JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

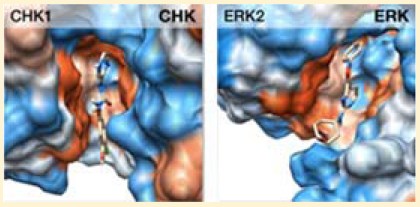
Automated Large-Scale File Preparation, Docking, and Scoring: Evaluation of ITScore and STScore Using the 2012 Community Structure–Activity Resource Benchmark

Sam Z. Grinter,[†,‡] Chengfei Yan,[‡,§] Sheng-You Huang,[‡] Lin Jiang,[‡] and Xiaoqin Zou*[,‡,§,∥]

[†]Informatics Institute, [‡]Dalton Cardiovascular Research Center, [§]Department of Physics & Astronomy, and [∥]Department of Biochemistry, University of Missouri, Columbia, Missouri 65211, United States

Ⓢ *Supporting Information*

ABSTRACT: In this study, we use the recently released 2012 Community Structure–Activity Resource (CSAR) data set to evaluate two knowledge-based scoring functions, ITScore and STScore, and a simple force-field-based potential (VDWScore). The CSAR data set contains 757 compounds, most with known affinities, and 57 crystal structures. With the help of the script files for docking preparation, we use the full CSAR data set to evaluate the performances of the scoring functions on binding affinity prediction and active/inactive compound discrimination. The CSAR subset that includes crystal structures is used as well, to evaluate the performances of the scoring functions on binding mode and affinity predictions. Within this structure subset, we investigate the importance of accurate ligand and protein conformational sampling and find that the binding affinity predictions are less sensitive to non-native ligand and protein conformations than the binding mode predictions. We also find the full CSAR data set to be more challenging in making binding mode predictions than the subset with structures. The script files used for preparing the CSAR data set for docking, including scripts for canonicalization of the ligand atoms, are offered freely to the academic community.



1. INTRODUCTION

The prospect of reliably predicting protein–ligand interactions has important implications for studies of protein function at the molecular level and for the design of therapeutic interventions.[1−4] Abundant, publicly accessible databases containing accurate protein–ligand structures and binding affinity information are invaluable tools to assess and improve the docking and scoring methods used to predict protein–ligand interactions.[5,6] The 2012 Community Structure–Activity Resource (CSAR) data set, publicly released on July third, contains 757 compounds with provided SMILES strings, 508 with known affinity values, 185 compounds designated to be inactive, and 57 compounds with an available protein–ligand crystal structure. These compounds span six protein targets. The compound affinities are given as either K_d, K_i, or IC_{50} and come from several different assays. For compounds with an available protein–ligand crystal structure, the CSAR data set provides the complex, the separated ligand and protein in their native bound conformations, and a set of unbound ligand conformations generated as MOL2 files. In this work, we use the 2012 CSAR data set to evaluate two knowledge-based scoring functions recently developed by our laboratory, ITScore and STScore. We also evaluate a Lennard-Jones potential as a point of reference.

ITScore[7,8] was developed using a statistical mechanics-based iterative method to deal with the challenging reference state problem[9,10] faced by knowledge-based scoring functions.[11−14] The approach starts by computing the native distance distributions for each atom pair type observed in the native ligand binding poses. This distribution is then compared to the same distribution but generated using a set of many decoy poses. Within the distribution that uses decoy ligand poses, each decoy is given a Boltzmann weight based on the free energy predicted for that decoy. The pair potentials can then be adjusted in a way that decreases the predicted free energy of the ligand in its native pose relative to the nearby decoy poses. This adjustment is done iteratively until the pair potentials are able to reliably distinguish the native ligand poses from the decoys.

STScore is a knowledge-based scoring function that we developed to introduce a new method of handling the sparse data problem (not yet published). The overall approach is to combine a potential of mean force (PMF) with a simple force-field-based potential where the relative weight given to either alternative is a function of their estimated inaccuracies. We use a Bayesian statistical model to estimate the inaccuracies in the PMF due to sparse count data, allowing the method to naturally increase the influence of the force-field-based alternative for any pairs or distances lacking training data. Both the PMF and force-field-based potentials used were simple, but the overall point of STScore is to demonstrate the concept of this sparse data method and to show that the method effectively combines the two component potentials, giving better predictions than either the PMF or force-field-based potential alone. The details

Special Issue: 2012 CSAR Benchmark Exercise

Received: January 21, 2013

dx.doi.org/10.1021/ci400045v | *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

of STScore will be described in a separate manuscript in preparation.

In this work, we use the 2012 CSAR data set to evaluate ITScore[7] and STScore. We also perform the same evaluations on a simple Lennard-Jones 6−12 potential[15,16] as a reference. This potential, labeled VDWScore, uses the same van der Waals radii and well depths assigned for the van der Waals force in the initial potential of ITScore.[7] Consequently, ITScore and STScore are similar to VDWScore for close atom pair distances (e.g., clashes), and STScore is also similar to VDWScore for rare atom pair types.

In the following section, we will discuss the details of our evaluations of these three scoring functions. We will explain in detail how we prepared the 2012 CSAR data set for all the docking calculations and introduce, in the Appendix, some scripts and data files for use by future users of the CSAR data set. The scripts may also be of interest to others dealing with large-scale file preparation for docking. Though similar scripts are frequently used by scientists in the pharmaceutical industry for large-scale docking, the scripts are not available to many academic users. In the Results, we will give special emphasis to the importance of the native ligand and protein conformations and discuss the challenge of handling protein and ligand flexibility in docking.

2. METHODS

We used the 2012 CSAR data set to evaluate two scoring functions, ITScore and STScore. We also performed the same evaluations on a simple force-field-based potential based on the Lennard-Jones 6−12 potential. These three scoring functions have been described in the Introduction. All docking calculations were performed using the MDock 1.2 software package (http://zoulab.dalton.missouri.edu/software.htm).[7,8] MDock uses a rigid sampling method that is based on DOCK 4.[16] First, spheres tangent to the protein surface are generated along the binding site. These are used to generate the initial putative ligand orientations by matching ligand atoms to the sphere points. These initial orientations are then minimized and assessed by the chosen scoring function. ITScore is the default scoring function used in MDock. For the STScore and VDWScore evaluations, the source code of MDock was modified only to include the STScore and VDWScore pair potentials and otherwise left in its original form.

2.1. Summary of Evaluations Performed.

We evaluated the performance of ITScore, STScore, and VDWScore based on binding mode predictions and binding affinity predictions. For binding mode predictions, we computed the heavy-atom root-mean-square deviation (RMSD) between each docked ligand and the native, bound-state ligand. These predictions were then evaluated in terms of the percent success rate (where a prediction is considered successful if the top ranked ligand pose has an RMSD less than 2.0 Å). For binding affinity predictions, we computed the Pearson correlation between the docking scores and the known binding affinities of the compounds. These correlations were computed for three separate groups according to the affinity measure provided for the compound: K_d, K_i, or IC_{50}. We also analyzed the ability of the scoring functions to distinguish between known actives and known inactives and present the data as a set of reciever operating characteristic (ROC) curves.

We analyzed the performance of the three scoring functions using both the native structures, where available, and various structure ensembles. For the ligands, the structure ensembles were generated using the software Omega (OpenEye, Inc.). For the proteins, the structure ensembles consist of either the available structures for a given protein group (for calculation on the full set of compounds in CSAR) or all of the these structures except the one bound to the ligand being docked (for subset of the CSAR data set that includes native protein−ligand structures). Because the full set of compounds includes many compounds with no available structure, we evaluated the scoring functions on the full set using only the Omega-generated ligand ensembles docked to the protein ensembles. For this case we evaluated the binding affinity predictions and the active/inactive compound discrimination (as ROC curves).

For the subset of the CSAR data set with structures, we further generated data for six other cases. Specifically, to each native protein structure we docked (1) the corresponding native ligand conformation, (2) an ensemble of conformations generated by Omega from the connection table of the ligand MOL2 file, and (3) an ensemble consisting of the one native conformation along with the set of Omega-generated conformations. For each protein ensemble, we then docked (4) the ligand in its native, bound conformation, (5) the ensemble of conformations generated by Omega from the MOL2 connection table, and (6) an ensemble of conformations generated by Omega from the SMILES string provided in the CSAR data set. For these six cases we evaluated the performance of the scoring functions on binding affinity predictions and binding mode predictions. These data series are depicted in detail in the Results, and the preparation thereof is described in detail below.

2.2. CSAR Data Set Preparation.

The CSAR data set contains SMILES strings for 757 compounds, 508 of which have known binding affinities. It also contains 57 protein−ligand crystal structures, some with known affinities and some without. In order to efficiently handle the data, we prepared a combined CSV datafile which contains a consistent identifier for each compound and as well as all of its associated nonstructural information. This associated information includes the known binding affinity, and a label specifying whether the affinity is given as K_d, K_i, or IC_{50}. We also included the SMILES string, the type of assay used to measure the affinity, and a three-letter label specifying which protein the compound is associated with: “CDC” for Cyclin-dependent kinase 2 bound to Cyclin A, “CDK” for Cyclin-dependent kinase 2, “CHK” for Checkpoint kinase 1, “ERK” for extracellular signal-regulated kinase 2, “LPX” for LpxC (a zinc-dependent bacterial deacetylase), and “URO” for Urokinase plasminogen activator (a serine protease).

We also included several binary labels for each compound to aid in defining useful subsets. For example, one of the labels specifies if a crystal structure is available containing the compound bound to its associated protein, and other labels specify if the compound is designated to be active or inactive. This labeling strategy allows one to easily apply commands to subsets of the CSAR data set defined according to these labels. The CSV datafile, which will be available at the CSAR Web site (http://www.csardock.org), will be of use to future users of the CSAR data set. More details about this datafile can be found in the Appendix.

We began with the set of protein−ligand complexes provided as MOL2 files in the CSAR data set and the datafile mentioned above. All other files were generated using a combination of shell scripting, Python scripting of Chimera[17] in its command-line format, and the tools distributed with the MDock software package. All statistics were done using R (http://cran.r-project.org).

Within each of the six protein groups, we aligned all of the crystal structures provided in the CSAR data set using Chimera's default MatchMaker settings. We also used Chimera to save the protein and the ligand into separate MOL2 files and to convert these MOL2 files into PDB format. All the ligand files were visually inspected. The CSAR data set includes separated ligand and protein files; however we did not use these files because they were generated from the unaligned structures, which are less convenient for evaluating binding mode predictions.

We used Omega 2.4.3 (OpenEye, Inc.) to generate sets of ligand conformations. For each compound, the SMILES string[18] was used to generate one set of ligand conformations. For each compound with an available crystal structure (57 compounds), the MOL2 connection table was used to generate another set of conformations. The conformations generated by these two methods are the same, except in 11 cases where the SMILES strings contained ambiguous stereochemistry. In these cases, Omega handled this ambiguity by generating extra conformations to explore the stereochemistry. The conformations generated from the MOL2 file connection tables use the native stereochemistry. We used both of these methods because the former method is necessary in order to generate conformations for the full CSAR data set (including both compounds with and without structures) and the latter method is necessary in order to generate conformations having atom ID numbers consistent with the native structure. This ID number consistency makes the binding-mode RMSD calculations easier. Evaluations using both of these types of Omega-generated ligand ensembles give nearly identical results (as shown in the Results section).

When running Omega, we used the `-fromCT` option to ensure that the native coordinates were not being used to generate new conformations and the `-flipper true` option so that SMILES strings with ambiguous stereochemistry would be handled. We set `-strictfrags` to true for accuracy and `-maxconfs` to 100 to keep the computational time reasonable for later docking calculations. (For SMILES strings with ambiguous stereochemistry, this conformation limit applies to each generated isomer). We also used Omega with the `-includeInput` option to provide a MOL2 file of the native conformation but with the atom IDs modified to be consistent with the other generated conformations. In summary, for each compound, Omega was used to produce an ensemble of ligand conformations and a renumbered version of the one native conformation. We also concatenated these conformations to generate an ensemble consisting of the one native conformation along with the set of Omega-generated conformations for docking.

2.3. MDock Docking Preparation.

We used UCSF dms[17] to generate the molecular surface of each protein structure, with the default probe radius of 1.4 Å. We used Sphgen_cpp[16] to generate spheres around the whole surface of each protein. Finally, we used get_sph,[16,19] included with the MDock software package (http://zoulab.dalton.missouri.edu/software.htm), to choose spheres in the vicinity of the protein binding site, which may be defined by a PDB file. We defined this binding site broadly by concatenated all the ligand PDB files from the native aligned structures into one file for each protein group, in order to reduce bias. For native-protein docking, all binding spheres within 3.0 Å of any ligand were included.

For non-native protein docking, we used the protein ensemble method,[19,20] consistent with our approach in the CSAR benchmark exercise. To test this method on the full set, we first produced six protein ensembles by combining all of the included protein structures for each protein group (CDC, CDK, CHK, ERK, LPX, and URO). For example, the URO protein ensemble consists of the seven URO protein structures available within the CSAR data set: URO_4, URO_6, URO_7, URO_8, URO_9, URO_15, and URO_18. To prepare the binding sphere files for these six ensembles, we first used get_sph to generate sphere files specific to each bound structure. For each structure, the corresponding binding sphere file contains only the spheres generated for the protein surface of that structure, restricted to the spheres within 3.0 Å of the native, bound ligand position. Then for each protein ensemble, we concatenated all of the sphere files corresponding to the protein structures that constitute the ensemble. For example, the sphere file for the URO ensemble was made by concatenating the sphere files specific to the seven URO protein structures: URO_4, URO_6, URO_7, URO_8, URO_9, URO_15, and URO_18. Finally, we used clu_sph,[16,19] included with MDock, to remove redundant spheres.

For the structure subset, we did not wish the docking calculations for any compound to benefit from the inclusion of the native protein structure for that compound within the ensemble. We therefore produced a different protein cross ensemble for each of the 56 compounds in the structure subset. Each cross ensemble includes all of the CSAR protein structures in the same protein group, except excluding the one protein structure that was bound to the compound in question. For example, the URO_6 cross ensemble consists of the six proteins structures that were not bound to the URO_6 ligand, that is: URO_4, URO_7, URO_8, URO_9, URO_15, and URO_18. This procedure was not possible for CDC (CDK2 bound to Cyclin A), for which only one crystal structure is available in the CSAR data set, CDC_260. Therefore, it was necessary to exclude CDC_260 from the protein cross-ensemble docking calculations performed on the structure subset, and for consistency, CDC_260 was also excluded from the native protein calculations on the structure subset. To generate a binding sphere file for each protein cross ensemble, we concatenated all of the sphere files corresponding to the protein structures that constitute that cross ensemble. For example, the sphere file for the URO_6 cross ensemble was made by concatenating the sphere files specific to the six protein structures in the URO_6 cross ensemble: URO_4, URO_7, URO_8, URO_9, URO_15, and URO_18. As with the full set ensembles, we used clu_sph[16,19] to remove redundant spheres.

2.4. Scoring Method Evaluation.

For binding mode predictions, we computed the RMSD between all non-hydrogen atoms in each docked ligand and all non-hydrogen atoms in the native, bound ligand. We evaluated the binding mode predictions in terms of the percent success rate. A prediction was considered successful if the RMSD between the docked ligand and the native ligand was less than 2 Å.

To evaluate the binding affinity predictions of the scoring functions, we computed the Pearson correlation between the docked scores and the experimentally determined binding affinities provided in the CSAR data set. Computing score—affinity correlations for all proteins in one group would not be appropriate, because the binding affinities for each group were obtained from assays that use different affinity measures. We

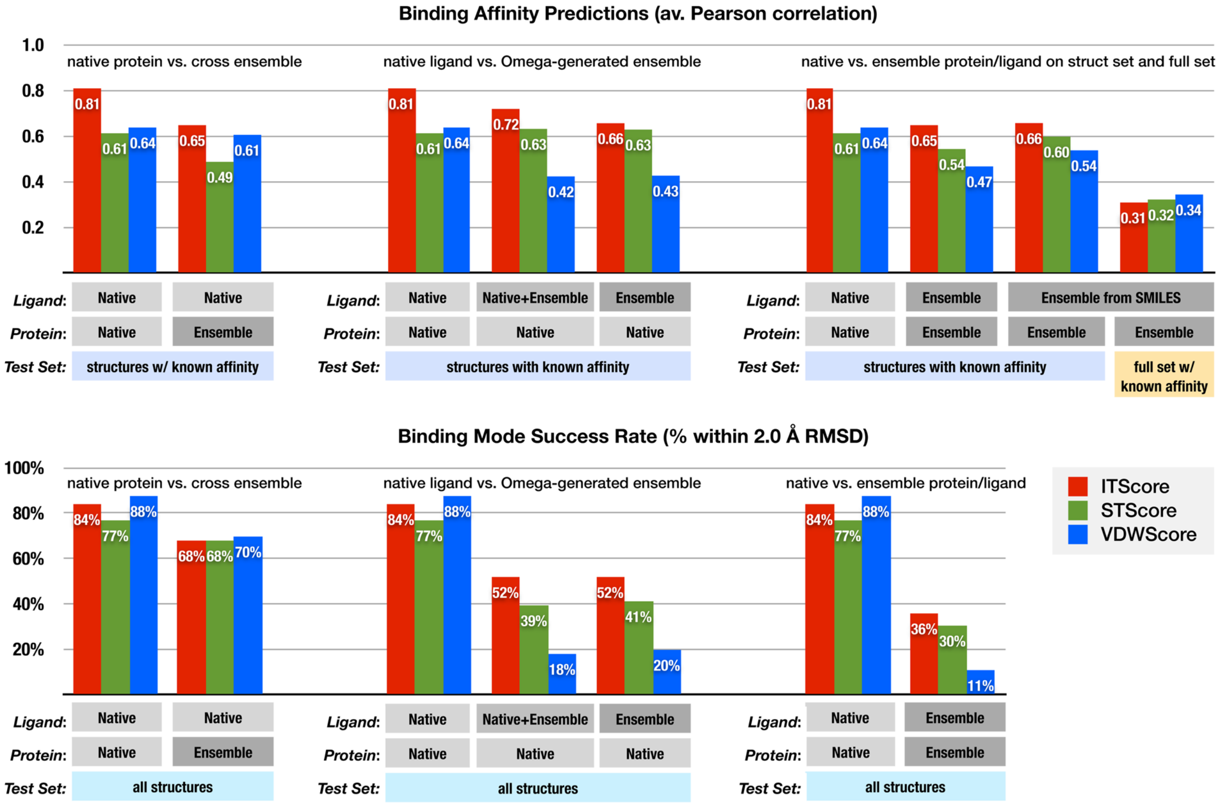


Figure 1. Binding affinity and binding mode predictions of the three scoring functions evaluated in this study: ITScore (red), STScore (green), and VDWScore (blue). The results are divided according to three attributes: the protein conformation (native or ensemble), the ligand conformation (native, native + ensemble, or ensemble), and the test set used (all structures, structures with affinities, or full set with affinities). The top panel gives the binding affinity accuracy. Each *y*-axis value is the mean of three Pearson correlations computed for each affinity group (K_d, K_i, or IC_{50}). The affinity groups are weighted according to the number of different proteins in the group: 2.0 for K_d, 2.0 for K_i, and 1.0 for IC_{50}. The bottom panel gives the binding mode prediction results in terms of percent success rate. The binding mode is considered successfully identified by a scoring function if the lowest-scored binding mode according to that scoring function is within 2.0 Å RMSD of the native binding mode.

therefore computed the correlation separately for each of the three affinity-measure groups: K_i, K_d, and IC_{50}. The three correlation values were then averaged to provide a generalization of the scoring functions' performance across all the proteins in the CSAR data set. The average was computed by taking a weighted mean of the three groups, where the weight given to a group is equal to the number of proteins in that group: two protein targets in the K_d group (considering CDC and CDK as one protein), two in the K_i group, and one protein in the IC_{50} group). We did not weight each affinity group equally, or according to the number of compounds in the group, because either of these methods would give about 30% of the weight to the IC_{50} group, which would give its single protein member, CHK, excessive influence on the averaged results. We present this weighted mean as the affinity performance measure used in the Results section. We also computed the score−affinity correlations separately for each protein group, and these correlations are provided in the Supporting Information.

In order to evaluate whether the scoring functions are able to discriminate active/inactive compounds, we labeled the compounds as active or inactive and analyzed the receiver operating characteristic (ROC) using the ROCR package,[21] available in R (http://cran.r-project.org). ROC curves, giving the ratio of the true positive rate to the false positive rate for various choices of score cutoffs, are given in the Results.

Details on the labeling of all the compounds as active or inactive are provided in the Appendix but are briefly summarized here: Compounds were considered active if their experimentally determined K_d, K_i, or IC_{50} value was less than 10 μM; compounds were considered inactive if their K_d, K_i, or IC_{50} value was known to be greater than 100 μM. A gap from 10−100 μM, within which compounds were considered neither active nor inactive, was used to reduce the risk that a compound might be labeled differently depending on the affinity measure and assay. This resulted in 491 compounds labeled active and 185 compounds labeled inactive. We provided the ROC curve for all protein groups combined, and for CDK2-Cyclin A and CDK2 separately, each of which have 22 actives and 84 inactives. Curves were not provided for the other four protein groups (CDK2, ERK2, LPXC, and Urokinase) because they all had 12 or fewer labeled inactives, and two of them had no inactives.

3. RESULTS

We evaluated the performance of ITScore, STScore, and VDWScore on binding affinity predictions, binding mode predictions, and active/inactive compound discrimination. The binding affinity predictions were evaluated by computing the Pearson correlation between the docking scores and the known binding affinities for three groups, separated by affinity measure: K_d, K_i, or IC_{50}. Binding mode predictions were evaluated based on the heavy-atom root-mean-square deviation

(RMSD) between each docked ligand and the native, bound ligand. Finally, ROC curves were plotted to show the active/inactive compound discrimination of the three scoring functions. These three sets of results are described as follows.

3.1. Binding Affinity Predictions. All binding affinity results are presented in the top panel of Figure 1. The *y*-axis values are the weighted averages of the three Pearson correlations for each affinity group (K_d, K_i, IC_{50}) where the weight given to a group is proportional to its number of protein members, as described in Methods.

We divided these results into three different sets in order to more clearly show three comparative relationships. The first set focuses on the effect of the native versus non-native protein conformation on the affinity prediction accuracy. The second set shows the importance of the native ligand conformation versus the Omega-generated ligand conformations. The third set presents the results for both the non-native protein and Omega-generated ligand ensembles used together, using either the structure subset (i.e., the subset in which the crystal structure is provided for each protein−ligand complex) or the full set with affinities. Each set begins with the native-ligand, native-protein results in order to make the visual comparisons easier.

In the first comparison set (top left of Figure 1, "native protein versus cross ensemble"), the native-ligand, native-protein affinity results are given in one group, followed by the native-ligand, protein cross ensemble results in a second group. ITScore and STScore are found to be sensitive to the non-native protein conformation. A substantial decrease in prediction accuracy is seen when the docking is done on the protein cross ensembles instead of the native protein structures. For ITScore, the average binding affinity correlation falls from 0.81 to 0.65, and for STScore it falls from 0.61 to 0.49. No significant difference is found for the accuracy of the affinity prediction by VDWScore as a result of the protein conformation.

In the second comparison set (top-middle of Figure 1, "native ligand versus Omega-generated ensemble"), the native-ligand, native-protein affinity results are given as the first group, followed by two groups that use Omega-generated ligand ensembles. The last group uses an ensemble of up to 100 Omega-generated ligand conformations ("Ligand: Ensemble"), while the middle group uses the same ensemble plus the native conformation ("Ligand: Native+Ensemble"). STScore performs similarly on all three groups with an average binding affinity correlation of 0.63 for both of the ensemble-ligand cases. ITScore shows a trend toward decreasing performance from the "Ligand: Native" group (R = 0.81) to the Ligand: Native +Ensemble group (R = 0.72) to the Ligand: Ensemble group (R = 0.66). VDWScore performs substantially worse on the Ligand: Native+Ensemble and Ligand: Ensemble groups with average binding affinity correlations of 0.42 and 0.43, respectively. Overall, the Ligand: Native+Ensemble and Ligand: Ensemble groups differ only slightly, suggesting that the decrease in the performance of ITScore or VDWScore from the Ligand: Native group is not purely the result of inadequate sampling of ligand conformations in the Omega-generated ensembles. The binding mode results (presented in the next subsection) between these two groups were also nearly identical.

In the top right of Figure 1 ("native versus ensemble protein/ligand on struct set and full set"), this third comparison set shows the native−native results again, followed by three groups of protein/ligand ensemble results. The first two sets of protein/ligand ensemble results use the structure subset ("Test Set: structures with known affinity"), the same subset used for all the results presented in the first and second comparison groups. The last group uses the full set of compounds with known binding affinities. The first two groups of protein/ligand ensemble results differ only in how the ligands were generated: in the first group (Ligand: Ensemble), the ligand ensembles were generated by Omega from the connection table of the native ligand MOL2 file, while in the second group ("Ensemble from SMILES"), the ligand ensembles were generated by Omega from the SMILES string provided by CSAR. These two sets are the same, except for 11 compounds with stereochemical ambiguities in the SMILES strings; for these compounds, extra conformations were generated by Omega to explore the stereochemical space. The binding affinity results between these two ligand generation methods were close, with ITScore having average binding correlations of 0.65 and 0.66 on the two groups respectively. For STScore, these values were 0.54 and 0.60, and for VDWScore they were 0.47 and 0.54. We conclude that the large difference in performance between the first ensemble−ensemble group on the structure subset and the last ensemble−ensemble group on the full set was due to the difficulty of the full test set versus the structure subset and not due to the method of generating the ensembles of ligand conformations.

In the first two groups of ensemble−ensemble results, a decrease in performance is seen for ITScore relative to the native−native results. This decrease, from R = 0.81 to R = 0.65 is the same decrease seen for the native-ligand, ensemble-protein case (R = 0.65), and the ensemble-ligand, native-protein case (R = 0.66). A similar decrease is seen for VDWScore (from R = 0.61 to R = 0.54) although the decrease is insignificant for the Ensemble from SMILES case (R = 0.60). Overall, the performance of STScore is close between the native−native (R = 0.61) and ensemble−ensemble cases (R = 0.54 for ligand conformations from the connection table, and R = 0.60 for the conformations from the SMILES strings).

In the last group of protein/ligand ensemble results, the full set of compounds (as ensembles generated from the CSAR-provided SMILES strings) were docked to the non-native protein ensembles for each protein group. As with the other groups, the *y*-axis value is the weighted average of the Pearson correlations for the K_d, K_i, and IC_{50} subsets, where the weight given to each group is proportional to the number of proteins within the group (two for K_d, two for K_i, and one for IC_{50}). A decrease in performance is seen for all three scoring functions on the full set of ensemble−ensemble results compared to the structure subset. For ITScore the average binding affinity correlation was 0.31, for STScore it was 0.32, and for VDWScore it was 0.34. For ITScore, which performed better than the other two scoring functions in every other case; this decrease in the average binding affinity correlation was large (from R = 0.65 for the structure subset to R = 0.31 for the full set with affinities). As discussed above, the ligand ensembles generated from the MOL2 connection table (Ligand: Ensemble) and the SMILES ligand ensembles ("Ligand: Ensemble from SMILES") give close results. This consistency in results suggests that the substantial decrease in performance seen with the full CSAR data set is due to greater difficulty in the full set itself. The full CSAR data set includes many compounds with similar activity values, and these activity values are less consistently related to the basic ligand parameters than

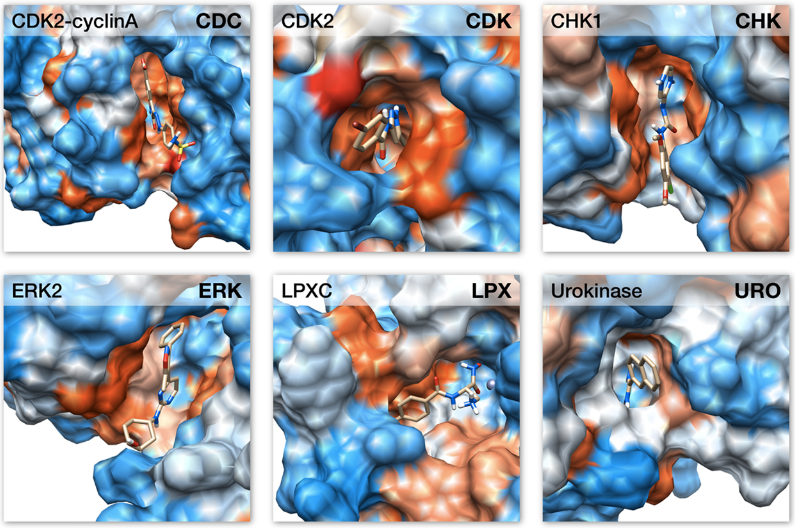


Figure 2. Example binding modes for each of the six protein groups. The protein surface around each residue is colored according to the Kyte and Doolittle scale of hydrophobicity.[22] Hydrophilic residues are colored blue while more hydrophobic residues are colored orange or red. The figure was generated using UCSF Chimera 1.6.2.

is the case for the structure subset. For example, the affinity values in the structure subset are moderately correlated to the number of atoms in each ligand; however, for the full set, this correlation is very weak.

Overall, ITScore made better binding affinity predictions than the other two scoring functions in every case except the ensemble−ensemble evaluation on the full set of compounds with affinities. The binding affinity prediction accuracy of STScore and of VDWScore were similar, with one doing better than the other in about half of the cases. Given the crudely simple functional form of VDWScore, its comparable affinity performance was unexpected. We suspect that its results may have been improved by the binding pocket hydrophobicity of several of the proteins in the 2012 CSAR data set, as depicted in Figure 2. While the surrounding regions can be seen to contain many hydrophilic residues, the actual contact residues are predominately hydrophobic for four of the proteins, and mixed for the other two proteins.

3.2. Binding Mode Predictions.

All binding mode results are presented in the bottom panel of Figure 1. The *y*-axis values are the percent of compounds for which the top ranked binding mode and the native binding mode are within 2.0 Å RMSD (root-mean-square deviation), excluding hydrogen atoms. As with the affinity results, the binding mode results are divided into three sets in order to more clearly show the three comparative relationships: "native protein versus cross ensemble," "native ligand versus Omega-generated ensemble," and "native versus ensemble protein/ligand." Because evaluating the binding mode requires knowledge of the native ligand position, the binding mode evaluations were restricted to the structure subset of compounds ("Test Set: all structures").

In the first comparison set (bottom left of Figure 1: "native protein versus cross ensemble"), the native-ligand, native-protein binding mode results are given in the first group, followed by the native-ligand, ensemble-protein results in the second group. All scoring functions show a decrease in binding mode predictions as a result of using the protein cross ensembles instead of the native protein conformations. For ITScore, the success rate drops from 84% to 68%. For STScore the native−native success rate is 77%, and drops to 68% when the protein ensemble is used instead. VDWScore does well in the native case, with a success rate of 88%, and the success rate drops to 70% with the protein ensemble.

In the second comparison set (bottom middle of Figure 1: "native ligand versus Omega-generated ensemble"), it is shown that the native ligand conformation is very important for successful binding mode predictions when testing the three scoring functions on the 2012 CSAR data set. All three scoring functions show a large decrease in performance when using an Omega-generated ligand ensembles (Ligand: Native+Ensemble or Ligand: Ensemble) rather than the native ligands, and this decrease in performance is substantially larger than the decrease seen when using the protein cross ensembles versus the native protein conformations. For ITScore the success rate drops from 84% to 52% for both Omega-generated ligand ensembles. For STStore the success rate drops from 77% to 39% for the Ligand: Native+Ensemble case and to 41% for the Ligand: Ensemble case. For VDWScore the decrease is the greatest, from 88% to 18% and 20%, suggesting that this scoring function (which does not consider electrostatics, explicitly or implicitly) is especially bad at distinguishing the native ligand conformation.

On the bottom right of Figure 1 ("native versus ensemble protein/ligand") is the third comparison set. It shows the native−native results as a point of comparison followed by the protein/ligand ensemble results. As with the ensemble-ligand, native-protein results, the ensemble−ensemble binding mode predictions are much worse than the native−native binding mode predictions for all three scoring functions. Between the ensemble-ligand, native-protein results and the ensemble-ligand, ensemble-protein results, a trend of decreasing performance is seen for all three scoring functions. The success rates decrease to 36% and 30% for ITScore and STScore, respectively. VDWScore, whose performance was already only 20% in the ensemble-ligand, native-protein case, does particularly poorly in the ensemble−ensemble case, with a success rate of 11%.

To summarize, when the native, bound ligand conformation is used, all three scoring functions give similar binding mode performance. For the native−native case, STScore performs

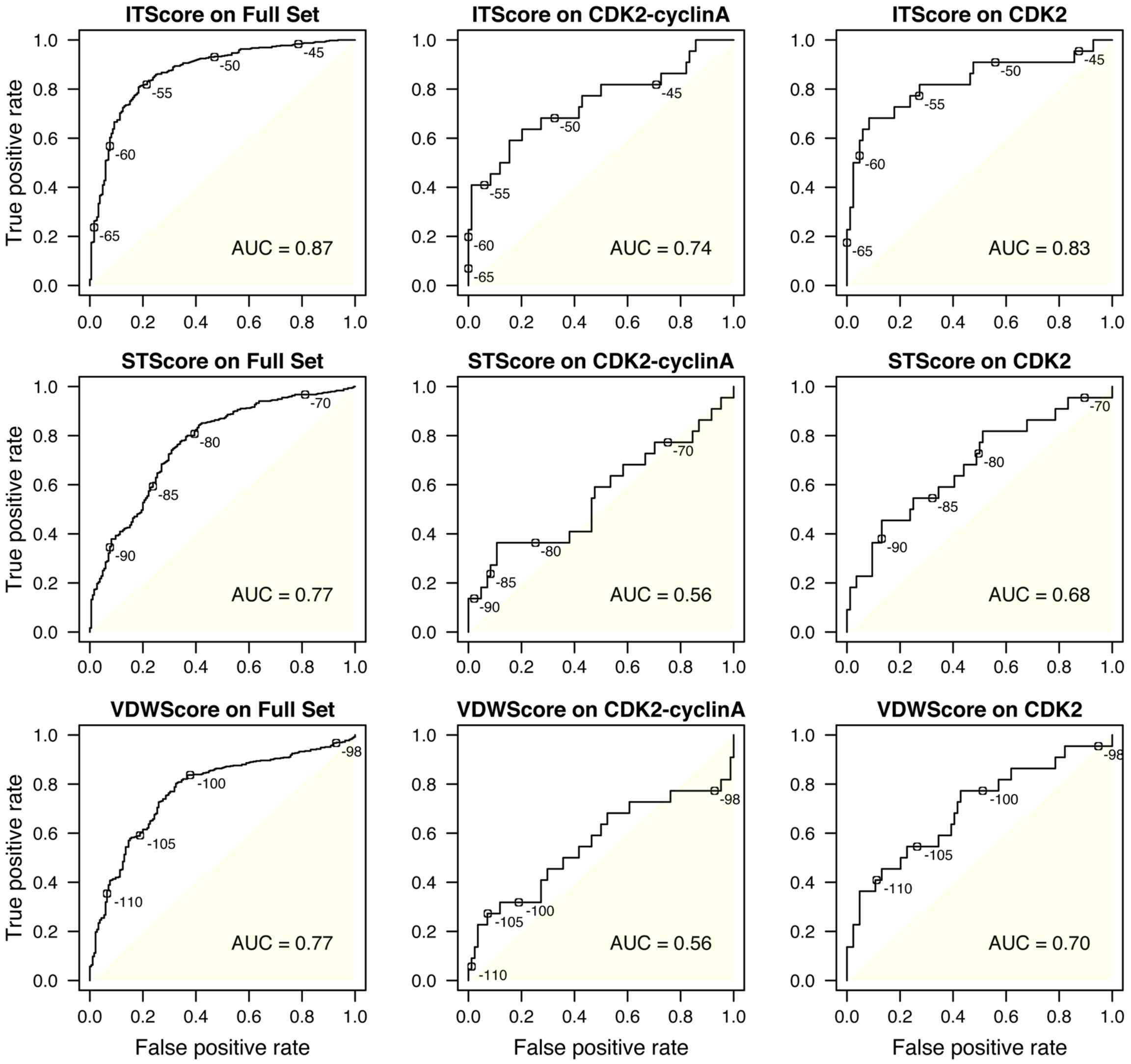


Figure 3. Active/inactive compound discrimination of ITScore (row 1), STScore (row 2), and VDWScore (row 3), presented as receiver operating characteristic (ROC) curves. These curves give the ratio of the true positive rate to the false positive rate for each possible choice of docking score cutoff. The labeled values on each curve give example cutoffs. The first column gives the ROC curves for the full set of active and inactive compounds. This column combines compounds from all six protein groups and tests the ability of each scoring function to discriminate between the actives and inactives of different proteins. The second and third columns give ROC curves for CDK2-Cyclin A and CDK2. Specific ROC curves for other protein groups were not provided because they all had 12 or fewer inactives.

slightly worse than the other two scoring functions. Whenever the Omega-generated ligand ensemble is used, the binding mode performance of all the three scoring functions decreases substantially. For VDWScore this decrease is very large, because the shape complementarity is not sufficient for docking in the absence of native ligand conformations, even with the pockets that are mostly hydrophobic.

3.3. Active/Inactive Compound Discrimination. The performance of the three scoring functions on active/inactive compound discrimination are presented as a set of three ROC curves for each scoring functions in Figure 3. These curves give the ratio of the true positive rate to the false positive rate for each possible choice of docking score cutoff. The labeled values on each curve give example cutoffs. For example, at the top left, the ROC curve titled “ITScore on Full Set” labels the curve at cutoffs of ITScore = −65, −60, −55, −50, and −45.

The first set of ROC curves (left panels) give the ROC curves for ITScore, STScore, and VDWScore on the full set of active and inactive compounds. This combines compounds from all six protein groups, testing the ability of each scoring function to discriminate between the actives and inactives of different proteins. The second and third columns give ROC curves for specific protein-group subsets of the full set. Only CDK2-Cyclin A and CDK2 are given because these are the only protein groups for which more than 20 actives and 20 inactives were available. The other protein groups have 12 or fewer inactives.

ITScore performs well in all three evaluations with an area under the curve (AUC) of 0.87 on the full set, and 0.74 and 0.83 for CDK2-Cyclin A and CDK2, respectively. STScore and VDWScore do not perform as well as ITScore on the full set (AUC = 0.77 for the two). On the individual protein evaluations, STScore and VDWScore perform poorly. For CDK2-Cyclin A, AUC equals 0.56 for both, narrowly better than random selection. For CDK2, AUC was 0.68 for STScore and 0.70 for VDWScore.

3.4. Summary of Results. In summary, ITScore performed relatively well in all binding affinity predictions on the structure subset. It also performed well in active/inactive compound discrimination. VDWScore performed less well than ITScore in the binding affinity predictions on the structure subset, but slightly better than ITScore in non-native protein/ligand ensemble binding affinity predictions on the full set.

In binding mode predictions, ITScore and VDWScore performed similarly when the native ligand conformation was used. When the Omega-generated ligand ensembles were used instead, ITScore performed much better than VDWScore. For all scoring functions, the generated ligand conformation ensembles gave worse binding mode results than the native ligand conformation, and in these cases, VDWScore did particularly poorly.

Overall, in the Omega-generated ligand ensemble cases, the performance of STScore was consistently better than VDWScore in both binding mode and binding affinity predictions. The exception was the full set with known affinities: in this case, the affinity predictions of STScore and ITScore were slightly worse than those of VDWScore.

In the native−native case of the CSAR structure subset, the binding affinities had average Pearson correlations of 0.81, 0.61, and 0.64 with the docking scores given by STScore, ITScore, and VDWScore, respectively. For the ensemble−ensemble case on the structure subset, these values fell to 0.65, 0.54, and 0.47. In the native−native case, the binding mode prediction success rates were 84%, 77%, and 88% for STScore, ITScore, and VDWScore. These values fell to 36%, 30%, and 11% in the ensemble−ensemble case (where the success criterion for binding mode prediction is that the top ranked binding mode and the native binding mode are within 2.0 Å RMSD). We also found the full CSAR data set to be more challenging in making binding mode predictions than the subset with structures. For the full set of compounds with known affinity, the binding affinities had average Pearson correlations of 0.31, 0.32, and 0.34 with the docking scores given by STScore, ITScore, and VDWScore, respectively. For the active/inactive compound discrimination all the scoring functions performed better. In evaluating the ROC on the full set, the area under the curve was 0.87 for ITScore, and 0.77 for both STScore and VDWScore.

4. DISCUSSION AND CONCLUSION

Our data supported some of our previous conclusions about ITScore and STScore and, in other cases, contradicted our expectations. Our results are consistent with our previous conclusion[23,24] that ITScore's iterative method of dealing with the reference state problem is able to substantially increase binding mode and binding affinity predictions compared to its initial potential (which is similar to STScore for abundant pair types and similar to VDWScore for close atom pair distances). All three scoring functions were tested using MDock, which positions an ensemble of pregenerated/rigid ligand conformations without further conformational sampling.

STScore is similar to the initial potential of ITScore for abundant pair types (which account for the majority of interactions) but is also similar to VDWScore for rare atom pair types. Consequently, it fell within our expectations that the performance of STScore was often in-between that of ITScore and VDWScore. Evidently, the knowledge-based component of STScore improves its binding mode predictions compared to its force-field based component alone, which is a slightly modified version of VDWScore. Its binding mode predictions were still not as good as ITScore, although our unpublished results suggest that ITScore-like iterations can further improve the binding mode predictions of STScore.

Considering its simple functional form, it was initially surprising that the binding affinity predictions of VDWScore exceeded STScore in some cases and ITScore in one case. It is possible that its performance was enhanced by the hydrophobicity of the protein binding pockets in the 2012 CSAR data set. This would make the task easier for VDWScore, because it must rely primarily on the shape complementarity of the protein and the ligand. In support of this view (that the performance of VDWScore is highly dependent on shape complimentarity), there was a significantly larger difference in the binding mode performance of VDWScore depending on whether the native or Omega-generated ligand conformation ensemble was used. This difference was much greater for VDWScore than for ITScore or STScore. When docking the native ligand conformation, VDWScore gave quite accurate binding mode predictions, but in those cases that use the Omega-generated ligand ensembles, its binding mode predictions were very poor. So in summary, these results suggest that the knowledge-based aspect of ITScore and STScore is able to increase their binding mode predictions beyond that of VDWScore by implicitly including other types of interactions. The contribution of other interactions is especially important when the native conformation of the ligand is unknown. Nevertheless, STScore and ITScore leave much room for improvement in binding affinity and binding mode predictions when using the non-native protein/ligand conformations.

In general, our scoring experiments with the full set of 2012 CSAR benchmark show that it is most interesting yet challenging to predict binding modes and affinities without the knowledge of native protein and ligand conformations. The van der Waals scoring function may be used as a reference for scoring comparison; van der Waals performs much better on predictions for native protein and ligand conformations than for non-native conformations. The use of the pregenerated ligand conformations seems to lower the success rates significantly more than the use of the non-native protein conformations for binding mode predictions. The corresponding difference is less for binding affinity predictions. This phenomenon may be due to the fact that the main conformational changes of the proteins are side chain flexibility in 2012 CSAR data set. Future work may include adapting these scoring functions and docking methods for use with on-the-fly ligand conformational sampling and side chain rotamer sampling.

6. APPENDIX

We have prepared a CSV data file and several script files which we believe will be of use to future users of the CSAR data set. The CSV datafile, which is available in the Supporting Information, specifies several labels for each of the 757 compounds in the CSAR data set. By providing the information for each compound in one file, it becomes simple to apply commands to relevant subsets of the CSAR data set. For copyright reasons, the affinity columns from the CSV file have been excluded. The full datafile, including all affinity data, is available on the CSAR Web site (http://www.csardock.org). The set of scripts we used to set up the CSAR data set for docking (i.e., the steps in the Methods before docking) are also available at the same Web site. One of these scripts can be used to generate a set of ligand conformation files with canonicalized atom orders and may be adapted for other docking applications. The CSV file is described in detail as follows.

We assigned each of the 757 compounds in the CSAR data set a consistent filename of the form `code_id` where code specifies the protein and id specifies the compound ID number. There are six protein groups: CDK2-Cyclin A, CDK2, CHK1, ERK2, LPXC, and Urokinase. The corresponding three-letter codes are CDC, CDK, CHK, ERK, LPX, and URO, respectively. For each compound, we built a CSV file which combines all the basic information available for each compound (other than its structural coordinates) into one file. The data series provided in this CSV file are summarized in Table 1. The first column gives the compound ID, and the second column gives its corresponding protein group. The compound IDs are unique only within each protein group. The third column, the base filename for each compound, combines the compound ID number and protein group name. These base filenames are unique.

Table 1. CSV Datafile Specification[a]

column	label	value
1	compound ID	integer
2	protein code	string (CDC, CDK, CHK, ERK, LPX, or URO)
3	base filename	string (e.g., CDK_4)
4	excluded?	boolean (0 or 1)
5	designated active?	boolean (0 or 1)
6	designated inactive?	boolean (0 or 1)
7	structure available?	boolean (0 or 1)
8	affinity available?	boolean (0 or 1)
9	in March structure subset?	boolean (0 or 1)
10	affinity measure	string (K_d, K_i, or IC_{50})
11	affinity assay	string (OctetRed, Abbott, Vertex, or Thermofluor)
12	affinity in μM	float
13	affinity in M	float
14	−log(affinity in *M*)	float
15	compound SMILES	string

[a] This table defines the 14 columns listed in our CSV datafile for the 2012 CSAR data set. Each column represents a compound label; they are explained in detail in the Appendix.

The next six columns in the CSV file provide binary labels, most of which describe what information is available for each compound. Logical combinations of these may be used to apply commands to relevant subsets. The first of the binary labels specifies whether or not a compound is excluded from the final results calculations. For our calculations, we only excluded one compound, CDK_5, because we had a doubt about its affinity data. In the CSAR data set, the affinity data spreadsheet for CDK2 includes a table of designated inactive compounds. For all of these inactives, the affinity is given as >100 μM. The other table of compounds includes those with known activity from 0.023 to 58.3 μM and two additional compounds CDK_5 and CDK_12. The affinity of CDK_12 is unknown, but the affinity of CDK_5 is given as K_d > 100 μM, the same value as given for the compounds in the list of designated inactives. Nevertheless, the K_d value given for this compound was not trustable, because the compound was reported to be insoluable. Therefore, we were suspicious of CDK_5 and excluded it from our results calculations. CDC_5 is the same compound and was also excluded.

The next two binary labels (columns 5 and 6) specify whether a compound is designated to be active or designated to be inactive. We considered a compound to be active only if its affinity was known to be less than 10 μM (whether K_d, K_i, or IC_{50}). Due to the differences between these affinity measures and assays, we left a gap from 10 to 100 μM between which a compound is considered to be neither nor inactive. Compounds with K_d, K_i, or IC_{50} > 100 μM were labeled as inactive.

Column 7 specifies whether or not a crystal structure is available in the CSAR data set containing the compound bound to its associated protein. This column defines the structure subset referred to in the Methods. Likewise, column 8 specifies if the precise affinity of the compound for its associated protein is available in the CSAR data set. We labeled this field as 0 (false) when the affinity is given as a comparison (e.g., >100 μM). Finally, the last binary label (column 9) specifies if the protein−ligand structure for a compound was one of the structures in the results of the 2011−2012 CSAR Benchmark Exercise, provided in March. The number of compounds in the CSAR Database satisfying each condition is given in Table 2.

Column 10 specifies which affinity measure was used (K_d, K_i, or IC_{50}) and Column 11 states which assay was used to produce the data (OctetRed, Abbott, Vertex, or Thermofluor). The next three columns are designated for the affinity data itself. Column 12 gives the affinity of the compound for its associated protein in units of micromolar, column 13 gives the affinity in units of molar, and column 14 gives the negative logarithm of the affinity. For three compounds, the original CSAR data set includes duplicate entries with slightly different affinities. These duplicate entries represent different salt forms of the same compound. For these compounds, we provide the mean of the two values. Lastly we included all the SMILES strings in the CSAR data set in column 15. The version of this file provided in the Supporting Information excludes the affinity data and SMILES strings. A version providing all data can be found on CSAR's Web site (http://www.csardock.org/).

Table 2. Size of CSAR Subsets[a]

protein name	protein code	total	doubted	active	inactive	structures	affinities	March
CDK2-Cyclin A	CDC	111	1	22	84	1	23	0
CDK2	CDK	111	1	22	84	15	25	0
CHK1	CHK	159	0	106	0	17	107	14
ERK2	ERK	298	0	293	0	12	298	12
LPXC	LPX	32	0	13	12	5	20	4
Urokinase	URO	46	0	35	5	7	35	4
	TOTALS	757	2	491	185	57	508	34

[a] This table gives the number of compounds in each protein group that satisfy each of the binary labels. These labels are listed in the column headers.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Additional figures and a CSV datafile (see Appendix). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: zoux@missouri.edu. Tel.: 573-882-6045. Fax: 573-884-4232.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Support to X.Z. from OpenEye Scientific Software Inc. (Santa Fe, NM) is gratefully acknowledged. This work is supported by NIH Grants R21GM088517 (X.Z.) and NSF CAREER Award DBI-0953839 (X.Z.). S.Z.G. is supported through the NLM Biomedical Informatics Research Training Program (T15 LM07089, PI: Simoes). Computing support by the HPC resources at the University of Missouri Bioinformatics Consortium (UMBC) is also gratefully acknowledged.

■ REFERENCES

(1) Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335−373.

(2) Huang, N.; Jacobson, M. P. Physics-based methods for studying protein-ligand interactions. *Curr. Opin. Drug Discovery Dev.* **2007**, *30*, 325−331.

(3) 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.

(4) Huang, S.-Y.; Zou, X. Advances and challenges in protein-ligand docking. *Phys. Chem. Chem. Phys.* **2010**, *11*, 3016−3034.

(5) Dunbar, J. B.; 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.

(6) Smith, R. D.; Dunbar, J. B.; 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.

(7) Huang, S.-Y.; Zou, X. An iterative knowledge-based scoring function to predict protein−ligand interactions: I. Derivation of interaction potentials. *J. Comput. Chem.* **2006**, *27*, 1866−1875.

(8) Huang, S.-Y.; Zou, X. An iterative knowledge-based scoring function to predict protein−ligand interactions: II. Validation of the scoring function. *J. Comput. Chem.* **2006**, *27*, 1876−1882.

(9) Thomas, P. D.; Dill, K. A. Statistical Potentials Extracted From Protein Structures: How Accurate Are They? *J. Mol. Biol.* **1996**, *257*, 457−469.

(10) Huang, S.-Y.; Zou, X. Mean-force scoring functions for protein-ligand binding. *Annu. Rep. Comput. Chem.* **2010**, *6*, 281−296.

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

(12) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. BLEEP − A Potential of mean force describing protein-ligand interactions: I. Generating potential. *J. Comput. Chem.* **1999**, *20*, 1165−1176.

(13) Zhang, C.; Liu, S.; Zhu, Q.; Zhou, Y. A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *J. Med. Chem.* **2005**, *48*, 2325−2335.

(14) Velec, H. F.; Gohlke, H.; Klebe, G. DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.* **2005**, *48*, 6296−6303.

(15) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated docking with grid-based energy approach to macromolecule-ligand interactions. *J. Comput. Chem.* **1992**, *13*, 505−524.

(16) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.* **2001**, *15*, 411−428.

(17) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera−a visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.

(18) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(19) Huang, S.-Y.; Zou, X. Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 399−421.

(20) Huang, S.-Y.; Zou, X. Efficient molecular docking of NMR structures: application to HIV-1 protease. *Protein Sci.* **2007**, *16*, 43−51.

(21) Sing, T.; Sander, O.; Beerenwinkel, N.; Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **2005**, *21*, 3940−3941.

(22) Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105−132.

(23) Huang, S.-Y.; Zou, X. Scoring and lessons learned with the CSAR benchmark using an improved iterative knowledge-based scoring function. *J. Chem. Inf. Model.* **2011**, *51*, 2097−2106.

(24) Huang, S.-Y.; Zou, X. Construction and test of ligand decoy sets using MDock: Community Structure-Activity Resource benchmarks for binding mode prediction. *J. Chem. Inf. Model.* **2011**, *51*, 2107−2114.