JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

CSAR Benchmark Exercise 2011–2012: Evaluation of Results from Docking and Relative Ranking of Blinded Congeneric Series

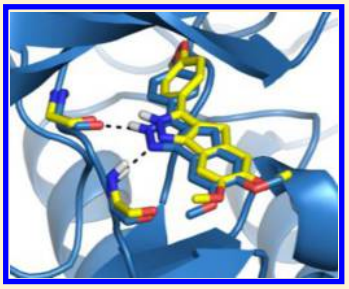
Kelly L. Damm-Ganamet,[†] Richard D. Smith,[†] James B. Dunbar, Jr.,[†] Jeanne A. Stuckey,[‡] and Heather A. Carlson*[,†]

[†]Department of Medicinal Chemistry, University of Michigan, Ann Arbor, Michigan 48109-1065, United States

[‡]Life Sciences Institute and the Department of Biological Chemistry, University of Michigan, Ann Arbor, Michigan 48109-2216, United States

Supporting Information

ABSTRACT: The Community Structure–Activity Resource (CSAR) recently held its first blinded exercise based on data provided by Abbott, Vertex, and colleagues at the University of Michigan, Ann Arbor. A total of 20 research groups submitted results for the benchmark exercise where the goal was to compare different improvements for pose prediction, enrichment, and relative ranking of congeneric series of compounds. The exercise was built around blinded high-quality experimental data from four protein targets: LpxC, Urokinase, Chk1, and Erk2. Pose prediction proved to be the most straightforward task, and most methods were able to successfully reproduce binding poses when the crystal structure employed was co-crystallized with a ligand from the same chemical series. Multiple evaluation metrics were examined, and we found that RMSD and native contact metrics together provide a robust evaluation of the predicted poses. It was notable that most scoring functions underpredicted contacts between the hetero atoms (i.e., N, O, S, etc.) of the protein and ligand. Relative ranking was found to be the most difficult area for the methods, but many of the scoring functions were able to properly identify Urokinase actives from the inactives in the series. Lastly, we found that minimizing the protein and correcting histidine tautomeric states positively trended with low RMSD for pose prediction but minimizing the ligand negatively trended. Pregenerated ligand conformations performed better than those that were generated on the fly. Optimizing docking parameters and pretraining with the native ligand had a positive effect on the docking performance as did using restraints, substructure fitting, and shape fitting. Lastly, for both sampling and ranking scoring functions, the use of the empirical scoring function appeared to trend positively with the RMSD. Here, by combining the results of many methods, we hope to provide a statistically relevant evaluation and elucidate specific shortcomings of docking methodology for the community.



INTRODUCTION

Structure-based drug design (SBDD) is a valuable technology that is seeing increased utilization to advance the process of drug discovery research.[1−6] A typical docking protocol is comprised of two components: the search algorithm and scoring function. An exhaustive search algorithm would account for all possible binding poses by allowing both the protein and ligand to be fully flexible; however, although ligand flexibility can be accurately reproduced, replicating the innumerable degrees of freedom of a protein is impractical due to the enormity of the conformational space that must be searched. Developing methods that incorporate protein flexibility in a computationally tractable manner has been recognized as a means to improve SBDD techniques.[2,7−13] The scoring function is used to evaluate and rank each pose by predicting the binding affinity between the ligand and protein. Many simplifications and assumptions are made to the scoring function to increase its speed, such as neglecting entropy and solvation, but these result in a loss of accuracy. Scoring function development is also an active area of SBDD research.[14−18]

Benchmark Docking Exercises. In order to facilitate the development of docking software, the Community Structure–Activity Resource (CSAR) center was funded by the National Institutes of Health (NIH) in 2008 to increase the amount of high quality experimental data publicly available for development, validation, and benchmarking of docking methodologies. CSAR conducted its first benchmark exercise in 2010 with the goal of (1) evaluating the current ability of the field to predict the free energy of binding for protein–ligand complexes and (2) investigating the properties of the complexes and methods that appear to hinder scoring.[19,20] This exercise illuminated that scoring functions are still not able to successfully predict binding affinity and, hence, are not capable of correctly rank ordering ligands.[20] Additionally, the size of the ligand did not appear to affect the scoring quality, but hydrogen bonding and torsional strain were found to be significantly different between well-scored and poorly scored complexes. Detailed results from most participants can be found in a special issue of the Journal of Chemical Information and Modeling [*J. Chem. Inf. Model.*

Special Issue: 2012 CSAR Benchmark Exercise

Received: January 11, 2013
Published: April 2, 2013

   dx.doi.org/10.1021/ci400025f | *J. Chem. Inf. Model.* 2013, 53, 1853–1870

2011, *51*, (9), 2025]. Previously in 2007, Nicholls and Jain organized a Docking and Scoring Challenge with an emphasis on developing standards for evaluation of methods, data set preparation, and data set sharing.[21] A second Challenge was conducted in 2011, where it was found that GLIDE[22,23] and Surflex-Dock[24,25] outperformed other methods tested in both pose prediction and virtual screening (enrichment).[26] A prevalent theme that emerged from the various participants was that optimizing the protein structures prior to docking improved performance.[27−29] Special issues of the Journal of Computer-Aided Molecular Design were dedicated to both the 2007 competiton [*J. Comput.-Aided Mol. Des.* **2008**, *22*, (3−4), 131] and 2011 competition [*J. Comput.-Aided Mol. Des.* **2012**, *26*, (6), 675]; detailed evaluations from participating groups can be found there. Moreover, OpenEye periodically runs the SAMPL Experiment to assess additional aspects of computational modeling relevant to SBDD such as prediction of vacuum−water transfer energies, binding affinities of aqueous host−guest systems, prediction of solvation energies, tautomer ratios, etc.[30−32]

Various groups have conducted independent evaluations of docking programs and have found that many search routines are capable of predicting the native binding pose of the ligand within a RMSD of 2 Å for a range of protein targets.[6,33] Furthermore, while not able to predict binding affinity well, current methods have proven to be successful at enriching hit rates (i.e., identifying active molecules from decoys).[6,33,34] However, consistently ranking inhibitors with nM-level affinity over those with μM-level affinity has proven to be a challenge, as is identifying "activity cliffs" where small changes result in significant increases or decreases in affinity. Most methods appear to do well at either pose prediction or enrichment; only a few are capable of successfully performing both.[33] A further caveat is that expert knowledge is necessary as small modifications to the software's parameters can have large effects on the docking results.[33,35]

A review by Cole et al. discusses how assessing and comparing the performance of docking software is a difficult task and can be misleading as one is not always comparing apples to apples.[35] An additional study found that the quality of the crystal structures in publicly available data sets can affect docking results; poor resolution structures led to unsuccessful docking and vice versa.[36] To that end, it has become increasingly clear that one major limitation to the field was a large, standardized, high quality data set of experimentally determined protein−ligand complexes.[27,37−39] Furthermore, computational chemists need reliable experimental data with the complexes such as binding affinity, solubility, pK_a, and logP/logD of the ligand. The CSAR center was created to fulfill this need, and details of our high quality data sets can be found in our data set paper in the same CSAR special issue of the Journal of Chemical Information and Modeling, along with a review of other publically available data sets.[40]

Assessment. In addition to high quality data, proper evaluation protocols are imperative to assess the performance of the docking methodology.[20,21,41] Pose prediction of the native or cognate ligand is a common approach for evaluating the search algorithm. Recently, this practice has been called into question; however, we must take a step back and recognize that while this task may not be performed regularly for drug discovery purposes, it is an essential positive control in the research lab. Cross-docking exercises are more relevant as they are the actual application of the docking software and should be conducted once it is confirmed that the method is capable of reproducing native binding poses. A variety of measures exist for evaluating pose predictions: RMSD (root-mean-square deviation), DPI (data precision index)/RDE (relative displacement error),[41−43] number of native contacts predicted,[44−46] RSR (real space R) and RSCC (real space correlation coefficient),[47] and coordinate error.[39]

RMSD is the standard for evaluating poses, but it can be misleading. Crystal structures are simply a static snapshot of the protein−ligand complex, but more importantly, the coordinates are only a model of the true experimental data. Furthermore, RMSD can be biased by the frame of reference; for example, binding-site residues and a ligand can shift just 1 Å in flexible docking and result in an artificially large RMSD despite maintaining all relevant contacts. Lastly, a random placement of a small ligand can have low RMSDs while symmetric molecules that are not handled properly can produce artificially high RMSD values.[45] Native contacts appear to be a robust measurement that can capture the complex interactions and, used in combination with the RMSD, provide a thorough evaluation of the predicted poses. Although noted in the literature that a drawback of native contacts is its inability to be automated,[45] we have created an automated tool in python to calculate both the percentage of native contacts correct between the protein and predicted ligand pose and a raw count of contacts (all contacts made between the protein and predicted ligand pose).

To assess the performance of the scoring function in a virtual screening-type application, enrichment and relative-ranking studies are commonly employed. The area under the curve (AUC) of receiver operator characteristic (ROC) curves[48,49] based on the rank score are typically reported for enrichment; this metric is able to assess how well a ranking function identifies known inhibitors as high ranking and discriminates them from inactive ligands. Standard correlation measures are used to evaluate the ability of scoring functions to rank-order active compounds, and sound statistical methods are necessary to identify true trends in the data.[20,35,39,41] Pearson's correlation (r) is typically employed to provide a linear relationship, whereas Spearman's rho (ρ) and Kendall's tau (τ) measures the strength of the nonparametric relationship between the ranks of the data. Hence, r is a better measure for assessing absolute predictions, while ρ and τ are more appropriate metrics for relative ranking.

Here, we present an evaluation of the results from the CSAR center's first blinded benchmark exercise. In order to avoid a winner-vs-loser mentality, participants were asked to submit two sets of results and focus on testing a hypothesis of their choice, rather than comparing their results to others. The exercise concluded with a symposium at the Fall 2012 National Meeting of the American Chemical Society (ACS) in Philadelphia, with eight speakers and an open discussion session.

Contributors. Most of the pose predictions and ranking values evaluated were calculated by authors featured in the same CSAR special issue of the Journal of Chemical Information and Modeling, some by the CSAR team, and a few by participants who spoke at the ACS symposium but were unable to submit papers to the special issue due to various time constraints. A variety of methods/codes were utilized in the exercise and include Gold, MOE, AutoDock, AutoDock Vina, MedusaDock, RosettaLigand, Schrödinger Induced Fit Dock, Q-Mol, OEDocking, CDOCKER, ICM-VLS, BlurDock, Glide,

MDock, Sol, and WILMA. Most are custom versions or in-house software and were expert-guided docking protocols. To provide anonymity to each group, they are denoted below as A−U, and each method submitted as 1−6. If a group submitted results using more than one docking program, they were separated into multiple groups (i.e., A, B, and C). We have done this to avoid a win−lose mentality as this benchmark exercise is not meant to be a contest but rather a means to elucidate important and common deficiencies across the methods in predicting and scoring binding poses. Additionally, a breakdown of the various sampling and ranking scoring functions and docking program used by each group is provided in Table 1. The docking programs are denoted a−r to once again provide anonymity, but this allows the reader to determine which groups used the same programs. Our hope is that this study will help direct the computational community to where the most significant effort is needed for future methodology development.

Table 1. Breakdown of Employed Sampling and Ranking Scoring Function and Docking Software for Each Group

group	sampling scoring function	ranking scoring function	docking software[a]
A	force field-based	force field-based	a
B	force field-based	force field-based	b
C	knowledge- and force field-based combined	knowledg- and force field-based combined	c
D	knowledge-based	knowledge-based	d
E	knowledge-based	knowledge-based	e
F	empirical-based	empirical-based	**f**
G	empirical-based	empirical-based	g
H	force field-based	force field-based	h
I	empirical-based	empirical-based	**f**
J	empirical-based	empirical-based	i
K	empirical-based	force field-based	j
L	empirical-based	empirical-based	k
M	force field-based	knowledge-based	**l**
N	knowledge- and empirical-based combined	knowledge- and empirical-based combined	m
O	force field-based	empirical-based	**n**
P	crude shape complementarity	knowledge-based	o
Q	knowledge-based	knowledge- and empirical-based combined	p
R	force field- and empirical-based combined	force field- and empirical-based combined	**l**
S	empirical-based	empirical-based	**n**
T	force field-based and shape/functionality-based complementarity	force field-based and shape/functionality-based complementarity	q
U	force field-based	force field-based	r

[a]Docking codes used by more than one group are shown in bold.

METHODS

Data Set and Participation. The goal of the 2011−2012 blinded exercise was to compare different improvements for docking and relative ranking of congeneric series of compounds, testing three areas: (1) pose prediction, (2) enrichment/discriminating active from inactives, and (3) relative ranking. The exercise was built around blinded, high-quality, experimental data from four protein targets: LpxC (University of Michigan data), Urokinase (Abbott data), Chk1 (Abbott data), and Erk2 (Vertex data). Participants were provided with a set of SMILES strings of the active and inactive ligands, the pH of the assay used to determine the binding data, and a PDB code to use for docking. Cross-docking studies, in addition to an analysis of the active site, were conducted in-house to determine the most appropriate PDB structure for use with each target (3P3E[50] for LpxC, 1OWE[51] for Urokinase, 2E9N[52] for Chk1, and 3I5Z[53] for Erk2, as shown in Table 2). Participants were asked to submit two sets of results and test a hypothesis of their choice. Twenty groups worldwide participated in the exercise where 17 sent pose predictions (the majority sent in the top three poses using multiple methods, resulting in 3250 total poses) and 18 sent in rankings (the majority sent in one set of rankings using multiple methods, resulting in 174 total rankings).

Obviously, participants were not told which ligands were active or inactive (inactive ligands will not have corresponding structures). As such, participating groups submitted poses of all ligands, but only those with corresponding high-quality CSAR structures were used in our pose prediction analysis. Active molecules that do not have corresponding crystal structures were also included in the enrichment/relative ranking portion of the exercise. A summary of the number of CSAR protein−ligand complexes and active/inactive molecules employed for each target is provided in Table 2 along with the number of predictions received for pose prediction and enrichment/relative ranking broken down by protein. A complete description of how the four data sets were curated and prepared for this exercise can be found in ref 40.

An online questionnaire was sent to all participating groups to gather additional data on the details of their methodology. We had a 100% response rate and gathered the following information: details on ligand setup, protein setup, and the molecular docking protocol, in addition to thoughts on the analysis conducted by the CSAR team. Having the methodology details used by each group allowed us to identify how particular aspects of docking programs across multiple groups' results affected the pose/ranking predictions.

Pose Prediction. RMSD and native contacts were used to evaluate the predicted poses. All poses, the best pose, and the top-scoring pose were evaluated for the predictions. We superimposed the submitted protein−ligand complexes using the wRMSD method[54] to the unpublished CSAR crystal structure to provide the same frame of reference. Groups were

Table 2. Summary of Data Sets Employed in Benchmark Exercise and Predictions Received

protein target	data source	PDB structure employed	# of structures	pose predictions received	# of active ligands	# of inactive ligands	ranking predictions received
LpxC	University of Michigan	3P3E	4	458	3	8	39
Urokinase	Abbott	1OWE	4	390	16	4	45
Chk1	Abbott	2E9N	14	1279	38	9	47
Erk2	Vertex	3I5Z	12	1123	39	−	43

dx.doi.org/10.1021/ci400025f | J. Chem. Inf. Model. 2013, 53, 1853−1870

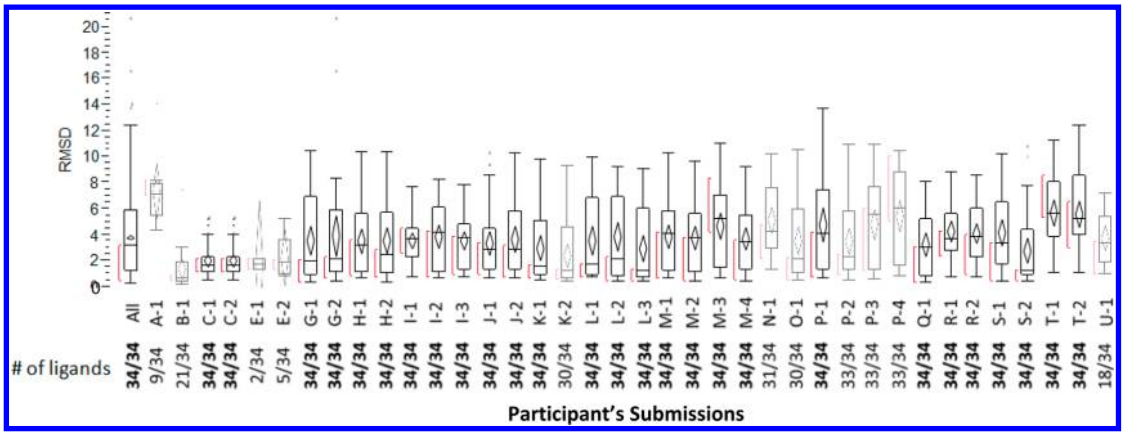


Figure 1. RMSD box plot of the best pose for each protein−ligand complex broken down by group−method. The rectangular box indicates the interquartile range (25−75%), and the bars the 1.5× interquartile range. The median is shown by the line in the box, and the diamond denotes the mean and 95% confidence interval around the mean. The red bracket signifies the shortest interval that contains 50% of the data, and outliers are indicated by squares above the bars. Group−method, which submitted scores for all ligands of LpxC, Urokinase, Chk1, and Erk2, are bolded.

asked to dock into a protein conformation found in the PDB, but the poses were compared back to the crystal structure solved by the CSAR group. If only the ligand pose was sent, we first placed it in the context of the suggested PDB structure for the exercise (i.e., 3P3E for LpxC) and then performed the superposition. Once the complex was superimposed, we calculated the RMSD between the predicted ligand pose and the crystallographic coordinates. Only heavy atoms were used, and symmetry was accounted for in the calculation. The script used was graciously donated by the Chemical Computing Group and run in MOE 2010.11.[55] A "correct" docking pose was defined as having a RMSD of less than 2 Å.[56,57] The RMSD script was utilized to pull out the atom correspondence for each ligand pose.

Additionally, a python script was written in-house to analyze for hetero−hetero, carbon−carbon, and packing contacts between the protein and predicted ligand pose. For the packing contact, atom types are ignored and all contacts are counted. Waters were not included in the analysis, but in the LpxC test case, the catalytic zinc was included as part of the protein. If a zinc atom was not present in any submission, the location of the 3P3E catalytic zinc was used. The cutoff values used for the interactions are as follows:

- O−O, O−N, O−F, N−N, and N−F ≤ 3.5 Å (approximate hydrogen bonding and electrostatic interactions)
- S-x, Br-x, Cl-x ≤ 3.8 Å, where x is O or N (approximate longer and weaker electrostatic interactions)
- Zn-x ≤ 2.8 Å, where x is O or N (ion coordination)
- C−C ≤ 4.0 Å (capture tight and loose van der Waals interactions)
- Packing ≤ 4.0 Å (capture all interactions)

Two types of analysis were conducted with the contacts: (1) the number of native contacts correctly predicted (i.e., same contacts as those in the co-crystal structure) and (2) a raw count of contacts (i.e., all contacts made between the protein and ligand). The number of native contacts correctly predicted can be thought of as an assessment metric: Is the predicted pose making the same contacts to the protein as in the native co-crystal structure? Is this pose right or wrong? The raw count is not an assessment of the pose per se but provides information on the actual contacts being made between the predicted pose and protein. Here, we are probing the reason why the pose is different to elucidate the cause of the problem; for example, which types of contacts are being sacrificed or overpredicted by the various scoring functions.

For the number of native contacts correctly predicted, only those contacts made between the predicted ligand pose and protein structure that are also present in the CSAR co-crystal structure (i.e., the native contacts) are summed and then broken down by hetero−hetero contacts (%Het−Het) and carbon−carbon contacts (%C−C) or as a total count (%Total). The contact script accounts for the symmetry of the ligand and histidine tautomers and residue equivalence. For example, a contact to either aspartic acid acidic oxygen would count equally. On the same note, we count all carbons in a residue equally, and hence, a contact to any carbon would count.

For the raw count of pose contacts, all of the hetero−hetero contacts made between the predicted ligand pose and protein structure are summed up (Het−Het), then carbon−carbon contacts (C−C), and finally packing contacts. We then determined whether the pose overpredicted, underpredicted, or had the same number of contacts in the co-crystal structure within 10%. To obtain a 95% confidence interval, 10,000 random bootstrap samples of the raw count contact data were taken with replacement. From each sample, the 95% confidence interval was determined by the 2.5 percentile and 97.5 percentile of the distribution of overpredicted, underpredicted, and same contacts from the bootstrap samples.

Furthermore, we have attempted to use RSRs and RSCCs for the assessment of the predicted ligand poses. RSR and RSCC provide a fit of the predicted ligand pose to the electron density and, as such, are an evaluation based on the raw experimental data. A crystal structure is a model. Hence, comparing back to the experiment data will remove a layer of bias from the evaluation—the error of the model is not propagated into the analysis. Unfortunately, neither the RSR or RSCC values were reproducible between the different versions of CCP4[58] used (4.1.2 and 4.2.0). We were also unable to reproduce the values reported by the Uppsala Electron Density Server (EDS)[59] for the original crystal structure. Because of this inconsistency, we feel it is difficult to trust these values, and as such, they are not used in our pose prediction analysis.

Ranking Evaluation and Statistical Analysis. We have evaluated the ability of scoring functions to properly identify the inactives in the series and their ability to rank-order the

dx.doi.org/10.1021/ci400025f | J. Chem. Inf. Model. 2013, 53, 1853−1870

Table 3. % Predictions <2 Å for Each Group−Method by Protein Target (Best Pose)[a]

	LpxC % predictions <2 Å	Urokinase % predictions <2 Å	Chk1 % predictions <2 Å	Erk2 % predictions <2 Å
group A-1	0.0	0.0	0.0	0.0
group B-1	**100.0**	**100.0**	**71.4**	**50.0**
group D-1	**100.0**	**100.0**	**71.4**	25.0
group D-2	**100.0**	**100.0**	**71.4**	25.0
group E-1	N/A	N/A	0.0	**100.0**
group E-2	N/A	**100.0**	33.3	**100.0**
group G-1	**100.0**	**50.0**	42.9	41.7
group G-2	**75.0**	**50.0**	28.6	41.7
group H-1	**75.0**	**50.0**	28.6	**50.0**
group H-2	**50.0**	**50.0**	35.7	**50.0**
group I-1	**100.0**	0.0	28.6	0.0
group I-2	**100.0**	**100.0**	35.7	8.3
group I-3	**100.0**	**75.0**	28.6	0.0
group J-1	**100.0**	**75.0**	42.9	25.0
group J-2	**100.0**	**75.0**	42.9	25.0
group K-1	**100.0**	**100.0**	**57.1**	25.0
group K-2	N/A	**100.0**	**78.6**	41.7
group L-1	**100.0**	**100.0**	**50.0**	33.3
group L-2	**100.0**	**100.0**	28.6	41.7
group L-3	**100.0**	**100.0**	**50.0**	**50.0**
group M-1	**100.0**	**75.0**	28.6	8.3
group M-2	**100.0**	**50.0**	21.4	25.0
group M-3	**100.0**	**100.0**	14.3	0.0
group M-4	**100.0**	**50.0**	28.6	16.7
group N-1	0.0	**100.0**	23.1	0.0
group O-1	**100.0**	N/A	**50.0**	33.3
group P-1	**100.0**	**100.0**	28.6	0.0
group P-2	**100.0**	**75.0**	30.8	16.7
group P-3	**75.0**	**100.0**	21.4	9.1
group P-4	**75.0**	**75.0**	21.4	9.1
group Q-1	**100.0**	**100.0**	14.3	25.0
group R-1	0.0	25.0	21.4	0.0
group R-2	25.0	**50.0**	21.4	8.3
group S-1	**100.0**	**75.0**	7.1	16.7
group S-2	**100.0**	**100.0**	**57.1**	**50.0**
group T-1	25.0	25.0	21.4	0.0
group T-2	**75.0**	25.0	0.0	0.0
group U-1	**50.0**	**66.7**	**100.0**	0.0

[a]Group−methods able to predict greater than or equal to 50% are bolded.

active compounds. ROC plots were generated to determine the AUC;[49] the greater the AUC, the better the ability of the method to identify active over inactive compounds. To obtain 95% confidence intervals around the AUC, bootstrap sampling was performed by randomly selecting samples with replacement 10,000 times. The size of each sample was the same as the size of the set used to generate the ROC plot. The AUC was calculated for each sample and the 2.5 percentile and 97.5 percentile of the resulting distribution of 10,000 AUC values were computed to give the 95% confidence interval.[25,60] Software was kindly provided by Ajay Jain to compute the confidence intervals.

Pearson's (r) parametric correlation coefficient and Spearman (ρ) and Kendall (τ) nonparametric correlation coefficients were calculated to determine the correlation between the predicted and known affinities. The software JMP[61] was used to calculate all statistics, unless otherwise noted. Fisher transformations combined with standard deviations were used to determine 95% confidence intervals around the Pearson correlation and Spearman correlation.[62] For the Kendall statistic, the Fisher transformation cannot be used; therefore, the approximation of $1.96 \times (1 - \tau^2)((2(2n + 5))/(9n(n - 1)))^{1/2}$ was used to determine the 95% confidence interval.[63] In our previous evaluation paper, we discussed the use of heavy atoms and SlogP as a "yardstick" to determine a baseline or null correlation.[20] Here, we have calculated the molecular weight of the ligand and SlogP using MOE 2010.11[55] as null control cases and identified groups that are statistically significant from these values. In order to compare the R^2 values across individual groups to the yardsticks, the variance in the residuals from the linear regression were compared using Levene's F-test using R.[64] A probability of a F-statistic less than 0.05 indicates that the error between the two fits is statistically different.

RESULTS AND DISCUSSION

How well did methods perform overall on predicting poses? Figure 1 shows a RMSD box plot of the best pose for each protein−ligand complex broken up by group−method, each method from each group. RMSD box plots provide the distribution of the RMSD values and the associated statistics

dx.doi.org/10.1021/ci400025f | J. Chem. Inf. Model. 2013, 53, 1853−1870

(mean, median, 95% confidence interval around the mean, interquartile range, outliers, etc.). For all protein targets combined, the median RMSD across all group−methods was 3.0 Å. Additionally, 37% of the group−methods have a median RMSD of less than 2.0 Å. The results were also broken down by protein target; the RMSD box plots are provided in the Supporting Information. The median RMSD was 1.14 Å for LpxC, 1.25 Å for Urokinase, 3.50 Å for Chk1, and 5.03 Å for Erk2. Table 3 provides a breakdown by protein on the percent of predictions less than 2.0 Å for each group−method. Only three groups (<10%) were able to predict well across all protein systems (of these groups, two used empirical-based scoring functions for sampling and ranking and one used force field-based). This is not surprising as it is known that most scoring functions are not robust enough to perform well across binding sites of various sizes, accessibility, and chemical properties.

The best pose is presented because it attempts to remove the bias of the scoring function on the results and asks if the method was able to find the correct pose in the top three predicted poses submitted by each group. We found that the best pose was also the top scoring pose 50.6% of the time for all protein targets combined, 58.6% for just LpxC, 46.7% for just Urokinase, 50.3% for just Chk1, and 48.9% for just Erk2. Essentially, scoring functions are predicting the best pose as the top pose better than random but still not at the rate that is necessary for drug discovery purposes. However, it is important to note that in most cases for this analysis the trends are essentially the same whether all poses, best poses, or top poses are utilized.

Which test sets were most challenging for predicting poses? Figure 2 shows a RMSD box plot of the best pose for each protein−ligand complex broken down by protein target. LpxC and Urokinase had the smallest median RMSD (1.14 Å and 1.25 Å, respectfully) with Chk1 a bit higher (3.50 Å) and then Erk2 (5.03 Å). As demonstrated in Table 3, the majority of groups were able to predict 75−100% of LpxC and Urokinase poses with a RMSD of less than 2.0 Å. Table 4 provides the distribution of RMSD for all, best, and top poses. Various benchmark studies have been conducted using the same test cases as discussed above.[6,25,26] However, a direct comparison cannot be made between our analysis and the published studies as different data sets of ligands were used. This also illustrates the need for standardized data sets such as those developed by CSAR; if groups were consistent with the benchmark data sets employed when evaluating their methodology developments, then the field would be able to assess whether positive improvements have actually been made.

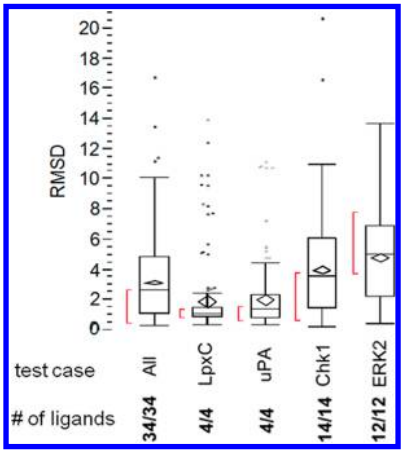


Figure 2. RMSD box plot of the best pose for each protein−ligand complex broken down by protein target. The rectangular box indicates the interquartile range (25−75%) and the bars the 1.5× interquartile range. The median is shown by the line in the box, and the diamond denotes the mean and 95% confidence interval around the mean. The red bracket signifies the shortest interval that contains 50% of the data, and outliers are indicated by squares above the bars.

LpxC and Urokinase only have one chemical series, while both Chk1 and Erk2 have three series within the ligand set provided to the participants. It appears that the most prominent reason that groups did not perform as well on Chk1 and Erk2 is because of the multiple chemical series. If the chemical series are broken out, performance across the different protein targets was very comparable when comparing the series that contained a chemically similar ligand to the co-crystal structure utilized for docking. Methods performed much better on series 1 than series 2 or 3 for both Chk1 and Erk2. The crystal structure suggested by the CSAR team for both Chk1 and Erk2 are co-crystallized with a ligand from their respective series 1. Accounting for the conformational changes that can occur within the binding pocket of the protein is a very difficult task.[2,8−13] In co-crystal structures, the prearrangement of the ligand binding site can lead to the cross-docking problem where the protein structure has adapted to bind a particular ligand or class of ligands but is unable to accommodate structurally diverse inhibitors as we found here. Incorporating protein flexibility is recognized as a means to overcome the cross-docking problem; however, not enough groups used protein flexibility to allow us to perform a statistically significant analysis on whether or not it affected the docking results.

How did RMSD correlate with native contacts? We first asked if the native contact metric agrees with RMSD and if it provides any additional useful information. Figure 3 shows native contact box plots of the best pose for each protein−ligand complex broken up by group−method for %Total, % Het−Het, and %C−C contacts correct. When comparing all series together, native contacts show the same trend as RMSD. Groups performed the best on LpxC and Urokinase; the median %Total was equal to 51% and 52%, respectively. Chk1 and Erk2 appeared to be more difficult; median %Total was equal to 25% and 13%, respectively. However, unlike a RMSD, native contacts can provide additional information on the specific type of contacts that are being made, as demonstrated in Figure 3B and C. When comparing the %Het−Het contacts correct versus %C−C contacts correct, it appears that groups were more successful at predicting the Het−Het contacts in Urokinase (%Het−Het = 77%). However, the C−C contacts groups were more successful at LpxC (%C−C = 54%). Another interesting trend that emerged is that methods had a more difficult time predicting the C−C contacts than the Het−Het contacts between the protein and ligand. In fact, 7.3% of the group−methods were able to predict 100% of the Het−Het contacts, while no group−methods were able to predict 100% of the C−C contacts. Additionally, 13.2% of the group−methods could predict greater than 80% of the Het−Het contacts, but only 1.6% of the group−methods could do the same for C−C.

Figure 4 shows the correlation between the calculated native contact and RMSD values. The data suggests that the values are exponentially correlated (r^2 = 0.75); as the ligand moves further away from the protein, contacts are lost at an exponential rate. Kroemer et al. also found that overall RMSD correlates with

Table 4. Distribution of Pose RMSD Values by Protein[a]

	<1 Å (%)	1−2 Å (%)	2−3 Å (%)	3−4 Å (%)	4−5 Å (%)	>5 Å (%)	median RMSD
LpxC							
all poses (n = 458)	22.05	**41.27**	15.50	4.15	1.31	15.72	
best poses (n = 174)	**34.48**	**47.70**	9.20	0.57	0.57	7.47	1.14
top poses (n = 174)	24.14	**49.43**	14.37	2.30	1.72	8.05	
Urokinase							
all poses (n = 390)	22.31	**29.74**	22.56	5.38	4.87	15.13	
best poses (n = 137)	**35.04**	**38.69**	13.87	3.65	2.92	5.84	1.25
top poses (n = 137)	24.09	**33.58**	22.63	5.84	3.65	10.22	
Chk1: all 3 ligand series combined							
all poses (n = 1279)	9.38	12.90	5.63	12.90	9.54	**49.65**	
best poses (n = 477)	16.35	18.24	7.76	17.19	9.01	**31.45**	3.50
top poses (n = 477)	12.58	15.09	5.24	13.00	9.43	**44.65**	
Chk1: series 1							
all poses (n = 364)	27.47	22.80	9.89	5.77	3.30	**30.77**	
best poses (n = 141)	**44.68**	26.95	8.51	4.96	0.71	14.18	1.08
top poses (n = 141)	**36.88**	24.82	8.51	4.26	2.84	22.70	
Chk1: series 2							
all poses (n = 442)	1.13	5.43	4.75	18.78	14.93	**54.98**	
best poses (n = 166)	3.01	10.84	8.43	24.10	15.06	**38.55**	4.20
top poses (n = 166)	1.20	4.22	3.61	18.07	15.06	**57.83**	
Chk1: series 3							
all poses (n = 473)	3.17	12.26	3.17	12.90	9.30	**59.20**	
best poses (n = 170)	5.88	18.24	6.47	20.59	10.00	**38.82**	3.94
top poses (n = 170)	3.53	17.65	4.12	15.29	9.41	**50.00**	
Erk2: all 3 ligand series combined							
all poses (n = 1123)	4.54%	9.35	5.43	5.43	8.90	**66.34**	
best poses (n = 411)	8.76	12.90	7.79	8.27	11.44	**50.85**	5.03
top poses (n = 411)	6.33	9.98	5.84	7.06	8.27	**62.53**	
Erk2: series 1							
all poses (n = 186)	8.60	**30.65**	6.45	9.68	11.83	**32.80**	
best poses (n = 71)	14.08	**40.85**	7.04	12.68	9.86	15.49	1.67
top poses (n = 71)	8.45	**43.66**	2.82	11.27	14.08	19.72	
Erk2: series 2							
all poses (n = 745)	1.74	3.89	4.56	5.50	9.93	**74.36**	
best poses (n = 276)	4.71	5.80	6.88	9.06	13.77	**59.78**	5.49
top poses (n = 276)	3.99	1.45	6.16	7.25	7.97	**73.19**	
Erk2: series 3							
all poses (n = 192)	11.46	9.90	7.81	1.04	2.08	**67.71**	
best poses (n = 64)	20.31	12.50	12.50	0.00	3.13	**51.56**	5.06
top poses (n = 64)	14.06	9.38	7.81	1.56	3.13	**64.06**	
All proteins							
all poses (n = 3250)	11.05	17.69	8.98	8.18	7.60	**46.49**	
best poses (n = 1199)	18.52	23.02	8.67	10.18	7.92	**31.69**	3.00
top poses (n = 1199)	13.43	20.43	8.76	8.59	7.26	**41.53**	

[a]RMSD values greater than 30% are bolded.

their interactions-based accuracy classification (with the exception of a few test cases).[45] As demonstrated in Figure 4A, for this data set, it is not possible to have a RMSD better than 3.45 Å when no contacts are being made. Furthermore, at 50% total contacts correct, the RMSD is 1.55 Å (on average) with a range up to ∼3 Å.

To date, the field uses 2 Å as the cutoff for a successful docking pose. This value was not determined quantitatively but rather through qualitative inspection over many years of evaluating docking programs and the desire to use a round number. Utilizing both native contacts and RMSD provides the researcher with a more complete picture of their docking performance and allows for a more quantitative analysis of the results. Here, we can use the %Total contacts correct at various RMSD cutoff values to examine if 2 Å is an appropriate metric, and if not, what is. At a RMSD cutoff of 2 Å, the %Total contacts correct ranges from 14% to 86% (for 499 data points). The same analysis was conducted at a RMSD cutoff of 1.5, 2.5, 3, and 3.5 Å, and the ranges along with the percentage of % Total contacts that fall within various cut-offs are provided in the Supporting Information. An examination of all data suggests that lowering the value does not gain significant contacts; however, 2.5 Å is just as reasonable of a cutoff as 2 Å for defining a correct pose.

The data for %C−C contacts correct (Figure 4C) essentially follows the same trend shown in %Total (Figure 4A). On the % Het−Het contacts correct graph (Figure 4B), there are interesting data points where 0% of the correct contacts are

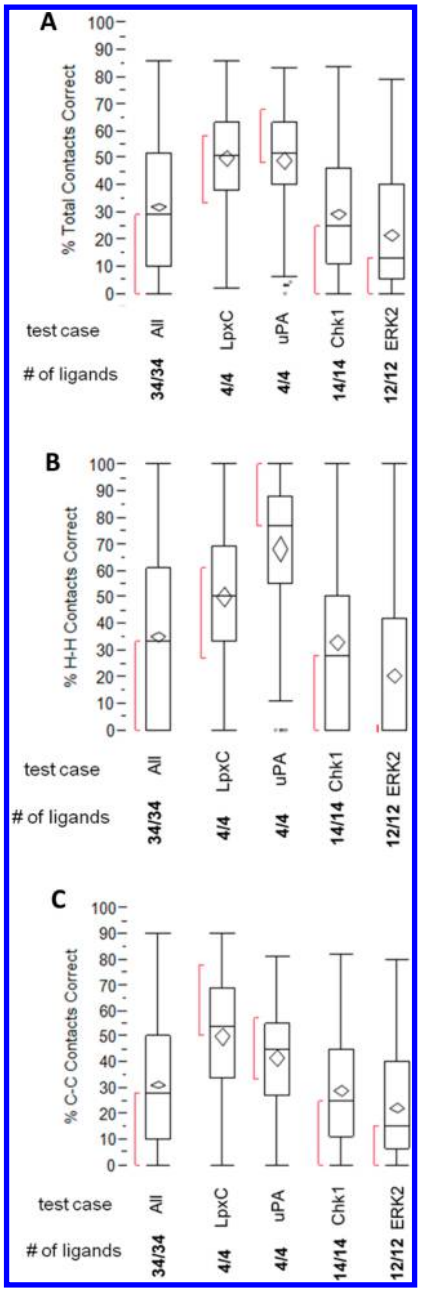


Figure 3. Native contacts box plot of the best pose for each protein−ligand complex broken down by protein target. The rectangular box indicates the interquartile range (25−75%) and the bars the 1.5× interquartile range. The median is shown by the line in the box, and the diamond denotes the mean and 95% confidence interval around the mean. The red bracket signifies the shortest interval that contains 50% of the data, and outliers are indicated by squares above the bars. (A) %Total contacts correct, (B) %Het−Het contacts correct, and (C) %C−C contacts correct.

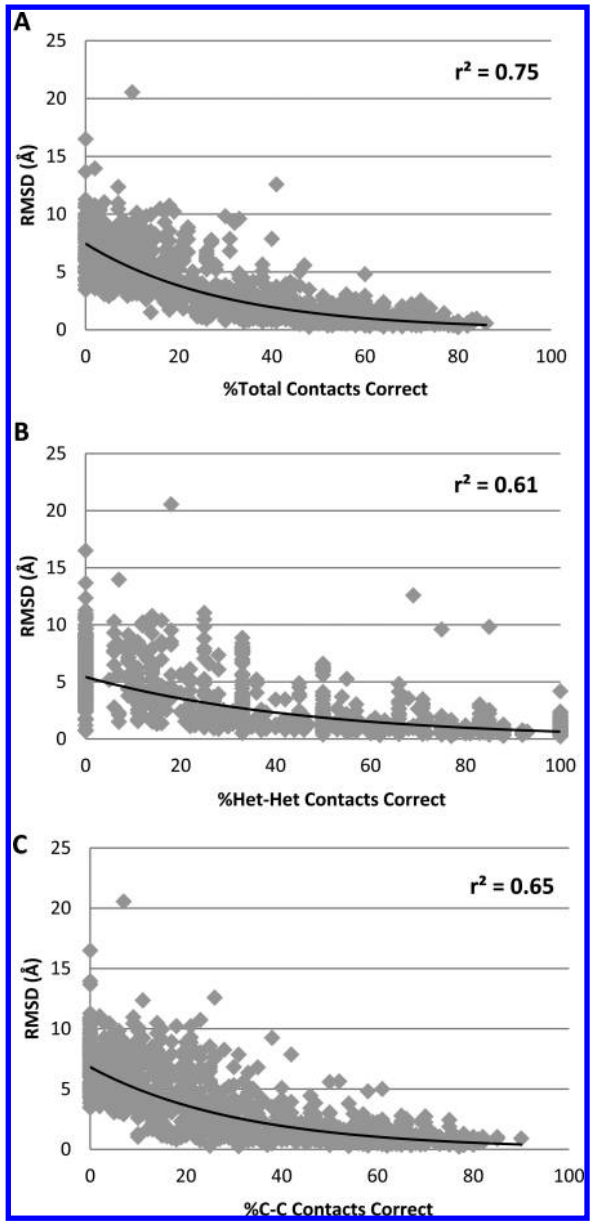


Figure 4. (A) %Total contacts correct, (B) %Het−Het contacts correct, and (C) %C−C contacts correct plotted again RMSD. The exponential fit is shown on each graph.

being made at a range of RMSD values (even less than 2 Å). Careful examination of the predicted poses elucidated that these points are all from Chk1. As shown in Figure 5, the ligand is just slightly shifted to the right. Although, the RMSD is equal to 0.702, both of the hinge region hydrogen bonds have been lost. One must be careful in the interpretation of native contacts data because the number of hydrogen bonds is typically much smaller than the number of carbon−carbon interactions, and the number of hydrogen bonds varies significantly from target to target. This data will also be influenced by the size and composition of the ligand.

Did the scoring functions overpredict or underpredict raw contacts? Additional information on whether scoring functions overpredict or underpredict contacts can be gathered by analyzing all raw contacts made between the protein and ligand (i.e., not just the percent native contacts correct as previously presented). This is an important question because it highlights the cause of the differences in the poses rather than just assessing whether or not the correct ligand pose was found. Consequently, it emphasizes weaknesses that could be addressed in scoring function development. Table 5 presents the percentage of raw Het−Het, C−C, and packing contacts that were overpredicted, underpredicted, or the same number of contacts (within 10%) for all protein targets combined and broken down by each protein. An important note is that "same-predicted" does not mean the same contacts are being made

dx.doi.org/10.1021/ci400025f | J. Chem. Inf. Model. 2013, 53, 1853−1870

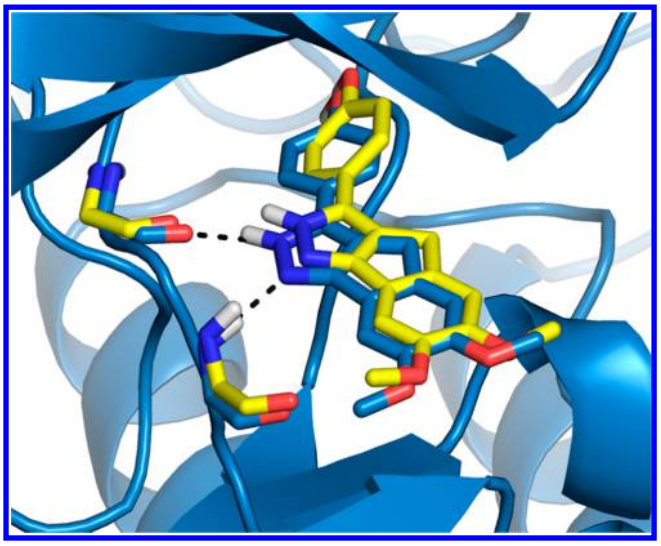


Figure 5. Predicted docking pose (submission; yellow) overlaid with the experimental co-crystal structure of Chk1−ligand 1 (blue). Dotted lines illustrate two important hydrogen bonds formed between the ligand and the hinge region of the protein backbone. The RMSD between the coordinates of the predicted pose and coordinates of the experimental structure is equal to 0.702, %Het−Het contacts correct is equal to 0%, and %C−C contacts correct is equal to 37%.

Table 5. Percentage of Raw Het−Het, C−C, and Packing Contacts Where the Number Was Overpredicted, Underpredicted, or Had the Same Contacts within 10% for Best Pose[a]

	Hetero−Hetero	Carbon−Carbon	packing
All (n = 1199)			
same predicted	17.70 ± 2.17	21.10 ± 2.25	32.53 ± 2.63
underpredicted	**50.37 ± 2.79**	**40.38 ± 2.84**	**40.71 ± 2.80**
overpredicted	31.93 ± 2.63	**38.53 ± 2.75**	26.60 ± 2.46
LpxC (n = 174)			
same predicted	21.26 ± 6.32	29.90 ± 6.90	**49.41 ± 7.47**
underpredicted	**73.00 ± 6.61**	19.00 ± 5.75	23.56 ± 6.32
overpredicted	5.75 ± 3.74	**51.20 ± 7.47**	27.03 ± 6.61
Urokinase (n = 137)			
same predicted	27.77 ± 7.30	19.75 ± 6.57	29.90 ± 7.66
underpredicted	**37.21 ± 8.03**	**59.09 ± 8.03**	**50.37 ± 8.39**
overpredicted	**35.02 ± 8.03**	21.16 ± 6.93	19.75 ± 6.93
Chk1 (n = 477)			
same predicted	15.33 ± 3.25	20.13 ± 3.46	31.23 ± 4.19
underpredicted	**49.48 ± 4.57**	**43.61 ± 4.51**	**46.95 ± 4.51**
overpredicted	35.19 ± 4.19	**36.26 ± 4.40**	21.82 ± 3.67
Erk2 (n = 411)			
same predicted	15.53 ± 3.53	19.00 ± 3.77	27.75 ± 4.28
underpredicted	**46.19 ± 4.87**	**39.41 ± 4.74**	**37.46 ± 4.77**
overpredicted	**38.27 ± 4.74**	**41.60 ± 4.87**	34.31 ± 4.52

[a]Values greater than 35% are bolded.

but rather that the same *number* of contacts has been predicted. For all proteins combined, the trend emerges that scoring functions have a bias to underpredict Het−Het and packing contacts but both underpredict and overpredict C−C contacts at the same rate. About half of all methods underpredict Het−Het contacts, while only 32% overpredict them. For C−C contacts, the same exact number of methods overpredict and underpredict (∼40%), and there is no general bias seen. We were very surprised to find that Het−Het and packing contacts were biased toward underpredicting contacts as the overall consensus in the field is that scoring functions tend to focus on optimizing both types of contacts and overpredicting the interactions.

In Figure 6A and B, the RMSD < 1 Å bin and RMSD = 1−2 Å bin for Het−Het contacts are provided, respectively; the population of each value is given by the size of the point. As one would expect, when the RMSD is quite small, the majority of the points are close to the identity line. It is also obvious from these graphs that when the RMSD is small, the trend that scoring functions are underpredicting contacts holds true. As the RMSD becomes larger, the data becomes more spread and moves away from the identity line (data for bins 2−4, 4−10, and >10 Å is provided in the Supporting Information). Furthermore, once the RMSD is greater than 10 Å, almost all of the contacts are off the identity line and being underpredicted. Figure 7A and B show the RMSD < 1 Å bin and RMSD = 1−2 Å bin for C−C contacts and Figure 8A and B for Packing contacts (again data for bins 2−4, 4−10, and >10 Å is provided in the Supporting Information). For C−C contacts, the points are spread almost evenly between underprediction and overprediction. The packing contacts agree with what was shown for Het−Het contacts, and at small RMSD values, the trend that the scoring function is underpredicting contacts remains. Again, this is a very interesting finding as most scoring functions use an additive term for van der Waals packing, and hence, more contacts should result in a better score.

Table 5 also shows the data broken down by protein target. For LpxC, the Het−Het contacts were significantly underpredicted, while the C−C contacts were overpredicted. The packing contacts were overpredicted and underpredicted at essentially the same rate. LpxC contains a catalytic zinc atom, which was included as part of the active site. Methods had a difficult time predicting these hydrogen bonds. However, the scoring function was still able to rank these predictions correct because it was overcompensating by overpredicting C−C contacts.

Did the docking metrics correlate with ligand descriptors? When utilizing new metrics for pose prediction, it is always prudent to determine if they correlate with size and chemical properties of the ligand. To assess this, MOE 2010 was utilized to calculate the Pearson (r), Spearman (ρ), and Kendall (τ) correlations between the ligand properties and RMSD, %Het−Het, and %C−C. The results are provided in the Supporting Information for molecular weight, # of atoms, # of heavy atoms, # of hetero atoms, # of hydrophobic atoms, # of acceptors, # of donors, # of acceptors and donors, # of carbons, and # of nitrogens. We found that there was no correlation between the ligand size and the metrics. Furthermore, we found that the metrics were not correlated with chemical properties of the ligand such as the number of hydrogen bond acceptors and donors, among others.

How did the methodology "features" correlate with RMSD? Each of the 20 groups employs their own protocols for protein and ligand setup and docking methodology. In order to understand how such methodology "features" across multiple groups' results affected the pose/ranking predictions, we asked each participant to fill out an online questionnaire to gather additional data on the details of their methodology. The pose

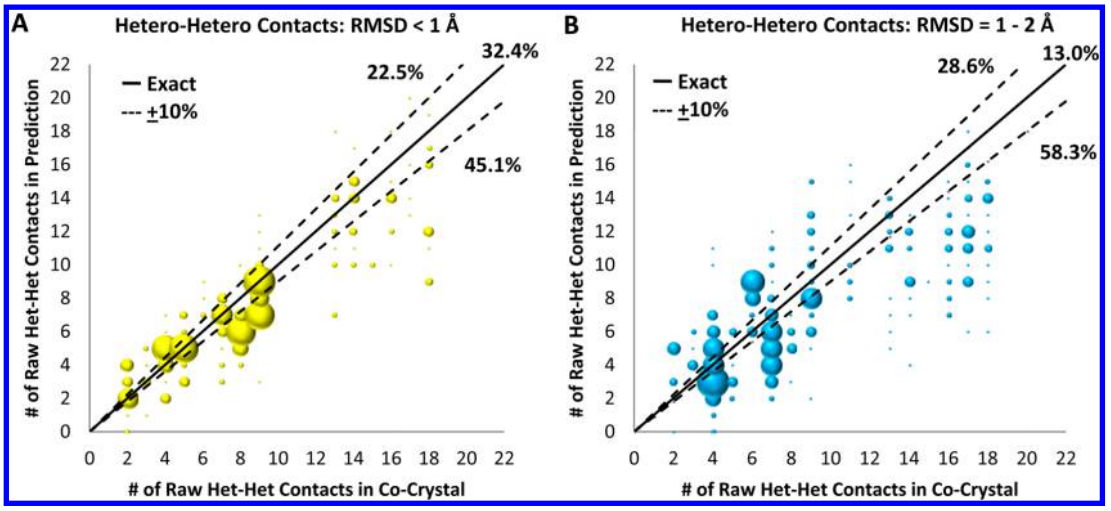


Figure 6. Number of raw Het−Het contacts in co-crystal versus number of raw Het−Het contacts in prediction. The solid line illustrates a perfect match, while the dotted lines show a ±10% range. (A) RMSD < 1 Å bin. (B) RMSD = 1−2 Å bin.

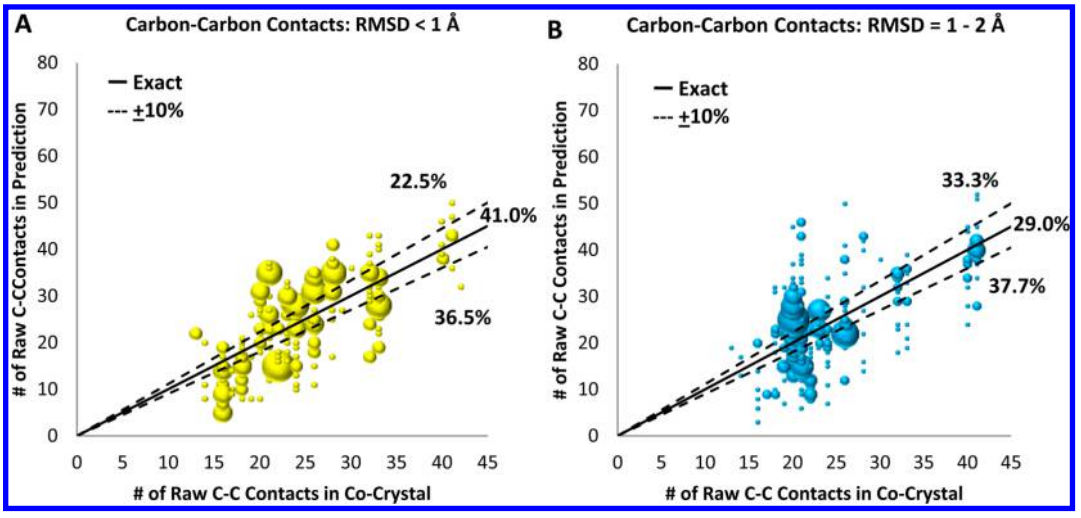


Figure 7. Number of raw C−C contacts in co-crystal versus number of raw C−C contacts in prediction. The solid line illustrates a perfect match, while the dotted lines show a ±10% range. (A) RMSD < 1 Å bin. (B) RMSD = 1−2 Å bin.

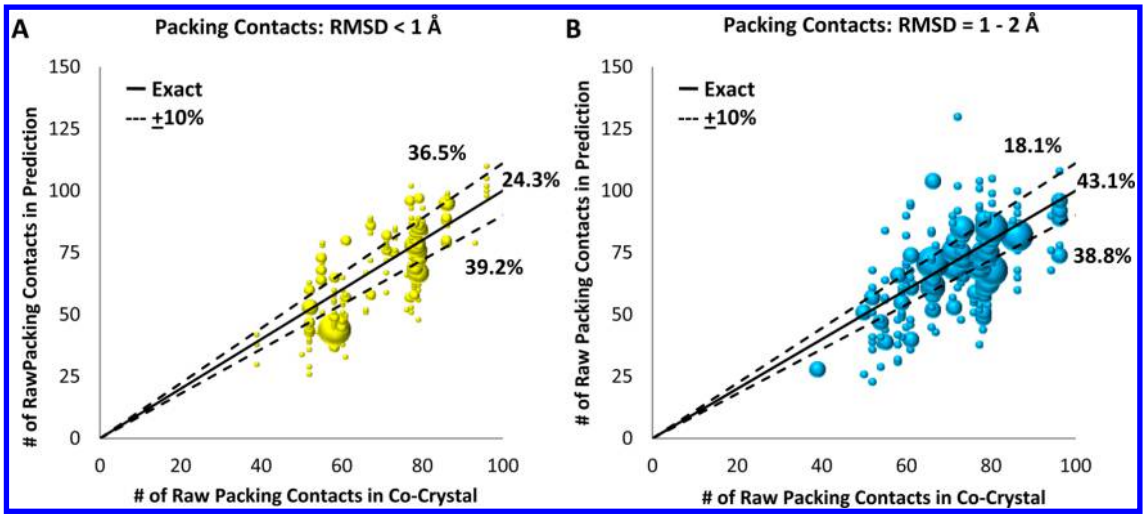


Figure 8. Number of raw packing contacts in co-crystal versus number of raw packing contacts in prediction. The solid line illustrates a perfect match, while the dotted lines show a ±10% range. (A) RMSD < 1 Å bin. (B) RMSD = 1−2 Å bin.

dx.doi.org/10.1021/ci400025f | J. Chem. Inf. Model. 2013, 53, 1853−1870

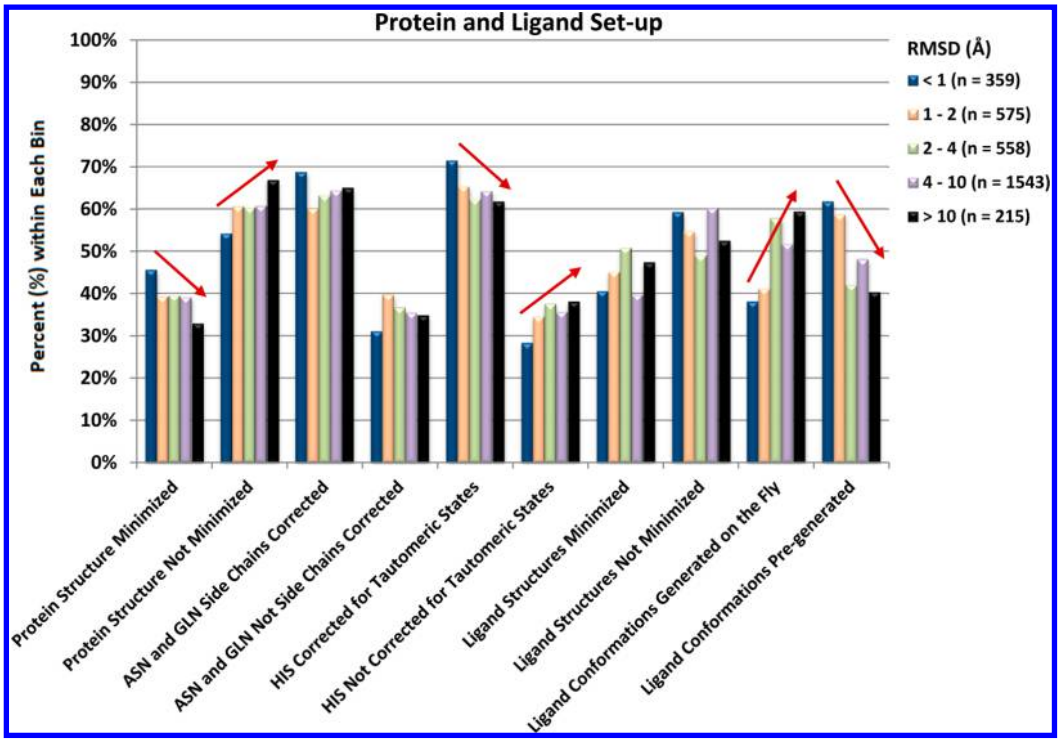


Figure 9. Outcome of the online questionnaire on protein and ligand setup for all poses. The pose prediction results were binned by RMSD and plotted as the percentage of time that a particular feature resulted in a pose within the RMSD bin. Distinct trends that are related to docking RMSD are noted with arrows.

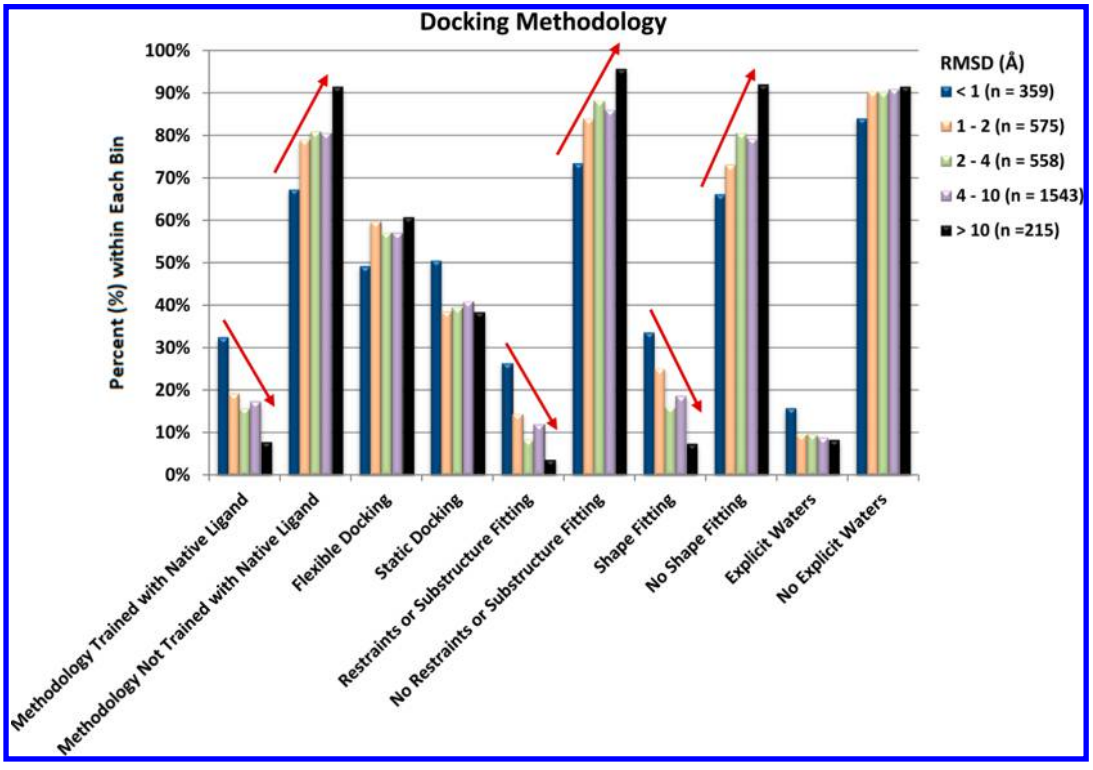


Figure 10. Outcome of the online questionnaire on docking methodology for all poses. The pose prediction results were binned by RMSD and plotted as the percentage of time that a particular feature resulted in a pose within the RMSD bin. Distinct trends that are related to docking RMSD are noted with arrows.

prediction results were binned by RMSD, and the percentage of time that a particular feature resulted in a pose within the RMSD bin is presented. It is important to note that although we received a 100% response rate, some participants did not answer every question. The data presented is for "all poses" as "best pose" when binned by RMSD did not always have enough data points to be statistically significant. Figure 9 shows how the various components of protein and ligand setup trend with

dx.doi.org/10.1021/ci400025f | J. Chem. Inf. Model. 2013, 53, 1853−1870

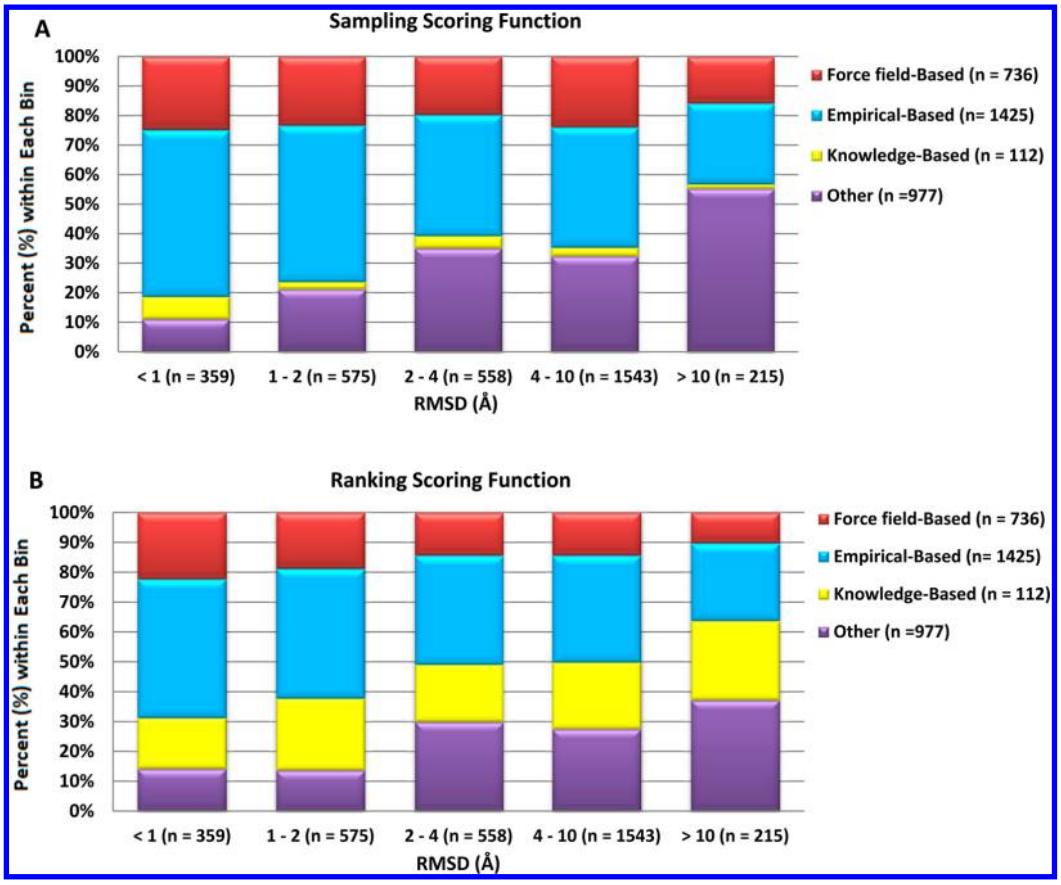


Figure 11. Outcome of the online questionnaire on scoring functions for all poses. The percentage of time that a scoring function was utilized is shown by RMSD bin. Distinct trends that are related to docking RMSD are noted with arrows.

the RMSD. Here, minimizing the protein and correcting the histidine tautomeric state had a positive effect on the docking results, while minimizing the ligand appeared to have a less positive effect. As previously mentioned, many groups that participated in the 2011 Docking and Scoring Challenge also found that optimizing the protein structure prior to docking improved their performance.[27−29] Most likely, this creates an internally consistent environment as the protein is now on the same energy landscape as the scoring function used in docking. Furthermore, scoring functions are typically parametrized for proteins but not for ligands, which may result in unrealistic ligand conformations such as bent aromatic rings. Lastly, pregenerated ligand conformations had better results than those that were generated on the fly. This may suggest an issue with ligand sampling.

Figure 10 shows the same analysis for the docking methodology that was employed by each group and how it trends with the RMSD. The data revealed that first training with the native ligand to determine optimal docking parameters significantly improved the docking performance as did using restraints, substructure fitting, and shape fitting. This data implies that results can be enriched when prior information about the system is known. Furthermore, marrying ligand- and structure-based approaches has been an area of active research in the field recently, and these hybrid techniques have been shown to outperform the use of structure-based methods on their own.[65−67] Interestingly, the use of special parameters for the catalytic zinc in LpxC did not improve docking performance (data not shown).

The type of scoring function utilized for both sampling and ranking was also analyzed, as shown in Figure 11. Many groups utilize their own scoring function or a combination of the different types, and these would fall into our "other" category. When empirical scoring functions are used as either the sampling or ranking scoring function, they appear to have a positive effect on the docking results as demonstrated by Figure 11A and B. However, the "other" category negatively affected pose prediction when utilized as the sampling or ranking scoring function. In general, the "other" category was primarily hybrid-type scoring functions where two or more types were combined. Knowledge-based scoring functions performed fairly consistent across the RMSD bins for both sampling and ranking scoring functions. Here, consistently means that this particular scoring function did not seem to have a negative or positive effect on the pose prediction performance. The force field-based scoring function was also fairly consistent across the RMSD bins when utilized as the sampling scoring function but appeared to have a positive effect when used as the ranking scoring function. The top three performing groups (best pose) all utilized empirical-based scoring functions for sampling, and two of the three also used empirical-based for ranking (the third was force field-based). One of the bottom three performing groups employed a force field-based scoring function for both sampling and scoring. The second used a shape and functionality-based complementarity, and the third used a crude shape-based complementarity for sampling and knowledge-based for scoring.

How well did methods perform overall on identifying actives from inactives and relative ranking? Tables 6 and

dx.doi.org/10.1021/ci400025f | J. Chem. Inf. Model. 2013, 53, 1853−1870

Table 6. Pearson r Parametric Correlation between the Predicted Scores and Experimental Binding Affinities by Protein for Group−Methods That Submitted Scores for All Ligands of LpxC, Urokinase, Chk1, and Erk2[a]

group−method	LpxC (3 lig)	Urokinase (15 lig)	Chk1 (29 lig)	Erk2 (38 lig)	sum r (max = 4.00)
median	**0.78**	**0.50**	−0.01	0.37	1.64
molecular wt	0.37	0.42	−0.14	0.40	1.05
SlogP	−0.51	0.06	−0.33	−0.15	−0.93
group A-1	−0.84	−0.28	0.01	0.11	−1.00
group C-1	**0.76**	**0.71**	0	0.30	1.77
group C-2	**0.73**	**0.71**	−0.12	0.34	1.66
group E-1	**0.87**	0.45	0.09	**0.66**	**2.07**
group E-2	**0.87**	0.45	0.01	**0.57**	1.90
group G-1	**1.00**	0.09	0.16	0.41	1.66
group H-1	**0.99**	0.32	0.02	0.46	1.79
group H-2	**0.92**	0.23	0.02	0.47	1.64
group I-1	**0.94**	0.48	−0.12	0.26	1.56
group I-2	**0.92**	0.49	−0.09	0.31	1.63
group J-1	**1.00**	0.35	0.30	−0.02	1.63
group L-1	**0.61**	**0.77**	−0.02	−0.02	1.34
group L-2	0.41	**0.60**	0.02	0.09	1.12
group M-1	0.67	0.29	−0.15	0.23	1.04
group M-2	0.31	0.33	−0.26	0.41	0.79
group M-3	0.40	0.47	−0.13	0.16	0.90
group M-4	0.20	0.43	−0.31	0.33	0.65
group N-1	−0.56	0.22	−0.12	0.03	−0.43
group P-1	**0.99**	0.36	0	0.41	1.76
group P-2	**0.59**	**0.50**	−0.22	**0.63**	1.50
group P-5	**0.99**	**0.72**	−0.03	0.48	**2.16**
group P-6	**0.79**	**0.58**	−0.14	**0.55**	1.78
group R-1	**0.89**	**0.54**	−0.12	0.22	1.53
group R-2	**0.84**	**0.61**	−0.11	0.15	1.49
group S-1	**0.92**	**0.57**	0.38	0.12	1.99
group S-2	**0.77**	**0.60**	0.16	0.44	1.97
group T-1	−0.56	0.17	0.32	−0.07	−0.14
group T-2	**0.85**	0.37	−0.17	0.40	1.45

[a]Values of r greater than 0.50 and sum r values greater than 2.00 are bolded.

Table 7. Spearman ρ Nonparametric Correlation between the Predicted Scores and Experimental Binding Affinities by Protein for Group−Methods That Submitted Scores for All Ligands of LpxC, Urokinase, Chk1, and Erk2[a]

group−method	LpxC (3 lig)	Urokinase (15 lig)	Chk1 (29 lig)	Erk2 (38 lig)	sum ρ (max = 4.00)
median	**0.50**	**0.52**	−0.03	0.31	1.30
molecular wt	**0.50**	0.41	−0.14	0.40	1.17
SlogP	−0.50	0.14	−0.29	−0.15	−0.80
group A-1	−0.87	−0.28	0.02	0.06	−1.07
group C-1	**0.50**	**0.64**	−0.02	0.32	1.44
group C-2	**0.50**	**0.63**	−0.09	0.34	1.38
group E-1	**0.50**	**0.50**	0.03	**0.67**	1.70
group E-2	**0.50**	**0.50**	0.05	**0.58**	1.63
group G-1	**1.00**	0.29	0.18	0.42	1.89
group H-1	**1.00**	0.20	0.06	0.45	1.71
group H-2	**1.00**	0.24	0.01	**0.50**	1.75
group I-1	**1.00**	**0.56**	−0.09	0.22	1.69
group I-2	**1.00**	**0.51**	−0.14	0.30	1.67
group J-1	**1.00**	0.44	0.10	−0.08	1.46
group L-1	**0.50**	**0.76**	0.01	−0.04	1.23
group L-2	**0.50**	**0.72**	0.08	0.05	1.35
group M-1	**0.50**	0.31	−0.16	0.21	0.86
group M-2	**0.50**	0.37	−0.24	0.39	1.02
group M-3	**0.50**	**0.50**	−0.05	0.14	1.09
group M-4	**0.50**	**0.50**	−0.30	0.28	0.98
group N-1	−0.50	0.22	−0.10	0.17	−0.21
group P-1	**1.00**	0.38	−0.04	0.41	1.75
group P-2	**0.50**	**0.52**	−0.22	**0.64**	1.44
group P-5	**1.00**	**0.55**	−0.02	0.45	1.98
group P-6	**0.50**	0.49	−0.11	**0.56**	1.44
group R-1	**0.50**	**0.60**	−0.17	0.24	1.17
group R-2	**0.50**	**0.62**	−0.15	0.14	1.11
group S-1	**1.00**	**0.63**	0.30	0.12	**2.05**
group S-2	**0.50**	**0.57**	0.13	0.40	1.60
group T-1	−0.50	0.32	0.28	−0.02	0.08
group T-2	**0.50**	0.40	−0.14	0.40	1.16

[a]Values of ρ greater than or equal to 0.50 and sum ρ values greater than 2.00 are bolded.

7 show the Pearson r and Spearman ρ, respectively, assessments of the correlation between the predicted scores and the experimental binding affinity for each group−method, broken down by protein target. Only the 28 group−methods that submitted scores for all ligands of LpxC, Urokinase, Chk1, and Erk2 were included in the analysis to ensure a fair evaluation and are shown in Tables 6 and 7. However, for completeness, all groups that submitted rankings are provided in the Supporting Information as well as the Kendall τ correlation. Pearson is parametric and a measure of the linear relationship between scores and binding affinities, while Spearman ρ and Kendal τ are nonparametric and reflect the correlation of the rank ordering of the data. As r compares the absolute values of each prediction, it is a much more difficult assessment metric than ρ and τ. While all correlations are worthwhile to compare, ρ and τ are more appropriate metrics for relative ranking and r for absolute binding affinities.

Overall, most groups did not perform well on relative ranking. This is not surprising as it is well documented that ranking ligands is very difficult.[6,20,33,34] The sum of r and ρ across all proteins is provided as a metric to assess each groups overall performance; a perfect ranking would result in a sum of r or ρ of 4.0 while random would be 0. We interpreted methods with sums of $\geq$2.0 as having good performance. Molecular weight and SlogP were calculated and used as "yardsticks" or null cases to determine a baseline value.[20] The sum of r was 1.05 for molecular weight and −0.93 of SlogP. We also calculated the F-statistic to determine if the fits found for the group−methods were statistically different from the fits found for molecular weight and also for SlogP. For the majority of the group−methods, the linear fits are statistically significant in their difference from the yardsticks both for methods with good correlation and methods with poor correlation.

Here, the maximum sum of ρ was 2.05 (group S-1, which used an empirical-based scoring function), and the minimum sum of ρ was −1.07 (group A-1, which used a force field-based scoring function). Only one of the group−methods resulted in a sum of ρ greater than 2.0, and two were anticorrelated. In summary, across all protein systems, most groups were not able to relatively rank the ligands. The maximum sum of r was 2.16 (group P-5, which used a knowledge-based ranking scoring function), and the minimum sum of r was −0.14 (group T-1, which used a force field-based scoring function). Only 2 of the 28 group−methods were able to attain a score of a sum of r

dx.doi.org/10.1021/ci400025f | J. Chem. Inf. Model. 2013, 53, 1853−1870

Table 8. AUC Values Derived from ROC Curves by Protein for Group−Methods that submitted Scores for All Ligands of LpxC, Urokinase, and Chk1[a]

group−method	LpxC (3 active, 8 nonactive)	Urokinase (15 active, 4 nonactive)	Chk1 (30 active, 9 nonactive)	sum AUCs (max = 3.00)	Urokinase + Chk1 sum AUCs (max = 2.00)
median	0.21	**0.83**	**0.56**	**1.60**	**1.39**
molecular wt	**0.83**	**0.55**	**0.69**	**2.08**	**1.24**
SlogP	0.13	**0.72**	**0.49**	1.33	**1.21**
group A-1	**0.71**	0.25	**0.69**	**1.64**	0.94
group C-1	0.04	**0.97**	**0.59**	**1.60**	**1.56**
group C-2	0.04	**0.78**	**0.60**	1.42	**1.38**
group E-1	0.08	**0.83**	**0.64**	**1.55**	**1.47**
group E-2	0.08	**0.83**	**0.54**	1.46	**1.37**
group G-1	0.33	**0.63**	0.41	1.38	**1.04**
group H-1	0.42	**0.83**	**0.55**	**1.80**	**1.38**
group H-2	0.25	**0.83**	**0.56**	**1.64**	**1.39**
group I-1	0	**0.68**	0.46	1.14	**1.14**
group I-2	0.33	**0.73**	0.38	1.44	**1.11**
group J-1	**0.92**	**0.97**	0.44	**2.32**	**1.41**
group L-1	**0.75**	**0.98**	0.49	**2.22**	**1.47**
group L-2	1	**0.97**	0.47	**2.44**	**1.44**
group M-1	0.13	**0.80**	**0.55**	**1.48**	**1.35**
group M-2	0.42	**0.78**	0.43	**1.63**	**1.21**
group M-3	0.13	**0.82**	0.47	1.41	**1.29**
group M-4	0.42	**0.78**	0.42	**1.62**	**1.21**
group N-1	**0.79**	**0.82**	**0.50**	**2.11**	**1.32**
group P-1	0.38	**0.72**	**0.62**	**1.71**	**1.34**
group P-2	0.21	**0.75**	**0.72**	**1.68**	**1.47**
group P-5	0.33	**0.97**	**0.75**	**2.05**	**1.71**
group P-6	0.21	**0.77**	**0.74**	**1.72**	**1.51**
group R-1	0.17	**0.53**	**0.53**	1.23	**1.06**
group R-2	0.13	**0.57**	**0.52**	1.21	**1.08**
group S-1	**0.79**	**0.83**	0.47	**2.10**	**1.31**
group S-2	**0.67**	**0.90**	0.43	**1.99**	**1.33**
group T-1	**0.92**	0.33	**0.53**	**1.78**	0.87
group T-2	0	**0.72**	**0.66**	1.38	**1.38**

[a]AUC values greater than 0.50 and sum AUC values greater than 1.50 and 1.00 (for Urokinase and Chk1 alone) are bolded.

greater than 2.0, and 3 of the 28 were anticorrelated. Using the rankings from ρ, two of the top three performing groups used empirical-based scoring functions for sampling and scoring, and the third used a crude shape-based complementarity for sampling and knowledge-based for scoring. One of the bottom three performing groups employed a force field-based scoring function for both sampling and scoring. The second used a shape and functionality-based complementarity, and the third used a combination of knowledge and empirical-based.

This finding illuminates another potential issue: Is the experimental data that the computational field considers to be "gold standard" actually correct? As discussed in our data set paper,[40] we have found that different experimental methods (e.g., Thermofluor, ITC, Octet Red, etc.) can give different values for the same protein−ligand complex, and furthermore, even the relative ranking between a chemical series can be dependent on the various experimental method employed. This is a very troublesome finding as it suggests that the best we may be able to do as a computational field is predict whether a compound is active or inactive but not absolute values or ranking between libraries of compounds. Our techniques are only as good as the data being used for development, and if the experiment data does not agree between methods, then we have no gold standard to judge our predictions. Researchers will need to first check the variance of their data and use only targets with the lowest to parametrize and validate their methods. We calculated correlations between the rankings for each group−method in this exercise and found no correlation, suggesting that inaccurate reference data is not the issue here. Most groups truly found ranking to be a difficult task. In our next benchmark exercise, we will try to build a data set to address this issue in more detail.

Table 8 shows the results for the enrichment or discriminating actives from inactives portion of the exercise, again only for the 28 group−methods that submitted ranks for all ligands (all group−methods are provided in the Supporting Information). A high AUC indicates that actives were clearly identified over inactives. The sum of the AUC is provided as a metric to assess each groups overall performance; a perfect ranking would result in a sum of 3.00, while random ranking would be 1.50 (there were no inactives for Erk2). Here, the evaluation was conducted again using only group−methods that sent in scores for all ligands of LpxC, Urokinase, and Chk1. The maximum sum AUC was 2.44 (group L-2, which used an empirical-based ranking scoring function), and the minimum sum AUC was 1.14 (group I-1, which used an empirical-based ranking scoring function). Thirteen of the 28 group−methods performance was less than random (sum AUC less than 1.50). All of the top three performing groups utilized empirical-based scoring functions for both sampling and scoring. Two of the

dx.doi.org/10.1021/ci400025f | J. Chem. Inf. Model. 2013, 53, 1853−1870

bottom three performing groups employed a combination of force field- and empirical-based scoring functions for both sampling and scoring, and the third used empirical-based.

We also conducted this analysis without LpxC as Urokinase and Chk1 have more ligands and suffer less from small-number statistical issues. For just Urokinase and Chk1, the sum of the median AUC was 1.39 (here perfect would be 2.00 and random would be 1.00). Only 3 of the 28 group−methods had sum AUCs greater than random, and 3 of the 28 were less than random. The majority of them were close to random. Similar to relative ranking, most methods were not able to do enrichment well across the board. Furthermore, no group−method performed the best in all three categories of pose prediction, discriminating actives from inactives, and relative ranking. The scoring functions utilized in each method have their own strengths and weaknesses that do not appear to be robust across the evaluation exercises.

Which test sets were most challenging for identifying actives from inactives and relative ranking? When using a Pearson correlation as the evaluation metric, the group−methods performed the best on LpxC (median r was 0.78), as provided in Table 6. However, when applying a nonparametric correlation coefficient (Table 7), the maximum median ρ was 0.52 for Urokinase, while LpxC was very close with a median ρ of 0.50. For this analysis, LpxC stands out as the protein system that group−methods performed the best at, while Chk1 appears to be the hardest test case. For the most part, the correlations found were not any better than random as compared to the null test cases. The three chemical series within the Chk1 and Erk2 ligands were also broken up (data is provided in the Supporting Information). For both Chk1 and Erk2, the correlations do improve within a few of the series but are still essentially not any better than random. Relative ranking is a difficult task and one where much work is needed.

When evaluating the group−methods on each protein system independently, the performance by the various group−methods was much better. Here, the median AUC for Urokinase was 0.83, indicating that groups were able to discriminate Urokinase actives from inactives quite well, as shown in Table 8. Chk1 followed with a median AUC of 0.56 (essentially random), and the median AUC of LpxC was 0.21 (worse than random). The three chemical series within the Chk1 ligands were also broken up, and the median AUCs were found to be 0.61 for series 1, 0.69 for series 2, and 0.56 for series 3 (tables are provided in the Supporting Information). Unlike with pose prediction, group−methods were still not able to discriminate actives from inactives within the three chemical series. However, there were cases where actives and inactives within a series could be identified, but overall performance suffered because inactives in one series outranked actives of another series.

Are scoring functions able to identify actives from inactives better than relative ranking or vice versa? Four protein systems were employed in the exercise. However, LpxC was too small of a data set to draw a statistically significant conclusion, and Erk2 did not have any inactives. As such, Chk1 and Urokinase were examined to determine if scoring functions were better at enrichment or relative ranking in this benchmark exercise. Unfortunately, most methods were not able to do either very well for Chk1 (data not shown). Figure 12 shows the ranking results (Pearson r or Spearman ρ correlations) plotted against enrichment results (AUC) for Urokinase. An AUC of less than 0.50 is considered random, while negative values of r or ρ signify that the rankings were anticorrelated.

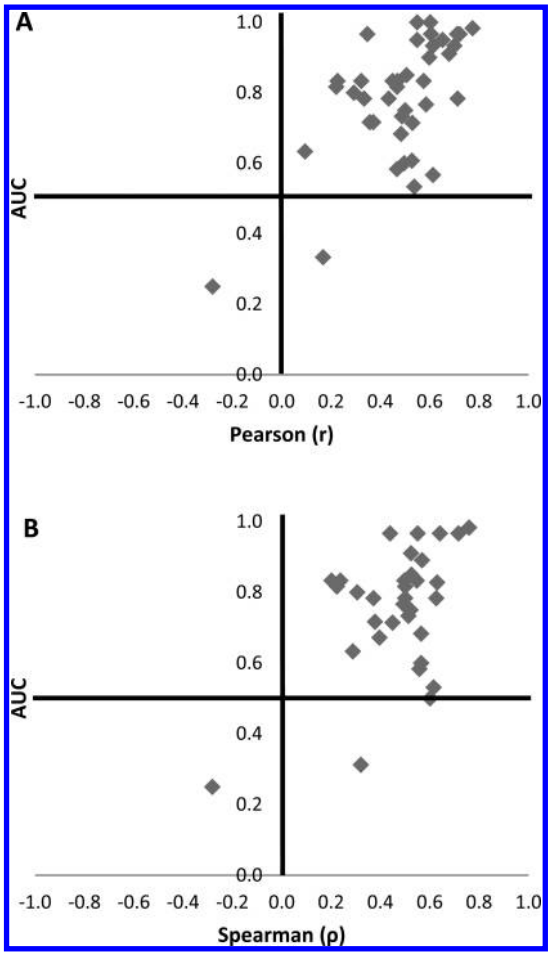


Figure 12. For the Urokinase test set, the ability to rank active molecules versus enriching hit lists is plotted. An AUC of less than 0.50 is considered random. Negative values of r or ρ signify that the data was anticorrelated. (A) Pearson r parametric correlation versus AUC and (B) Spearman ρ nonparametric correlation versus AUC.

Here, we see that the trend is the same whether a parametric or nonparametric correlation is applied; most groups are able to predict the active Urokinase ligands from the inactives better than rank the active molecules. At a value of AUC equal to 1.0 (perfect enrichment), there is a large spread of r and ρ. However, we see that at high values of r and ρ, the AUC is close to 1.0. This demonstrates that if a group is able to rank well, then they are usually able to discriminate actives from inactives well but not vice versa. Ranking is typically thought of as the harder task of the two, so if the scoring function is fined tuned enough to rank, then it should also be able to do enrichment.

How do predicted poses correlate with ranking? Do the programs typically get the pose correct when the ranking is correct or vice versa? Docking programs are faced with two major tasks: (1) predicting the pose of the ligand in the presence of the protein and (2) scoring the predicted the pose. Although two separate tasks, they should be correlated, which brings about the question "Are scoring functions getting the rankings correct for the right reasons?". One would assume that if the pose is ranked the highest, it should be the pose seen in the crystal structure. In Figure 13, the RMSD of the top-ranking pose for molecule X is shown versus the percentage of *inactive* molecules ranked higher than molecule X for both Urokinase and Chk1 targets (Was the scoring function able to rank the active molecule higher than

dx.doi.org/10.1021/ci400025f | J. Chem. Inf. Model. 2013, 53, 1853−1870

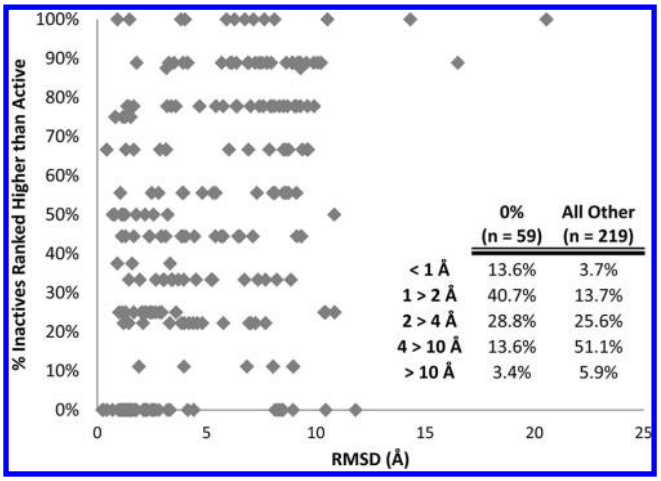


Figure 13. RMSD is plotted against the percentage of inactive molecules ranked higher than an active molecule for both Urokinase and Chk1 targets. The insert shows the percentage of ligands that fall with each RMSD bin for two groups: (1) active molecules that have no inactives ranked higher (0%) and (2) active molecules that have one or more inactives ranked higher (all other).

the inactive molecules, and if not, how many inactive molecules were ranked higher?). Again, only Urokinase and Chk1 were used because of the reasons stated above. While there is a spread in the data suggesting that the scoring function is not always ranking for the correct reason, there also is a large group of molecules (13.6%) near the 0,0 point on the graph. There are multiple reasons that scoring functions may be ranking an incorrect pose higher than the correct pose. First, it may be due to incorrect capturing of the contacts being made between the protein and ligand. Additionally, it may be because of the terms that are not explicitly accounted for in the scoring function (e.g., entropy and/or solvation are typically not included).

The data was also binned by RMSD for two groups: (1) active molecules that have no inactives ranked higher (0%) and (2) active molecules that have one or more inactives ranked higher (all other). Of the "0%" group, 54.2% of the poses had a RMSD of less than 2 Å, the cutoff for a successful docking prediction. For the "all other" group, only 17.4% fall within this cutoff. It appears from this data that the correct pose is not a necessity for ranking correctly, but there is a better chance to be scored correctly if the pose is correct as well.

CONCLUSION

In this benchmark exercise, participants were asked to compare different improvements for pose prediction, enrichment, and relative ranking of congeneric series of compounds across four protein targets. Here, we have provided a thorough analysis across all groups' results to determine common limitations to many docking programs to help the field prioritize where effort should be made. Additionally, much emphasis was placed on the pose prediction evaluation metrics to help set standards in the field. When developing computational methods, proper evaluation of the results is just as important as the high quality experimental data used in the data set.

Using best pose, the median RMSD across all group−methods was 3.0 Å, and 37% of group−methods had a median RMSD < 2 Å. LpxC and Urokinase had the smallest median RMSD with Chk1 following, and Erk2 was the most challenging. Native contacts are exponentially correlated to RMSD. Additionally, they provide a breakdown of contact types (Het−Het vs C−C) and information on atom packing. For all proteins combined, raw Het−Het and packing contacts were underpredicted and raw C−C contacts were both overpredicted and underpredicted at the same rate. No correlations were found between the pose prediction metrics and chemical properties or size of ligand. For protein and ligand setup, minimizing the protein and correcting the histidine tautomeric state had a positive effect on the docking results, while minimizing the ligand appeared to have a less positive effect. Pregenerated ligand conformations had better results than those that were generated on the fly. Additionally, first training with the native ligand to determine optimal docking parameters significantly improved the docking performance as did using restraints, substructure fitting, and shape fitting. Lastly, for both sampling and ranking scoring functions, the use of the empirical scoring function appeared to have a positive effect on the docking results, while the "other" category negatively affected pose prediction.

For the most part, methods were not very successful at relative ranking or enrichment. The sum of the median ρ was 1.28/4.00, r was 1.65/4.00, and the sum of the median AUCs was 1.60/3.00. For relative ranking, group−methods performed the best on LpxC, and Chk1 was found to be the most challenging. In the enrichment study, Urokinase proved to be the most straightforward, while LpxC was the most difficult. Compared to relative ranking, group−methods were able to identify actives from inactives better for Urokinase. However, LpxC was too small of a data set to draw conclusions, and for Chk1, group−methods were not able to do either very well. Lastly, the correct pose is not a necessity for ranking correctly, but there is a better chance to be scored correctly if the pose is correct as well.

Future benchmark exercises from CSAR will involve ranking pregenerated poses to separate the two major components of the docking algorithm and, hence, focus only on the ability of the scoring function to correctly rank the "real" binding pose. As always, we will strive to conduct a blinded exercise with as many systems and ligands as possible to not only avoid system dependent insights but also to provide statistical significance to the results. We greatly appreciate the efforts of all of our colleagues in the pharmaceutical industry for the donation of this data and for future benchmark exercises.

ASSOCIATED CONTENT

Supporting Information

(1) RMSD box plot of the best pose for each protein−ligand complex broken down by group method for each protein target, (2) percent of %Total contacts within various RMSD bins and broken down by various %Total contact cut-offs, (3) number of raw Het−Het contacts in co-crystal versus number of raw Het−Het contacts in prediction, (4) number of raw C−C contacts in co-crystal versus number of raw C−C contacts in prediction, (5) number of raw packing contacts in co-crystal versus number of raw packing contacts in prediction, (6) Pearson, Spearman, and Kendall correlations between ligand descriptors and pose prediction metrics (RMSD and %Native contacts correct), (7) Pearson r parametric correlation between the predicted scores and experimental binding affinities by protein for all group methods, (8) Spearman ρ nonparametric correlation between the predicted scores and experimental binding affinities by protein for all group methods, (9) Kendall τ nonparametric correlation between the predicted scores and experimental binding affinities by protein for all group methods, (10)

Pearson r parametric correlation between the predicted scores and experimental binding affinities by chemical series for all group methods, (11) Spearman ρ nonparametric correlation between the predicted scores and experimental binding affinities by chemical series for all group methods, (12) Kendall τ nonparametric correlation between the predicted scores and experimental binding affinities by chemical series for all group methods, (13) AUC values derived from ROC curves by protein for all group methods, and (14) AUC values derived from ROC curves by all series of Chk1 for all group methods. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: 734-615-6841. Fax: 734-763-2022. E-mail: carlsonh@umich.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank all participants in this year's benchmark exercise. Whether you submitted a paper for the upcoming CSAR special issue of the Journal of Chemical Information and Modeling, gave a talk at the symposium, submitted scores for this analysis, or just attended the talks and the discussions at the symposium, everyone's feedback was valuable to our efforts. We thank numerous colleagues for helpful discussions, particularly Greg Warren (OpenEye) for his insights on creating the exercise data sets. Additionally, we thank the CSAR advisory board for their valuable comments and feedback. The CSAR Center is funded by the National Institute of General Medical Sciences (U01 GM086873). We also thank the Chemical Computing Group and OpenEye Scientific Software for generously donating the use of their software.

REFERENCES

(1) Cheng, T.; Li, Q.; Zhou, Z.; Wang, Y.; Bryant, S. H. Structure-based virtual screening for drug discovery: A problem-centric review. *AAPS J.* **2012**, *14*, 133−141.
(2) Huang, S. Y.; Zou, X. Advances and challenges in protein−ligand docking. *Int. J. Mol. Sci.* **2010**, *11*, 3016−3034.
(3) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813−1818.
(4) 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.
(5) Lyne, P. D. Structure-based virtual screening: An overview. *Drug. Discovery Today* **2002**, *7*, 1047−1055.
(6) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.
(7) Carlson, H. A. Protein flexibility and drug design: How to hit a moving target. *Curr. Opin. Chem. Biol.* **2002**, *6*, 447−452.
(8) Carlson, H. A. Protein flexibility is an important component of structure-based drug discovery. *Curr. Pharm. Des.* **2002**, *8*, 1571−1578.
(9) Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. A. Target flexibility: an emerging consideration in drug discovery and design. *J. Med. Chem.* **2008**, *51*, 6237−6255.
(10) Damm, K. L.; Carlson, H. A. Exploring experimental sources of multiple protein conformations in structure-based drug design. *J. Am. Chem. Soc.* **2007**, *129*, 8225−8235.
(11) Durrant, J. D.; McCammon, J. A. Computer-aided drug-discovery techniques that account for receptor flexibility. *Curr. Opin. Pharmacol.* **2010**, *10*, 770−774.
(12) Jain, A. N. Effects of protein conformation in docking: Improved pose prediction through protein pocket adaptation. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 355−374.
(13) Spyrakis, F.; BidonChanal, A.; Barril, X.; Luque, F. J. Protein flexibility and ligand recognition: Challenges for molecular modeling. *Curr. Top. Med. Chem.* **2011**, *11*, 192−210.
(14) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.* **2009**, *49*, 1079−1093.
(15) Huang, S. Y.; Grinter, S. Z.; Zou, X. Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys.* **2010**, *12*, 12899−12908.
(16) Jain, A. N. Scoring functions for protein−ligand docking. *Curr. Protein Pept. Sci.* **2006**, *7*, 407−420.
(17) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. *Br. J. Pharmacol.* **2008**, *153* (Suppl 1), S7−26.
(18) Pham, T. A.; Jain, A. N. Customizing scoring functions for docking. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 269−286.
(19) Dunbar, J. B., Jr.; Smith, R. D.; Yang, C. Y.; Ung, P. M.; Lexa, K. W.; Khazanov, N. A.; Stuckey, J. A.; Wang, S.; Carlson, H. A. CSAR benchmark exercise of 2010: Selection of the protein−ligand complexes. *J. Chem. Inf. Model.* **2011**, *51*, 2036−2046.
(20) Smith, R. D.; Dunbar, J. B., Jr.; Ung, P. M.; Esposito, E. X.; Yang, C. Y.; Wang, S.; Carlson, H. A. CSAR benchmark exercise of 2010: Combined evaluation across all submitted scoring functions. *J Chem Inf Model* **2011**, *51*, 2115−2131.
(21) Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133−139.
(22) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.
(23) 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.
(24) Jain, A. N. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499−511.
(25) Jain, A. N. Surflex-Dock 2.1:Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 281−306.
(26) Warren, G. M.; McGaughey, G. B.; Nevins, N. Editorial. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 674.
(27) Corbeil, C. R.; Williams, C. I.; Labute, P. Variability in docking success rates due to dataset preparation. *J. Comput.-Aided Mol .Des.* **2012**, *26*, 775−786.
(28) Repasky, M. P.; Murphy, R. B.; Banks, J. L.; Greenwood, J. R.; Tubert-Brohman, I.; Bhat, S.; Friesner, R. A. Docking performance of the glide program as evaluated on the Astex and DUD datasets: A complete set of glide SP results and selected results for a new scoring function integrating WaterMap and glide. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 787−799.
(29) Spitzer, R.; Jain, A. N. Surflex-Dock: Docking benchmarks and real-world application. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 687−699.
(30) Guthrie, J. P. A blind challenge for computational solvation free energies: Introduction and overview. *J. Phys. Chem. B.* **2009**, *113*, 4501−4507.

(31) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. Predicting small-molecule solvation free energies: An informal blind test for computational chemistry. *J. Med. Chem.* **2008**, *51*, 769−779.
(32) Skillman, A. G.; Geballe, M. T.; Nicholls, A. SAMPL2 challenge: Prediction of solvation energies and tautomer ratios. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 257−258.
(33) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455−1474.
(34) Kim, R.; Skolnick, J. Assessment of programs for ligand binding affinity prediction. *J. Comput. Chem.* **2008**, *29*, 1316−1331.
(35) Cole, J. C.; Murray, C. W.; Nissink, J. W.; Taylor, R. D.; Taylor, R. Comparing protein−ligand docking programs is difficult. *Proteins* **2005**, *60*, 325−332.
(36) Nissink, J. W.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein−ligand interaction. *Proteins* **2002**, *49*, 457−471.
(37) Diago, L. A.; Morell, P.; Aguilera, L.; Moreno, E. Setting up a large set of protein-ligand PDB complexes for the development and validation of knowledge-based docking algorithms. *BMC Bioinf.* **2007**, *8*, 310.
(38) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein−ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726−741.
(39) Warren, G. L.; Do, T. D.; Kelley, B. P.; Nicholls, A.; Warren, S. D. Essential considerations for using protein−ligand structures in drug discovery. *Drug Discovery Today* **2012**, *17*, 1270−1281.
(40) Dunbar, J. B., Jr.; Smith, R. D.; Damm-Ganamet, K. L.; Ahmed, A.; Esposito, E. X.; Delproposto, J.; Chinnaswamy, K.; Kang, Y.-N.; Kubish, G.; Gestwicki, J. E.; Stuckey, J. A.; Carlson, H. A. CSAR Data Set Release 2012: Ligands, affinities, complexes, and docking decoys. *J. Chem. Inf. Model.* **2013**, DOI: 10.1021/ci4000486.
(41) Hawkins, P. C.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: Pitfalls and traps. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 179−190.
(42) Goto, J.; Kataoka, R.; Hirayama, N. Ph4Dock: pharmacophore-based protein−ligand docking. *J. Med. Chem.* **2004**, *47*, 6804−6811.
(43) Hawkins, P. C.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer generation with OMEGA: Algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.
(44) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): A novel method for analyzing three-dimensional protein−ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337−344.
(45) Kroemer, R. T.; Vulpetti, A.; McDonald, J. J.; Rohrer, D. C.; Trosset, J. Y.; Giordanetto, F.; Cotesta, S.; McMartin, C.; Kihlen, M.; Stouten, P. F. Assessment of docking poses: Interactions-based accuracy classification (IBAC) versus crystal structure deviations. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 871−881.
(46) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195−207.
(47) Yusuf, D.; Davis, A. M.; Kleywegt, G. J.; Schmitt, S. An alternative method for the evaluation of docking performance: RSR vs RMSD. *J. Chem. Inf. Model.* **2008**, *48*, 1411−1422.
(48) Swets, J. A.; Dawes, R. M.; Monahan, J. Better decisions through science. *Sci. Am.* **2000**, *283*, 82−87.
(49) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*, 2534−2547.
(50) Lee, C. J.; Liang, X.; Chen, X.; Zeng, D.; Joo, S. H.; Chung, H. S.; Barb, A. W.; Swanson, S. M.; Nicholas, R. A.; Li, Y.; Toone, E. J.; Raetz, C. R.; Zhou, P. Species-specific and inhibitor-dependent conformations of LpxC: Implications for antibiotic design. *Chem. Biol.* **2011**, *18*, 38−47.
(51) Wendt, M. D.; Rockway, T. W.; Geyer, A.; McClellan, W.; Weitzberg, M.; Zhao, X.; Mantei, R.; Nienaber, V. L.; Stewart, K.; Klinghofer, V.; Giranda, V. L. Identification of novel binding interactions in the development of potent, selective 2-naphthamidine inhibitors of urokinase. Synthesis, structural analysis, and SAR of N-phenyl amide 6-substitution. *J. Med. Chem.* **2004**, *47*, 303−324.
(52) Tong, Y.; Claiborne, A.; Stewart, K. D.; Park, C.; Kovar, P.; Chen, Z.; Credo, R. B.; Gu, W. Z.; Gwaltney, S. L., 2nd; Judge, R. A.; Zhang, H.; Rosenberg, S. H.; Sham, H. L.; Sowin, T. J.; Lin, N. H. Discovery of 1,4-dihydroindeno[1,2-c]pyrazoles as a novel class of potent and selective checkpoint kinase 1 inhibitors. *Bioorg. Med. Chem.* **2007**, *15*, 2759−2767.
(53) Aronov, A. M.; Tang, Q.; Martinez-Botella, G.; Bemis, G. W.; Cao, J.; Chen, G.; Ewing, N. P.; Ford, P. J.; Germann, U. A.; Green, J.; Hale, M. R.; Jacobs, M.; Janetka, J. W.; Maltais, F.; Markland, W.; Namchuk, M. N.; Nanthakumar, S.; Poondru, S.; Straub, J.; ter Haar, E.; Xie, X. Structure-guided design of potent and selective pyrimidylpyrrole inhibitors of extracellular signal-regulated kinase (ERK) using conformational control. *J. Med. Chem.* **2009**, *52*, 6362−6368.
(54) Damm, K. L.; Carlson, H. A. Gaussian-weighted RMSD superposition of proteins: A structural comparison for flexible proteins and predicted protein structures. *Biophys. J.* **2006**, *90*, 4558−4573.
(55) *Molecular Operating Environment (MOE)*, version 2010.10; Chemical Computing Group: Montreal, Canada, 2010.
(56) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(57) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FLEXX incremental construction algorithm for protein−ligand docking. *Proteins* **1999**, *37*, 228−241.
(58) Winn, M. D.; Ballard, C. C.; Cowtan, K. D.; Dodson, E. J.; Emsley, P.; Evans, P. R.; Keegan, R. M.; Krissinel, E. B.; Leslie, A. G.; McCoy, A.; McNicholas, S. J.; Murshudov, G. N.; Pannu, N. S.; Potterton, E. A.; Powell, H. R.; Read, R. J.; Vagin, A.; Wilson, K. S. Overview of the CCP4 suite and current developments. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2011**, *67*, 235−242.
(59) Kleywegt, G. J.; Harris, M. R.; Zou, J. Y.; Taylor, T. C.; Wahlby, A.; Jones, T. A. The Uppsala Electron Density Server. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 2240−2249.
(60) Lei, S.; Smith, M. R. Evaluation of several nonparametric bootstrap methods to estimate confidence intervals for software metrics. *IEEE Trans. Software Eng.* **2003**, *29*, 996−1004.
(61) *JMP*, version 9.0.0; SAS Institute, Inc.: Cary, NC.
(62) Bonett, D. G.; Wright, T. A. Sample size requirements for estimating Pearson, Kendall and Spearman correlations. *Psychometrika* **2000**, *65*, 23−28.
(63) Long, J. D.; Cliff, N. Confidence intervals for Kendall's tau. *Br. J. Math. Stat. Psychol* **1997**, *50*, 31−41.
(64) R Development Core Team *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2009.
(65) Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74−82.
(66) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular shape and medicinal chemistry: A perspective. *J. Med. Chem.* **2010**, *53*, 3862−3886.
(67) Sastry, G. M.; Dixon, S. L.; Sherman, W. Rapid shape-based ligand alignment and virtual screening method based on atom/feature-pair similarities and volume overlap scoring. *J. Chem. Inf. Model.* **2011**, *51*, 2455−2466.