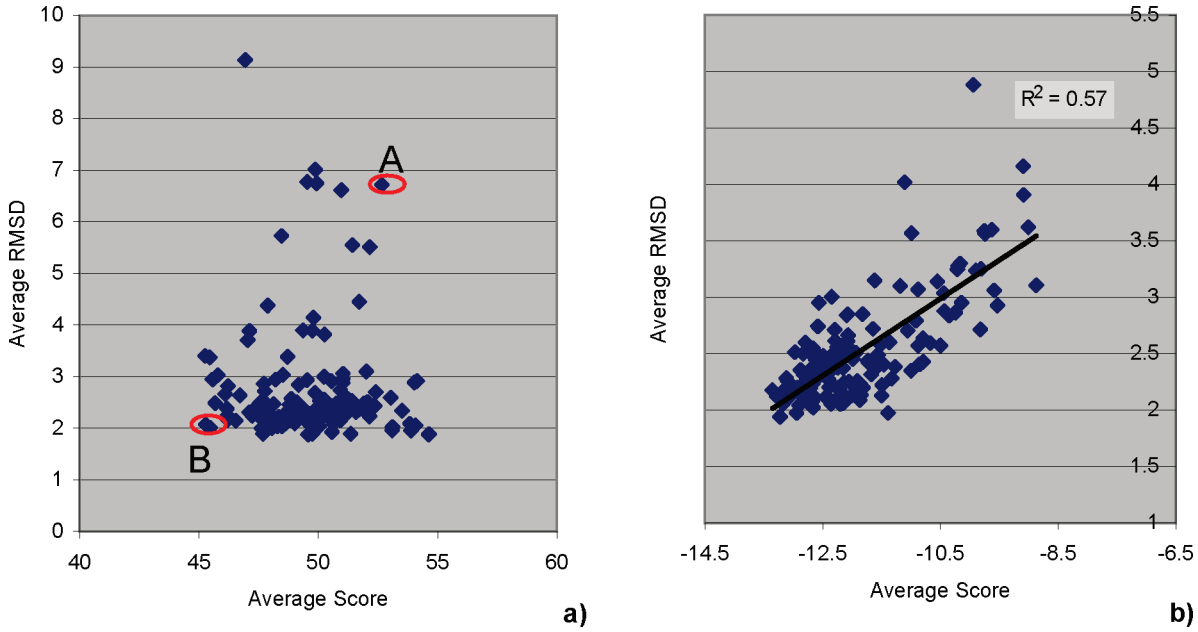

Figure 9. Average RMSD versus average score per grid a) Gold and b) Glide.

Can the Average Score or the Average RMSD per Grid Predict which Grid To Use for Docking? The correlation between the average RMSD and the average score per grid for Glide could be used to identify the best single grid or grids to use for docking. In order to probe the robustness of this trend the ligand set was split 10 times into training sets of 75 and test sets of 65. For each grid the average RMSD values and the average scores for training and test sets were calculated, and the correlation of the training set average RMSD and the test set average RMSD was calculated (see Table 3) Furthermore, the correlation of the training set average score and the test set average RMSD was calculated.

The high average R^2 of 0.74 for the training set average RMSD and the test set average RMSD suggests that in cases where several crystallographic structures are available one could perform the matrix cross-docking experiment, calculate the average RMSD values for all cognate ligands, and then choose the grid(s) with the best RMSD value(s) for docking of new compounds. Coincidently, this is one way in which an experienced modeler could use extensive crystallographic data in a lead optimization program.

Table 3. Correlation of the Training Set Average RMSD and the Test Set Average RMSD and Correlation of the Training Set Average Score and the Test Set Average RMSD for 10 Different 75/65 Splits and Average

run	R^2 [average RMSD(train) − average RMSD(test)]	R^2 [average score(train) − average RMSD(test)]
1	0.67	0.33
2	0.68	0.37
3	0.69	0.41
4	0.72	0.45
5	0.75	0.46
6	0.76	0.48
7	0.76	0.54
8	0.78	0.58
9	0.78	0.58
10	0.78	0.65
average	**0.74**	**0.48**

The observed R^2 (average RMSD/average score) of 0.56 for all 140 ligands is confirmed by the average R^2 (10 splits) of 0.48 for the the training set average score and the test set average RMSD. One could take advantage of this relationship in hypothetical scenarios with multiple protein conformations for a target but no structure of a bound ligand. Here the set of ligands for which the binding modes are to be elucidated could be docked to the panel of crystal structures, and the crystal structures with the lower average score should give a higher percentage of correctly docked ligands.

Whether these average RMSD/average score trends are specific to this particular CDK2 data set where induced-fit effects can play a significant role needs to be further investigated. If not, then these trends can be exploited for other kinases or general targets.

SUMMARY

There are several fundamental questions that scientists face every time they perform a docking experiment of a noncognate compound into a protein complex, especially when multiple crystal structures for this are available: 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? e) Should the docking results from multiple protein structures be combined?

The first part of this study focused on docking results and performance for single protein structures, while the present paper addressed in particular questions c) through e) and scrutinized the statistical significance of log(IC50)/score correlations important for a) and b).

Using docking results from multiple protein structures yields a docking accuracy improvement only for Glide. Though 90% of the three-grid combinations show an improvement of docking accuracy over the best single grid, cross-validation experiments with subsets showed a lower predictivity. The observation that Gold multistructure docking results did not robustly improve the accuracy over that of single structures led to the conclusion that the Gold scoring function should not be applied to multiple structures for the same ligand, while the Glide scoring function allows this. Apparently, the fitting to reproduce affinity for the Glide scoring function had the consequence that it is scaled or normalized. This is expressed in the high average RMSD/average score correlation over all ligands for all grids. For Glide, the average score over multiple ligands could suggest which protein structures to choose.

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

For the multiple-grid cross-docking experiments we have established that random choice is responsible for high correlations obtained for three- and four-grid experiments (see Table 1). These combinations did not have high correlations when applied to a blind validation set. Moreover, extensive scrambling experiments clearly indicate that correlation values such as the ones presented in Table 1 can be reproduced or improved using scrambled data. Furthermore, it was demonstrated that for Glide almost 10% of the grids produced negative correlations for correctly docked ligands (RMSD <1.5 Å) and that only 37% of protein structures at this cutoff display a statistically significant correlation (99% confidence interval), a strong indication that random choice is dominating even single-grid docking experiments.

The dangers of overfitting must not be overlooked in cross-docking experiments. Moreover, overfitting can also be present in similar methodologies commonly adopted by computational chemists, such as consensus scoring, for example. The use of min or max functions in addition to other variables (fudge factors to adjust different molecular weights or sizes, etc.) to improve the docking score could seem reasonable from an intuitive viewpoint, but as it has been shown in this paper, it could also lead to useless results tainted by overfitting.

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.

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

(3) Duca, J. S.; Elkin, C.; Madison, V. S.; Voigt, J. H. Cross docking of inhibitors into CDK2 structures. 1. *J. Chem. Inf. Model.* **2008**, *48*, 659−668.

(4) Schrödinger, LLC, New York, 2007.

(5) 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.

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

(7) Tripos International, 1699 South Hanley Rd., St. Louis, MO 63144, U.S.A.

(8) 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.

(9) 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.

(10) 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.

(11) 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.

(12) (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.
(13) Iyer, M.; Zheng, T.; Hopfinger, A. J.; Tseng, Y. J. QSAR Analyses of Skin Penetration Enhancers. *J. Chem. Inf. Model.* **2007**, *47*, 1130−1149.
(14) (a) Matsumoto, M.; Kurita, Y. Twisted GFSR generators. *ACM Trans. Model. Comput. Simul. (TOMACS)* **1992**, *2*, 179−194. (b) Matsumoto, M.; Nishimura, T. Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator. *ACM Trans. Model. Comput. Simul.* (*TOMACS*) **1998**, *8*, 3−30.
(15) Spiegel, M. R.; Stephens, L. J. *Schaum's Outline of Theory and Problems of Statistics*, 2nd ed.; McGraw-Hill: 1999; p 206.
(16) All histograms are centered, which means, e.g., a bin labeled with 10% contains items from 5−15%.

CI700428D