JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results

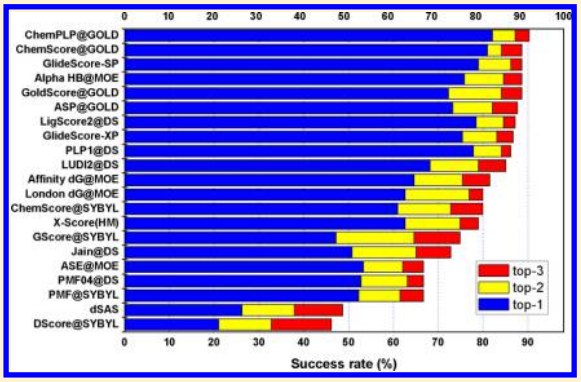
Yan Li,[†] Li Han,[†] Zhihai Liu,[†] and Renxiao Wang*[,†,‡]

[†]State Key Laboratory of Bioorganic and Natural Products Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, 345 Lingling Road, Shanghai 200032, People's Republic of China

[‡]State Key Laboratory of Quality Research in Chinese Medicine, Macau Institute for Applied Research in Medicine and Health, Macau University of Science and Technology, Macau, People's Republic of China

Supporting Information

ABSTRACT: Our comparative assessment of scoring functions (CASF) benchmark is created to provide an objective evaluation of current scoring functions. The key idea of CASF is to compare the general performance of scoring functions on a diverse set of protein−ligand complexes. In order to avoid testing scoring functions in the context of molecular docking, the scoring process is separated from the docking (or sampling) process by using ensembles of ligand binding poses that are generated in prior. Here, we describe the technical methods and evaluation results of the latest CASF-2013 study. The PDBbind core set (version 2013) was employed as the primary test set in this study, which consists of 195 protein−ligand complexes with high-quality three-dimensional structures and reliable binding constants. A panel of 20 scoring functions, most of which are implemented in main-stream commercial software, were evaluated in terms of "scoring power" (binding affinity prediction), "ranking power" (relative ranking prediction), "docking power" (binding pose prediction), and "screening power" (discrimination of true binders from random molecules). Our results reveal that the performance of these scoring functions is generally more promising in the docking/screening power tests than in the scoring/ranking power tests. Top-ranked scoring functions in the scoring power test, such as X-Score[HM], ChemScore@SYBYL, ChemPLP@GOLD, and PLP@DS, are also top-ranked in the ranking power test. Top-ranked scoring functions in the docking power test, such as ChemPLP@GOLD, Chemscore@GOLD, GlidScore-SP, LigScore@DS, and PLP@DS, are also top-ranked in the screening power test. Our results obtained on the entire test set and its subsets suggest that the real challenge in protein−ligand binding affinity prediction lies in polar interactions and associated desolvation effect. Nonadditive features observed among high-affinity protein−ligand complexes also need attention.



1. INTRODUCTION

Molecular docking is probably the most applied technique in structure-based drug design. Many molecular docking programs have been developed since the 1980s. Examples of today's popular molecular docking programs include DOCK,[1,2] AutoDock,[3] GOLD,[4] Glide,[5,6] and Surflex-Dock.[7,8] In order to predict the preferred binding pose of a ligand molecule to a given molecular target, computational docking methods sample possible ligand binding poses and often rely on scoring functions to rank them. Thus, many scoring functions are implemented in molecular docking programs, such as GoldScore[4] and GlideScore.[5,6] There are also some standalone scoring functions, such as X-Score[9] and DrugScore.[10,11]

Docking/scoring methods are applied in structure-based drug design for various purposes.[12−20] For example, large compound libraries can be virtually screened by molecular docking programs to identify the candidates that fit into the binding site on a given target. Then, only those selected candidates need to be examined in experiments to save cost and labor. Once some true binders to the given target are verified, molecular docking can be applied to predict their binding modes. Knowledge of ligand binding modes in turn can be used by scoring functions to predict ligand binding affinities to the target. Experimental means, such as X-ray crystal diffraction and NMR spectroscopy, are able to derive real binding modes. But they all have certain technical bottlenecks so that they often fail to provide timely feedback. Thus, hypotheses produced by molecular docking serve as valuable guidance for rational optimization of lead compounds. More reliable and more efficient docking/scoring methods are still needed in structure-based drug design as well as other molecular recognition studies.

Due to the wide popularity of docking/scoring methods, many comparative assessments of docking/scoring methods have been conducted in the past.[21−33] These studies typically

Received: February 10, 2014
Published: April 7, 2014

evaluated the performance of docking/scoring methods in ligand binding pose/binding affinity prediction and virtual screening. One of the earliest studies of this type was conducted by Bissantz et al., who evaluated three docking programs (DOCK, FlexX, and GOLD) in combination with seven scoring functions on 10 thymidine kinase complexes and 10 estrogen receptor complexes.[21] Bursulaya et al. evaluated five docking programs (DOCK, FlexX, AutoDock, GOLD, and ICM) on 37 complexes formed by 11 different proteins.[22] Later, the data sets used in evaluations were expanded rapidly. To cite a few recently published examples, Chen et al. evaluated four docking programs (FlexX, GOLD, GLIDE, and ICM) using a set of 164 protein−ligand complexes and virtual screening trials on 12 drug targets.[23] Warren et al. evaluated 10 docking programs (Dock, DockIt, FlexX, Flo, FRED, GLIDE, GOLD, LigandFit, MOE, and MVP) in combination with 37 scoring functions on 136 complexes formed by eight pharmaceutical target proteins.[24] Cross et al. evaluated six docking programs (DOCK, FlexX, GLIDE, ICM, PhDOCK, and Surflex) with respect to docking accuracy on 68 selected protein−ligand complexes and virtual screening enrichment on 40 protein targets.[25] A brief summary of the comparative studies of docking/scoring methods to our knowledge is given in the Supporting Information (part I).

Outcomes of such comparative studies provide valuable guidance for the users to make a reasonable choice among available methods. They also help the developers get a better understanding of the strengths and weaknesses of current docking/scoring methods. To serve either purpose, the study must be conducted objectively on a solid basis. However, previous comparative studies of docking/scoring methods often have two problems. One lies in the data sets employed in those studies. Those data sets are often assembled in a random manner rather than with the outcomes of a systematic approach. Besides, some low-quality samples are not filtered out. Evaluation results obtained on these types of data sets could be somewhat misleading. This problem has been addressed in more detail in the companion work of this series[34] and thus will not be repeated here.

Another problem lies in the evaluation methods employed in those studies, in which scoring methods were often tested in the context of molecular docking or even virtual screening trials. We refer to this approach as the "black-box approach" in this work. Molecular docking is a sophisticated procedure combining binding pose sampling and scoring. In particular, sampling of binding poses is controlled by multiple parameters. Different settings of adjustable parameters and optional methods for binding pose sampling also have impact on the final outcomes of a molecular docking job. A virtual screening trial is even more sophisticated because, in addition to the complications in molecular docking, a library of molecules as well as a set of rules for selecting the top candidates have to be considered. If one only cares for the final outcomes of a molecular docking program, the black-box approach is well-acceptable. But if one wants to evaluate the performance of a scoring function in this process, this approach is not effective.

To tackle this problem, we proposed in the early 2000s that "scoring" needed to be separated from "docking" to obtain a more objective evaluation of scoring function per se. In our first published study following this approach,[35] a total of 11 scoring functions was used on 100 selected protein−ligand complexes. Each scoring function was tested directly on the crystal structures of those protein−ligand complexes to examine if they could produce binding scores correlated with experimental binding data. Besides, a set of decoy ligand binding poses was generated for each complex by using a separate molecular docking program. Then, each scoring function was evaluated by its success rate of identifying the native binding pose among computer-generated decoys. Our method seems to be well-accepted by the community as indicated by a large number of citations in the literature. It should be mentioned that Mitchell et al.[36] and Brooks et al.[37] conducted comparative studies of scoring functions with similar methods at approximately the same time.

Our efforts on establishing better benchmarks for scoring function evaluation later evolved into what we call the comparative assessment of scoring functions (CASF) project. This project has been greatly aided by the creation of the PDBbind database.[38,39] The first publicly revealed study, i.e., CASF-2007, employed several data sets selected from the PDBbind database (version 2007) to test a total of 16 popular scoring functions.[40] The primary test set consisted of 195 diverse protein−ligand complexes with high-resolution crystal structures and reliable binding constants, which were selected through a systematic nonredundant sampling of the PDBbind database. In addition to the use of a new test set, evaluation methods were also refined in CASF-2007. All scoring functions were evaluated in three aspects, i.e., "docking power", "ranking power", and "scoring power". Our study demonstrated that only a few scoring functions achieved modest correlation between computed binding scores and experimental binding data, but a number of scoring functions performed much better in detecting the correct ligand binding poses with success rates over 70%. Another notable observation was that no scoring function consistently outperformed others in all three aspects. Thus, it is wise to choose appropriate scoring functions for different purposes in practice.

We recently completed another update study, i.e., CASF-2013. Compared to CASF-2007, major improvements have been made in all three major aspects of our benchmark. First, we have reformed the methods for compiling the primary test set. An updated test set, i.e., the PDBbind core set (version 2013), was compiled based on the latest release of PDBbind. Second, a few more scoring functions in popular commercial software have been added to our test. A näive scoring function was also introduced as a reference, which produced somewhat unexpected results. Third, evaluation methods have been expanded to cover four aspects, i.e., "scoring power", "ranking power", "docking power", and "screening power". Evaluations were conducted on the entire test set as well as some subsets of protein−ligand complexes. In the companion work of this series,[34] we have described how the primary test set used in CASF-2013 was compiled. In this work, we will describe the evaluation methods used in CASF-2013 and report the evaluation results. We will also make comparison to other well-known benchmarks in this field, such as the DUD/DUD-E benchmark[41,42] and the CSAR exercise.[43−46] All of the data sets employed in CASF-2013 will be released to the public so that other researchers can utilize them in their own studies.

2. METHODS

2.1. Primary Test Set. The basis of our CASF-2103 benchmark is a set of 195 protein−ligand complexes, i.e., the PDBbind core set (version 2013). These complexes were selected through a systematic sampling from over 8,300 complexes formed between protein molecules and small-

molecule ligands recorded in the PDBbind database. Binding constants of these complexes span nearly 10 orders of magnitude (log K_a = 2.07−11.52). These complexes are grouped into 65 clusters by protein sequences. Each cluster consists of three complexes, which are referred to as "the best", "the median", and "the poorest" by their binding affinities. Binding affinity of the best complex is required to be at least 100 times higher than that of the poorest. The technical methods employed in this process are described in the companion work of this series.[34] A list of the protein−ligand complexes included in this data set together with several basic features are given in the Supporting Information (part II) of this work.

2.2. Scoring Functions under Assessment. In this study, we tested a panel of 20 scoring functions on our new benchmark (Table 1). Among them, 18 scoring functions are implemented in several mainstream commercial molecular modeling software. X-Score, a scoring function developed by ourselves, was also considered. Besides, we introduced a näive scoring function as a reference in our benchmark. This scoring function uses a single descriptor, i.e., the buried solvent-accessible surface area of the ligand molecule upon binding (ΔSAS), which is an estimation of the size of the protein−ligand binding interface. We have shown in the companion work[34] that the correlation between ΔSAS and protein−ligand binding constants is obviously better than other descriptors of molecular size, such as molecular weight or the number of non-hydrogen atoms. Brief descriptions of all 20 scoring functions are given in the Supporting Information (part III). Other scoring functions reported in the literature were not considered in this study, which will be explained later in this work.

Table 1. Summary of the Scoring Functions Evaluated in This Study

scoring function	software	classification	ref
LigScore2	Discovery Studio (version 3.5)	empirical function	47
PLP1/PLP2		empirical function	48, 49
PMF		statistical potential	50−53
Jain		empirical function	54
LUDI1/LUDI2/LUDI3		empirical function	55, 56
GoldScore	GOLD (version 5.1)	energy-based function	4
ChemScore		empirical function	57, 58
ChemPLP		empirical function	59
ASP		statistical potential	60
G-Score[a]	SYBYL (version 8.1)	energy-based function	4
PMF[b]		statistical potential	50−53
D-Score		energy-based function	2
ChemScore		empirical function	57, 58
GlideScore-SP/XP	Schrödinger (version 2011)	empirical function	5, 6, 61
London-dG	MOE (version 2011)	empirical function	MOE user manual
ASE		empirical function	
Affinity		empirical function	
Alpha-HB		empirical function	
X-Score	Academic software (version 1.3)	empirical function	9
ΔSAS	in-house software	empirical function	

[a]:SYBYL's implementation of GoldScore. [b]:SYBYL's implementation of PMF.

Two things regarding X-Score need to be clarified here. First, the current release of X-Score is version 1.3, which was calibrated with an early version of the PDBbind database. What was tested in this study is a special modification of X-Score to avoid possible overlaps between its training set and the primary test set employed in this study. This modification was recalibrated on the remaining 2,764 protein−ligand complexes in the PDBbind refined set (version 2013) after removing the 195 protein−ligand complexes in the PDBbind core set. In fact, this recalibration produced rather trivial changes in the coefficients before each energy term in X-Score (see the Supporting Information, Table S1 in part III). The other 18 scoring functions in commercial software were tested as is even if their original training sets have overlaps with our test set. Second, there are three optional modes in X-Score (version 1.3), i.e., X-Score^{HM}, X-Score^{HP}, and X-Score^{HS}, which differ from each other only in the hydrophobic effect term. The performance of X-Score^{HM} was slightly better than the other two modes in most tests in CASF-2013. Therefore, X-Score^{HM} is chosen to represent X-Score in this work if not specified otherwise.

2.3. Preparation of Structures. *2.3.1. Processing Protein−Ligand Complex Structures.* Coordinates of the 195 protein−ligand complexes in the primary test set were all downloaded from RCSB Protein Data Bank (http://www.rcsb.org/pdb/). The structural files from PDB were then processed into standard formats so that they could be utilized by the molecular modeling software considered in our study. The technical methods for this task have been described in the companion work of this series.[34] For each complex, the processed protein structure was saved in a PDB-format file, and the processed ligand structure was saved in a Mol2-format and a SD-format file. In the above process, no structural optimization was conducted on either the protein or the ligand to retain their original coordinates from PDB.

Another set of complex structures was prepared by optimizing the ligand binding pose in each complex structure. This treatment was necessary for resolving the steric clashes between the ligand molecule and the protein molecule. This task was performed with the "in situ minimization" module in the Discovery Studio software. The minimization was performed by using the "smart minimization algorithm" with the CHARMm force field.[62] The number of minimization cycles was set to 500. While the ligand structure was allowed to relax, the protein structure was kept fixed during minimization. This set of locally optimized complex structures as well as the original crystal structures were used in the scoring power and ranking power tests in our benchmark.

2.3.2. Generation of Decoy Ligand Binding Poses. Decoy ligand binding poses were needed in our benchmark to evaluate the docking power and screening power of each scoring function. We used three popular molecular docking programs, including GOLD (version 5.1, Cambridge Crystallographic

dx.doi.org/10.1021/ci500081m | J. Chem. Inf. Model. 2014, 54, 1717−1736

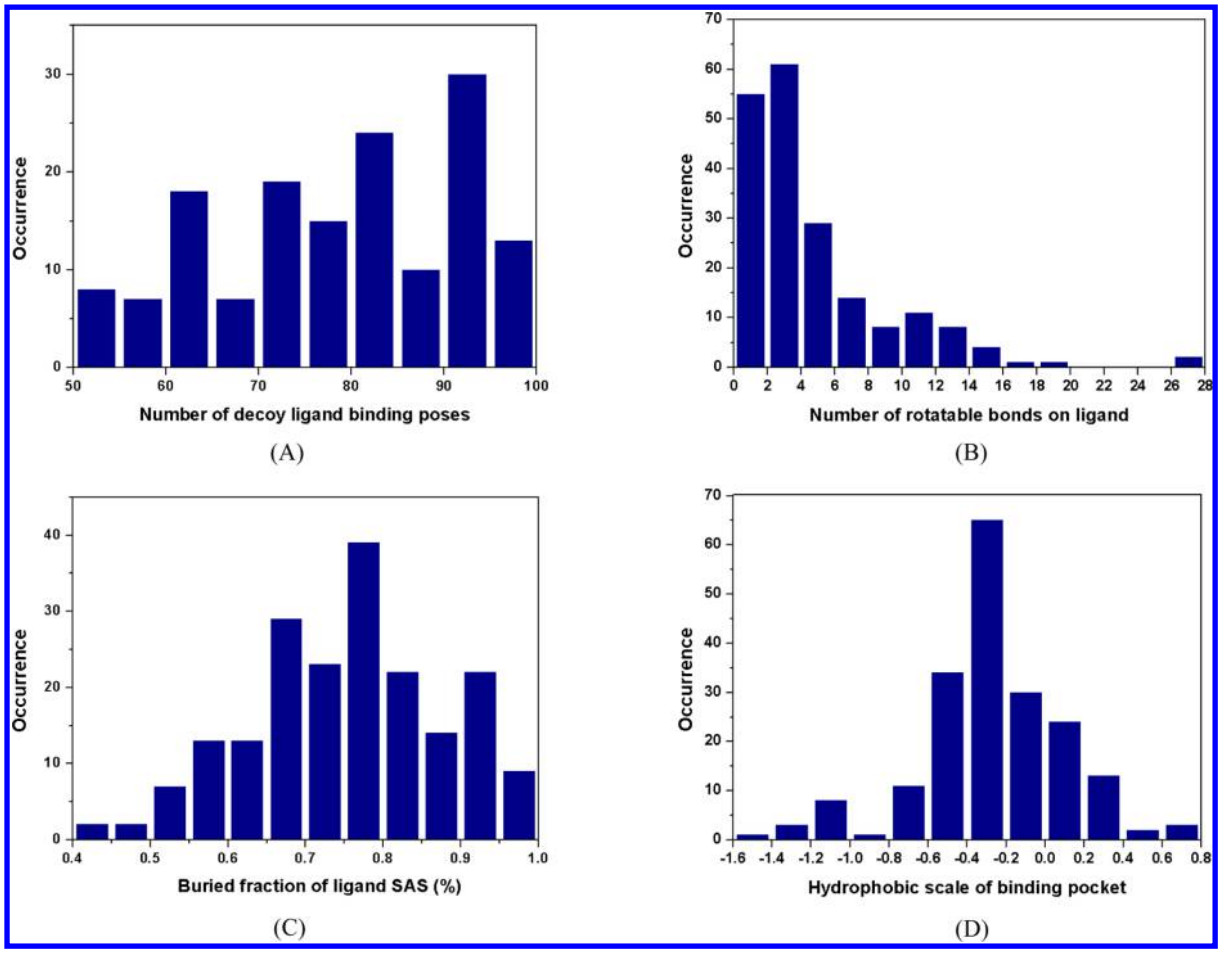


Figure 1. Distributions of several key properties of the protein−ligand complexes in the primary test set: (A) number of decoy ligand binding poses; (B) number of rotatable bonds on the ligand molecule; (D) fraction of ligand solvent-accessible surface buried upon binding; (C) hydrophobic scale of the binding pocket on the protein molecule.

Data Center), Surflex implemented in the SYBYL software (version 8.1, CERTARA Inc.), and the molecular docking module implemented in the MOE software (version 2011, Chemical Computing Group). These three docking programs use different algorithms for sampling ligand binding poses. Their outcomes were combined to obtain a set of ligand binding poses that was as complete as possible. Key parameters and settings used in docking are described in subsequent sections. It needs to be emphasized that, unlike regular molecular docking jobs, these parameters were chosen to generate diverse rather than converged binding poses.

GOLD. GOLD adopts a genetic algorithm for sampling ligand binding poses. In our study, the binding pocket on protein was defined as the residues on the protein within 10 Å from the bound ligand. Structure of the protein molecule was kept rigid during docking. The "automatic" parameter set was adopted, and the program was set to generate up to 100 binding poses with 10% "searching efficiency" in each docking job. No early termination was enabled. Four parallel docking jobs were conducted for each ligand by using one of the four scoring functions implemented in GOLD, i.e., GoldScore, ChemScore, ASP, and ChemPLP (Table 1). Therefore, a total of 4 × 100 = 400 binding poses were generated for each given ligand molecule.

Surflex. Surflex adopts a similarity-based searching algorithm for sampling ligand binding poses. In our study, three separate docking jobs were conducted for each ligand molecule by setting the "additional starting conformations per molecule" parameter to 2, 5, and 10, respectively. The "bloat" parameter for binding site definition was set to the default value of zero. Other parameters were set as follows: "angstroms to expand search grid", 6; "max conformations per fragment", 20, and "max number of rotatable bonds per molecule", 100. The maximal binding poses produced by each docking job was set to 100. Therefore, up to 3 × 100 = 300 binding poses were generated for each ligand molecule.

MOE. The molecular docking module in MOE provides several sampling algorithms and scoring functions, and thus there can be multiple possible combinations of them. In our study, we only employed the default sampling method, i.e., the "triangle matcher" algorithm. The binding site was defined by using the entire ligand molecules as reference. Three parallel docking jobs were conducted for each complex by employing one of the three available scoring functions, i.e., London-dG, ASE, and Alpha-HB (Table 1), respectively. In each docking job, up to 100 binding poses were kept. Therefore, up to 3 × 100 = 300 binding poses were generated for each ligand molecule.

Then, the outputs of all three docking programs were combined to give an ensemble of ligand binding poses for each complex. Only the binding poses with root mean square deviation (RMSD) values lower than 10 Å (with respect to the

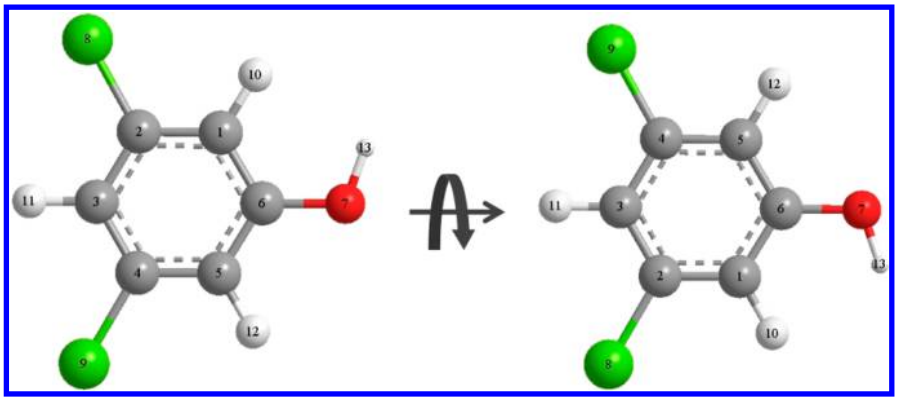


Figure 2. Example illustrating how the standard RMSD algorithm fails to handle symmetric structures. Rotation of this molecule along the x axis by 180° switches the positions of atom nos. 1, 2, 4, 5, 8, and 9 but only reproduces its coordinates. The standard algorithm yields a RMSD value of 3.0 Å between these two poses, while our property-matched RMSD algorithm yields the correct result of zero.

native binding pose) were considered at subsequent steps. All of these binding poses were grouped by their RMSD values (0−10 Å) into 10 bins with an interval of 1 Å. The binding poses in each bin were further grouped into up to 10 clusters according to their internal similarities. This task was performed by using the "rms_analysis" tool in the GOLD software. In each cluster, the binding pose with the lowest internal strain energy was selected as the representative of that cluster. The strain energy of each binding pose was calculated using the MMFF94 force field in the SYBYL software.

Through the preceding process, ideally a total of 100 representative decoy ligand binding poses would be obtained for each complex. However, the number of final selected decoy binding poses was actually lower than 100 for many protein−ligand complexes because of the geometrical constraints of the binding site or the particular shape of the ligand molecule. For example, a deep, narrow binding site or a small, rigid ligand molecule does not allow too many possible binding poses. The binding poses generated for such a complex do not necessarily fill up all of the 10 RMSD bins, or the binding pose in each bin is not always grouped into precisely 10 clusters. In fact, the total number of decoy ligand binding poses ranged from 50 to 100 case by case in our test set, with an average number around 83 (Figure 1A). These decoy sets were employed in the docking power test in our benchmark.

2.4. Evaluation Methods. All 20 scoring functions were tested in four aspects, namely, scoring power, ranking power, docking power, and screening power. The basic concepts and evaluation methods are explained below.

2.4.1. Scoring Power. It refers to the ability of a scoring function to produce binding scores in a linear correlation with experimental binding data. In our benchmark, this feature was evaluated directly on the known three-dimensional structures of the 195 protein−ligand complexes in the test set, including the original crystal structures and the locally optimized structures. Scoring power of a scoring function was quantitatively evaluated by the classic Pearson's correlation coefficient (R) between its binding scores and experimental binding constants and the standard deviation (SD) in regression:

$$R = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

$$\mathrm{SD} = \sqrt{\frac{\sum [y_i - (a + bx_i)]^2}{N - 1}}$$

Here, x_i is the binding score computed by a certain scoring function on the ith complex; y_i is the experimental binding constant of this complex; a and b are the intercept and the slope of the regression line between experimental data and computed data, respectively. Binding constants were all given in logarithm units ($\log K_a$). A special note is that not all 20 scoring functions could compute the entire test set successfully in this test. Some scoring functions actually produced negative (unfavorable) binding scores on certain complexes. All such cases were removed from correlation analysis for those scoring functions. Also, Pearson's correlation is employed to measure a linear relationship. There are other statistical indicators, such as Spearman's ρ or Kendall's τ, that measure the relationship between relative ranks of data. According to our experience,[45] Spearman's ρ or Kendall's τ produce basically the same trend as Pearson's R in scoring function evaluation. Thus, we did not compute those indicators in this study because they do not provide additional value.

2.4.2. Ranking Power. It refers to the ability of a scoring function to correctly rank the known ligands of the same target protein by their binding affinities when the precise binding poses of these ligands are given. Our test set consists of 65 clusters of complexes, each of which has three complexes formed by the same protein with significantly different binding affinities. This special construction is suitable for performing the ranking power test. If a scoring function correctly ranked the three members in a complex cluster as the best > the median > the poorest, one point was recorded for this scoring function. An overall success rate was computed accordingly once this analysis was completed over the entire test set. The above standard emphasizes a completely correct rank on each cluster, which is referred to as a "high-level" success rate. In order to provide additional information, a "low-level" success rate was also introduced; i.e., a scoring function is only required to rank the best complex as the top one to gain a point, regardless of the order of the median and the poorest in ranking. Both sets of success rates were used as indicators of ranking power in our benchmark. Similar to the scoring power test, two sets of results were obtained for each scoring function on original crystal structures and optimized complex structures, respectively.

dx.doi.org/10.1021/ci500081m | J. Chem. Inf. Model. 2014, 54, 1717−1736

Table 2. Summary of the Cross-Binding Protein−Ligand Pairs Found in the Test Set

PDB code of the cross-binding ligand	primary target protein	secondary target protein	binding data to the secondary target as recorded in ChEMBL
1SLN	stromelysin-1	growth factor receptor-bound protein 2	K_i = 2 nM
1BCU	thrombin	coagulation factor X	K_i = 90 nM
1O5B	urokinase-type plasminogen activator	coagulation factor X	K_i > 5 μM
1BCU	thrombin	trypsin	K_i = 320 nM
1O5B	urokinase-type plasminogen activator	trypsin	K_i = 8 nM
1MQ6	coagulation factor X	trypsin	K_i = 78 nM
3E93	mitogen-activated protein kinase 14	cell division protein kinase 2	K_d > 10 μM
3JVS	serine/threonine-protein kinase Chk1	cell division protein kinase 2	K_d > 10 μM
3OWJ	casein kinase II	cell division protein kinase 2	K_d > 10 μM
4DES	transthyretin	cell division protein kinase 2	K_d = 1.1 μM
4DJV	β-secretase 1	endothiapepsin	K_d = 660 μM
1BCU	thrombin	angiotensin converting enzyme	K_i = 48 nM
2FVD	cell division protein kinase 2	casein kinase II	IC_{50} = 3.39 μM
2GSS	glutathione *S*-transferase P1-1	casein kinase II	IC_{50} = 77.1 μM
3DD0	carbonic anhydrase II	casein kinase II	K_i = 2.18 μM
2FVD	cell division protein kinase 2	glycogen phosphorylase	K_d = 200 nM
3E93	mitogen-activated protein kinase 14	glycogen phosphorylase	K_d > 10 μM
3JVS	serine/threonine-protein kinase Chk1	glycogen phosphorylase	K_d > 10 μM
3OWJ	casein kinase II	glycogen phosphorylase	K_d > 10 μM
2FVD	cell division protein kinase 2	mitogen-activated protein kinase 14	K_d = 8.7 μM
3JVS	serine/threonine-protein kinase Chk1	mitogen-activated protein kinase 14	K_d > 10 μM
3OWJ	casein kinase II	mitogen-activated protein kinase 14	K_d > 10 μM
1BCU	thrombin	urokinase-type plasminogen activator	K_i = 940 nM
2FVD	cell division protein kinase 2	serine/threonine-protein kinase 6	K_d > 10 μM
3E93	mitogen-activated protein kinase 14	serine/threonine-protein kinase 6	K_d > 10 μM
3JVS	serine/threonine-protein kinase Chk1	serine/threonine-protein kinase 6	IC_{50} = 13 nM
1SLN	stromelysin-1	transthyretin	IC_{50} = 16.2 μM
1UTO	trypsin	transthyretin	IC_{50} = 35.3 μM
2FVD	cell division protein kinase 2	transthyretin	IC_{50} = 20 μM
3DD0	carbonic anhydrase II	transthyretin	K_i = 4.31 μM

2.4.3. Docking Power. It refers to the ability of a scoring function to identify the native binding pose among computer-generated decoys. Ideally, the native binding pose should be identified as the one with the best binding score. As described earlier in this work, a set of decoy binding poses (up to 100) was generated for each protein−ligand complex by using several molecular docking programs. The native binding pose was also mixed into the decoy set to ensure that the ensemble contained at least one correct binding pose. Then, each scoring function was applied to score the decoy set of each complex. The RMSD value between the native binding pose and the best-scored binding pose among all decoys plus the native one was computed using the following standard algorithm:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}[(x_i - x_i')^2 + (y_i - y_i')^2 + (z_i - z_i')^2]}{N}}$$

Here, (x_i, y_i, z_i) and (x'_i, y'_i, z'_i) are the Cartesian coordinates of the ith atom in two binding poses. Only non-hydrogen atoms in the molecule were considered in our computation.

Note that the standard RMSD algorithm given above may produce incorrect results on symmetric molecular structures (see Figure 2 for an example). Thus, we designed a modified algorithm to compute what we called property-matched RMSD values, which will be referred to as $\mathrm{RMSD}^{\mathrm{PM}}$ in this work. Briefly, these algorithms match atom pairs between two binding poses by atom types rather than atom IDs. The atom typing scheme in XLOGP3[63] was employed for this purpose. This scheme considers not only the nature of the given atom itself (e.g., element type and hybridization state) but also the neighboring atoms that are covalently connected to this atom. Our tests indicated that this algorithm handled symmetric molecular structure correctly. The $\mathrm{RMSD}^{\mathrm{PM}}$ values computed by our algorithm are always lower than or equal to the standard RMSD values.

By either algorithm, if the RMSD value between the best-scored binding pose and the native binding pose fell below a predefined cutoff, e.g., RMSD < 2.0 Å, it was recorded as a successful prediction. Once this analysis was completed over the entire test set, an overall success rate was computed for the given scoring function. This success rate was used as an indicator of docking power in our benchmark.

2.4.4. Screening Power. It refers to the ability of a scoring function to identify the true binders to a given target protein among a pool of random molecules. In our benchmark, screening power was evaluated in a cross-docking trial. Our test set includes 65 clusters of complexes, each of which consists of three complexes formed by a certain protein. For each protein, the three known ligands were taken as the positives, whereas the other 195 − 3 = 192 ligand molecules in the test set were taken as the negatives. The assumption here is that since each protein in our test set is distinctive from the others in terms of sequence similarity, the ligands of other proteins are not likely to be the binders of this protein. But this assumption of course should be verified. For this purpose, we searched through the ChEMBL database,[64,65] which is perhaps the largest public

resource of protein−ligand binding data, for possible cross-binders, i.e., those ligand molecules that bind to other proteins besides its primary target. This task was completed by submitting the chemical structures of the ligand molecules and the sequences of the protein molecules in our test set as queries. By examining the searching results from ChEMBL, we indeed identified a total of 30 cross-binding protein−ligand pairs, which are summarized in Table 2. These cross-binders are related to 12 target proteins in our test set. Without exception, the binding affinity of a cross-binder to its secondary target is always lower than its binding affinity to its primary target. Thus, discovery of these cross-binders did not change the definition of the best ligand in each complex cluster. Some of the cross-binders actually have inaccurate binding data from ChEMBL, such as K_d > 10 μM. We treated them as true binders anyway.

For each of the 65 proteins, all 195 ligand molecules were docked into its binding site, yielding a total of 65 × 195 = 12,675 protein−ligand pairs. Since each protein has three different structures in our test set, technically the structure of the best complex in each cluster was selected for this task. Ligand binding poses were generated through essentially the same approach for generating the decoy ligand binding poses in the docking power test. The major difference here was that the binding poses generated by three molecular docking programs were not distributed into 10 RMSD bins. For instead, these binding poses were directly grouped into 50 clusters by their conformational similarity. The binding pose with the lowest strain energy in each cluster was selected as the representative. Thus, up to 50 representative ligand binding poses were selected for each protein−ligand pair. The limit of 50 was chosen to reduce computational costs. It was also because this number of binding poses could be obtained for most protein−ligand pairs.

Each scoring function was applied to compute the binding poses of all 195 ligand molecules (including true binders and negatives) on each target protein. For any given ligand, the best-scored binding pose among all available poses was taken as the predicted binding pose and the corresponding binding score was taken as the predicted binding affinity by this scoring function. All 195 ligand molecules were then ranked according to their binding sores in a descending order. The screening power of a scoring function was measured by counting the total number of true binders among the 1%, 5%, and 10% top-ranked molecules. Enhancement factors (EF) were computed using the following equations:

$$EF_{1\%} = \frac{NTB_{1\%}}{NTB_{total} \times 1\%} \qquad EF_{5\%} = \frac{NTB_{5\%}}{NTB_{total} \times 5\%}$$

$$EF_{10\%} = \frac{NTB_{10\%}}{NTB_{total} \times 10\%}$$

Here, $NTB_{1\%}$, $NTB_{5\%}$, and $NTB_{10\%}$ are the number of true binders observed among the top 1%, 5%, and 10% candidates selected by a given scoring function. NTB_{total} is the total number of true binders for the given target protein, which is typically three for each target protein if there are no cross-binders.

In addition to enhancement factors, we also used the success rate of identifying the best ligand in each complex cluster as an indicator of screening power. For each target protein, if the best ligand was found among the 1%, 5%, and 10% top-ranked candidates, a point was counted for the scoring function under test. Then, the overall success rates on the entire test set at three different levels could be computed accordingly.

2.5. Evaluation on Subsets of the Primary Test Set. In our study, three orthogonal descriptors were used to divide the entire test set into subsets. Each resulting subset thus consists of a group of protein−ligand complexes sharing a similar property. Evaluation results obtained on these subsets supplement those obtain in a nondiscriminatory manner on the entire test set.

The first descriptor is the total number of rotatable bonds on the ligand molecule. Here, a rotatable bond is defined as an acyclic sp^3−sp^3 or sp^3−sp^2 single bond between two non-hydrogen atoms. Single bonds connecting terminal groups, such as -CH_3, -NH_2, -OH, and -X (X = halogen atoms), whose rotation does not produce new rearrangement of heavy atoms are not counted as rotors. This descriptor is determined by the chemical structure of the ligand molecule only.

The second descriptor is the fraction of ligand solvent-accessible surface buried upon binding. The algorithm for computing this property has been described in the companion work of this series.[34] This descriptor is determined by how the ligand binds to the protein.

The third descriptor is a "hydrophobic scale" (*H*-scale) of the binding pocket on the protein molecule. It is computed by summing up the fragmental log *D* values of all pocket residues and then dividing it by the total number of pocket residues. Here, pocket residues refer to those in direct contact with the bound ligand molecule. An amino acid residue on the protein is considered to be in direct contact with the bound ligand if any heavy atom on its side chain is within 4.5 Å from any heavy atom on the ligand. This distance cutoff is approximately the sum of the atomic radii of two sulfur atoms. log *D* is the octanol−water distribution coefficient of an organic molecule measured at a particular pH level. log *D* values are adopted instead of log *P* values (octanol−water partition coefficients) to reflect the protonation states of amino acid residues under a physiological condition. The fragmental log *D* values of each type of amino acid residue are summarized in Table 3. These parameters are cited from the work by Tao et al.,[66] in which they analyzed experimentally measured log *D* values of over 200 oligopeptides with an additive model. Conceptually, a more positive fragmental log *D* indicates a more hydrophobic residue, while a more negative fragmental log *D* indicates a more hydrophilic residue. The hydrophobic scale is determined only by the composition of the binding pocket on the protein molecule.

Table 3. Fragmental log *D* Values of 20 Amino Acid Residues Used in *H*-Scale Computation[a]

residue	log *D*	residue	log *D*	residue	log *D*
ALA	−0.27	GLY	−0.22	PRO	+0.15
ARG	−1.65	HIS	−0.44	SER	−0.45
ASN	−0.98	ILE	+0.69	THR	−0.26
ASP	−2.06	LEU	+0.80	TRP	+1.46
CYS	+0.82	LYS	−2.27	TYR	+0.55
GLN	−1.00	MET	+0.51	VAL	+0.32
GLU	−2.19	PHE	+1.16		

[a]These parameters are cited from the work by Tao et al.[66]

The above three descriptors were all compute using our in-house computer programs. Distributions of these descriptors are given in Figure 1. Classification of the test set by these three

dx.doi.org/10.1021/ci500081m | J. Chem. Inf. Model. 2014, 54, 1717−1736

descriptors is summarized in Table 4. The cutoffs in each set of classification were hand-selected to make each subset distinctive

Table 4. Classification of the Test Set by Three Properties

classification standard	range	no. of complexes	subset symbol
no. of rotatable bonds on the ligand molecule	<4	116	A1
	[4,8]	47	A2
	>8	32	A3
fraction of ligand solvent-accessible surface buried upon binding	<0.65	37	B1
	[0.65−0.85]	117	B2
	>0.85	41	B3
hydrophobic scale of the binding pocket on the protein molecule	<−0.50	42	C1
	[−0.50−0.00]	116	C2
	>0.00	37	C3

from the others and consist of adequate samples at the same time. Classification of each protein−ligand complex under these three criteria are given in the Supporting Information (part II). In our study, docking power and scoring power of each scoring function were also evaluated on each set of subsets using the same methods described in section 2.4. Evaluation of ranking power and screening power were not applicable on these subsets because of the evaluation methods designed by us.

3. RESULTS AND DISCUSSION

3.1. Benchmark Design. Various benchmarks for evaluating docking/scoring methods have been reported in the literature. Our CASF benchmark is different in many aspects from them. It is important to explain the main features of our benchmark at the first place. It will also help the readers interpret the outcomes of our benchmark correctly.

First of all, our CASF benchmark is designed for testing scoring functions. It is true that scoring functions are often applied in combination with molecular docking methods in structure-based drug design. But scoring functions are actually independent from binding pose sampling methods and thus deserve independent treatments. As mentioned in the Introduction, it is a common problem in many other comparative studies that scoring functions were tested in the context of a molecular docking process. In contrast, a basic idea of our benchmark is to separate "scoring" from "docking" (or "sampling") to deduce objective judgment of the performance of scoring functions. All of our evaluation methods are designed accordingly to fulfill this goal.

In our CASF-2013 benchmark, all scoring functions were tested in four different aspects, i.e., scoring power, ranking power, docking power, and screening power. These tests are designed to match the typical applications of scoring functions in practice. For example, the very essential goal of molecular docking is to derive the ligand binding mode to a given target protein, and this process is typically guided by a scoring function. The docking power test is designed accordingly to evaluate if a scoring function can identify the correct binding pose. Technically, we used decoy ligand binding poses generated in prior so that all scoring functions are evaluated on the same ground. These decoy binding poses were generated by combining the outcomes of several molecular docking programs to ensure that the final decoy set was not dominated by any particular docking program.

Besides binding pose prediction, one may want the scoring function to provide reasonable estimation of absolute binding affinities. The scoring power test is designed to reflect the performance of a scoring function in this aspect. In other scenarios, one may simply want to separate tight binders from weak binders or to rank known binders correctly. This task does not require binding scores to be in a linear correlation with binding data. Thus, a separate ranking power test is necessary. Two things need to be emphasized here: (i) Scoring power was evaluated across various protein−ligand complexes in our benchmark, whereas ranking power is evaluated on protein−ligand complexes formed by the same protein. (ii) Both scoring power and ranking power were evaluated directly on experimentally resolved protein−ligand complex structures in order to avoid uncertainties introduced by docking methods.

As a new development of our benchmark, a screening power test has been introduced. Since docking-based virtual screening is widely applied in lead discovery, it is desirable to evaluate scoring functions in this aspect. Separating true binders from random molecules, however, is a much more complicated issue. To perform such a test, one would need some true binders (positives) as well as some random molecules that do not bind to the given target protein (negatives). While true binders can be selected relatively easily, how to select a standard set of random negatives is not so straightforward. The DUD/DUD-E benchmark[41,42] developed by Schoichet's group is perhaps the most popular answer to this problem. Originally, DUD assembled a total of 40 hand-picked targets and 2,950 ligands of them. For each ligand, a set of molecules are retrieved from the "druglike" subset of ZINC[67] as negatives. As a key idea of DUD, these negatives are selected intentionally to be physically similar but topologically distinct from the true ligand. Recently, this benchmark has been upgraded to DUD-E,[42] which includes a total of 102 target proteins as well as 22,886 known ligands for them, which are drawn from ChEMBL.[64,65] Fifty negatives are selected from ZINC for each known ligand, which have most dissimilar chemical structures to the true ligands but yet share similar physicochemical properties.

In our study, we decided not to use an external set of negatives. Instead, for each protein included in the test set, the 192 ligand molecules in other complex clusters were taken as the negatives. Therefore, we do not need an additional set of rules for selecting the negatives. As implied above, those rules themselves could be subject of debates. As compared to the comprehensive DUD-E benchmark, our test set includes a rather limited number of true binders, i.e., normally three, for each target protein, whereas each target protein in DUD-E has several dozens to several hundreds of known binders. This shortcoming of our test set will be overcome gradually in the future. As mentioned in the companion work of this series,[34] we have practical plans to expand the PDBbind core set. The positive/negative ratio in DUD-E is 1:50, whereas it is normally 1:64 in our test set. Thus, our test set is also very challenging for a virtual screening trial. A major advantage of our test set is that the quality of protein−ligand complex structures and binding data is at a higher level. Besides, diverse binding poses of each ligand to each target protein, a total of 12,675 protein−ligand pairs, were generated through an extensive effort. Based on this, scoring functions alone can be evaluated in virtual screening trials without the need of molecular docking methods. In contrast, DUD/DUD-E is more suitable for testing docking/scoring schemes in the black-box mode.

Table 5. Performance of All 20 Scoring Functions in the Scoring Power Test

	on crystal structures			on optimized structures		
scoring function[a]	N[b]	R[c]	SD[d]	N	R	SD
X-Score^{HM}	195	0.614	1.78	195	0.624	1.76
ΔSAS	195	0.606	1.79	195	0.596	1.81
ChemScore@SYBYL	195	0.592	1.82	194	0.546	1.89
ChemPLP@GOLD	195	0.579	1.84	195	0.538	1.90
PLP1@DS	195	0.568	1.86	195	0.557	1.87
G-Score@SYBYL	195	0.558	1.87	189	0.456	2.01
ASP@GOLD	195	0.556	1.88	195	0.529	1.92
ASE@MOE	195	0.544	1.89	195	0.547	1.89
ChemScore@GOLD	189	0.536	1.90	187	0.473	1.99
D-Score@SYBYL	195	0.526	1.92	194	0.530	1.91
Alpha-HB@MOE	195	0.511	1.94	193	0.487	1.97
LUDI3@DS	195	0.487	1.97	195	0.448	2.02
GoldScore@GOLD	189	0.483	1.97	192	0.479	1.97
Affinity-dG@MOE	195	0.482	1.98	193	0.480	1.98
LigScore2@DS	190	0.456	2.02	183	0.390	2.07
GlideScore-SP	169	0.452	2.03	155	0.425	2.11
Jain@DS	191	0.408	2.05	194	0.341	2.13
PMF@DS	194	0.364	2.11	193	0.357	2.10
GlideScore-XP	164	0.277	2.18	149	0.336	2.22
London-dG@MOE	195	0.242	2.19	195	0.252	2.18
PMF@SYBYL	191	0.221	2.20	190	0.189	2.19

[a]Scoring functions are ranked by the Pearson correlation coefficients obtained on the original crystal structures. [b]Number of complexes receiving positive (favorable) binding scores by this scoring function. [c]The Pearson correlation coefficient between experimental binding data and computed binding scores. [d]The standard deviation (in log K_a units) in the linear correlation between experimental binding data and computed binding scores.

The last issue is about what the readers should expect in the following sections. Our discussion will focus on the overall trends revealed in the outcomes of our tests. Discussion on why a particular scoring function fails or succeeds in certain cases is certainly related to how that scoring function is designed. Some scoring functions implemented in commercial software lack detailed documentations, which makes an in-depth analysis of their algorithms impossible. Besides, that type of discussion would better be given by the developers of that scoring function. Our CASF benchmark is a third-party evaluation after all. The readers also should be aware that we processed the protein−ligand complex structures in the test set without paying enough attention to some special factors, such as alternative protonation states and critical bridging water molecules. These factors may affect protein−ligand interactions for a certain number of samples in our test set. It is possible to treat these factors in more appropriate ways if one works on a handful of proteins. But for a diverse data set such as ours, it is not very practical to do so. Therefore, our benchmark measures the "default" performance rather than the "optimal" performance of those scoring functions.

3.2. Evaluation Results of Scoring Power and Ranking Power. We start our discussion with scoring power and ranking power tests because they are conducted directly on known three-dimensional structures of protein−ligand complexes. In our benchmark, scoring power refers to the ability of a scoring function to generate binding scores in linear correlation with experimental binding data for a set of protein−ligand complexes. The correlations between the binding scores calculated by 20 scoring functions and experimental binding constants on the entire test set are given in Table 5. The best-scoring function in this test is X-Score^{HM}, which produced a correlation coefficient (R) of 0.614 and a standard deviation (SD) of 1.78 log K_a units (corresponding to 2.43 kcal/mol in binding free energy). The next three top-ranked scoring functions are ChemScore@SYBYL, ChemPLP@GOLD, and PLP1@DS, which produced correlation coefficients in the range of 0.568−0.592. In our previous CASF-2007 benchmark, the top four scoring functions in the same test were X-Score^{HM}, DrugScore^{CSD}, ChemScore@SYBYL, and PLP1@DS. Thus, these two sets of results are basically consistent except that DrugScore^{CSD} has been replaced by ChemPLP@GOLD. This finding indicates the robustness of our evaluation considering that the test set used in this study has only 13% overlap with the one used in CASF-2007.

Scatter plots of the experimental binding constants versus the binding scores produced by the top four scoring functions in this test are given in Figure 3. Correlations produced by each scoring function are actually in different patterns. But there are a number of common outliers, for which virtually all scoring functions significantly overestimate or underestimate their binding constants. An obvious trend here is that most significant outliers, e.g., residuals > $\pm\sigma$, are among protein−ligand complexes with medium-level or high-level affinities. Distributions of fitting residuals generated by these four scoring functions are shown in Figure 4. One can see that they all deviate from a normal distribution more or less. It gives a fair warning that statistical means for comparing two normal distributions, such as t test and F test, may not be applicable here. The residuals associated with the best complexes, the median complexes, and the poorest complexes are indicated in different colors in Figure 4. It is obvious that binding constants of the poorest (i.e., the one with the lowest binding affinity in each complex cluster) tend to be overestimated, while binding constants of the best (i.e., the one with the highest binding affinity in each complex cluster) tend to be underestimated. All scoring functions encounter a greater problem at the high-affinity end than at the low-affinity end, which verifies the

dx.doi.org/10.1021/ci500081m | J. Chem. Inf. Model. 2014, 54, 1717−1736

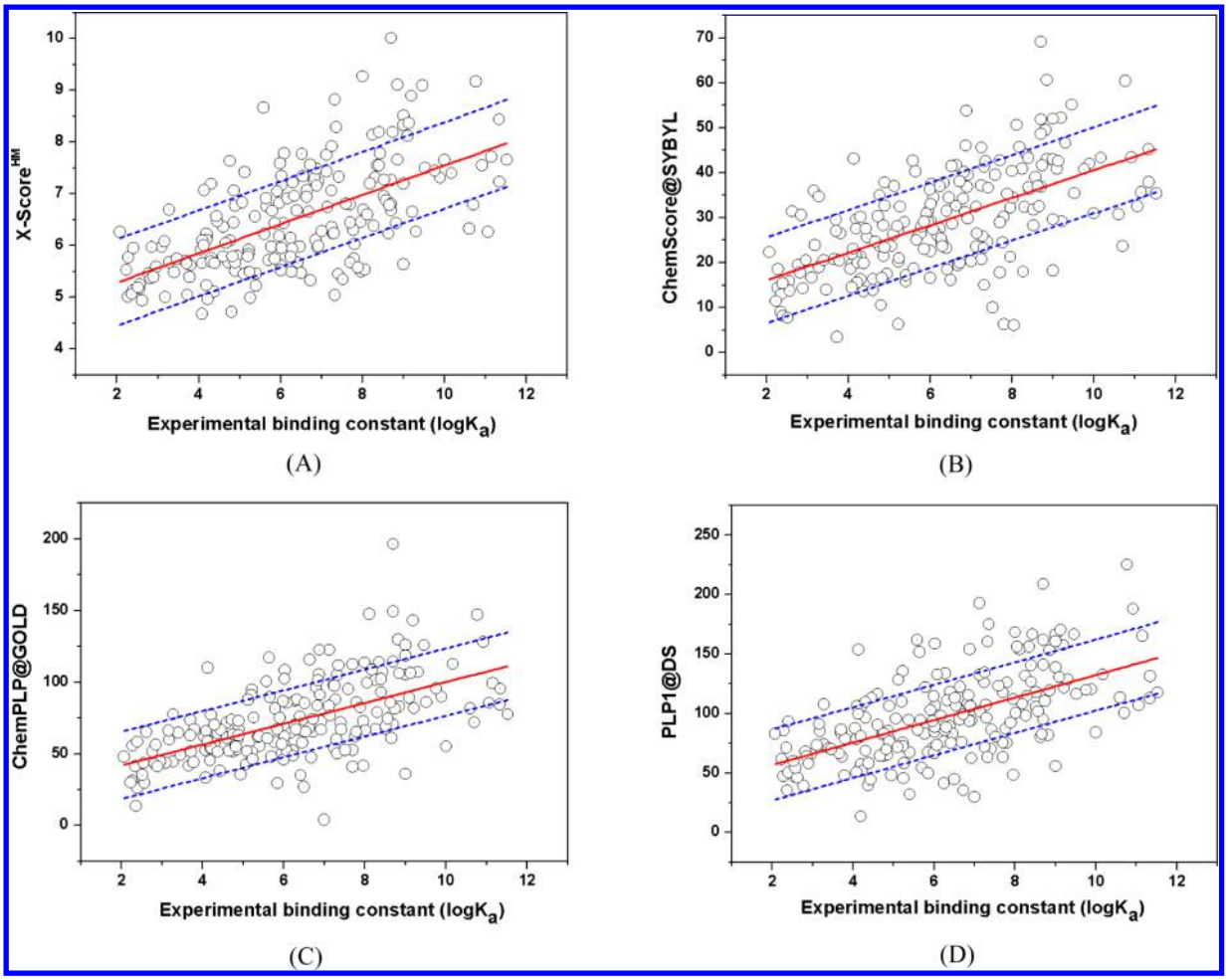


Figure 3. Scatter plots of experimental binding constants versus binding scores produced by the top four scoring functions in the scoring power test: (A) X-Score^{HM}; (B) ChemScore@SYBYL; (C) ChemPLP@GOLD; (D) PLP1@DS. Here, the regression line is indicated by the solid red line, while the residual range within $\pm\sigma$ is indicated between two blue dashed lines.

qualitative observation in Figure 3. In fact, there are approximately 10 most significant outliers with residual > $\pm 2\sigma$ in our test set, and they are all "the best" complexes without exception (Figure 4). These outliers have crippled the statistical data of all scoring functions by a considerable extent.

Another set of evaluations were conducted on complex structures in which the ligand binding pose was optimized in situ. The results are also summarized in Table 5. One can see that the performance of most scoring functions retains at basically the same level or gets slightly worse. Thus, ranking of scoring functions is essentially the same as that obtained on the original crystal structures. But the performances of several force-field-based scoring functions are affected more obviously. In particular, the correlation coefficient of G-Score@SYBYL reduced from 0.558 to 0.456 on optimized complex structures. At the same time, the total number of complexes that could be processed successfully by G-Score@SYBYL reduced from 195 to 189. The same trend was also observed for LigScore@DS. This is understandable since force fields include distance-sensitive energy terms so that small conformational changes in ligand binding pose may lead to a significant change in the final binding score. Apparently, scoring functions that are not sensitive to minor conformational changes are more welcome in practice.

In our benchmark, ranking power is defined as a different feature from scoring power. To rank the known binders to a given target protein correctly, scoring functions under test do not have to produce binding scores in linear correlation with experimental binding data. Thus, higher success rates are expected in the ranking power test. The success rates of all 20 scoring functions in this test are summarized in Table 6. In terms of high-level success rates, a total of 10 scoring functions achieved success rates over 50%. X-Score^{HM} and ChemPLP@GOLD achieved the highest success rate of 58.5%. They are followed by PLP2@DS, GoldScore@GOLD, and ChemScore@SYBYL. Note that these scoring functions are also the top five in the scoring power test (Table 5).

One probably has expected even higher success rates in the ranking power test. It is because in each complex cluster, the binding affinity gap between best/median and median/poorest is at least 10-fold. Thus, an incorrect ranking of the three members in a cluster will happen only if a scoring function makes terrible predictions of binding affinities. The low-level success rates provide additional information about the outcomes of this test (Table 6). Take X-Score^{HM} for example, its high-level and low-level success rates are 58.5% and 72.3%, respectively. It indicates that in 72.3 − 58.5 = 13.8% of cases, this scoring function messed up the rank between the median complex and the poorest complex, which we call "softcore"

dx.doi.org/10.1021/ci500081m | J. Chem. Inf. Model. 2014, 54, 1717−1736

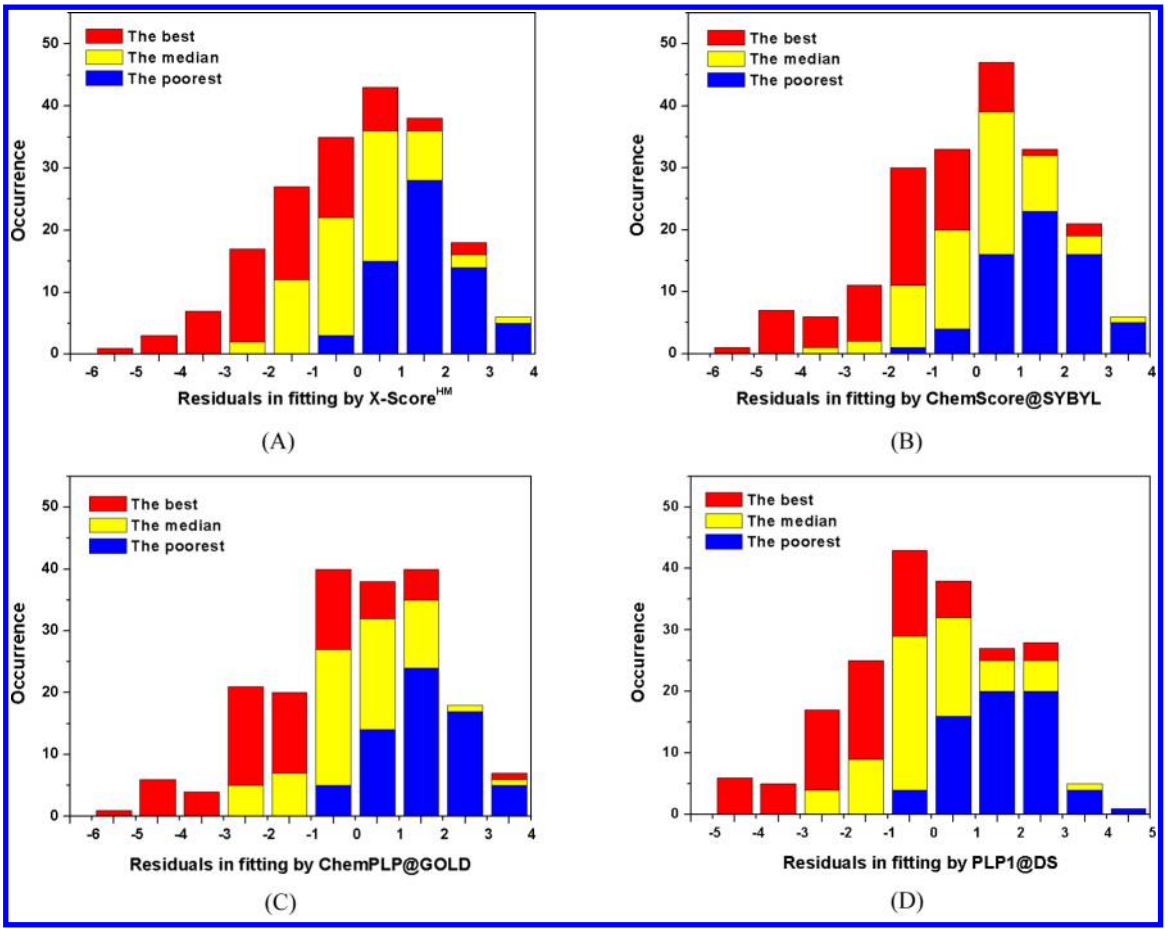


Figure 4. Residuals in fitting binding scores produced by four scoring functions to experimental binding constants: (A) X-Score^{HM}; (B) ChemScore@SYBYL; (C) ChemPLP@GOLD; (D) PLP1@DS. Residuals associated with the best, the median, and the poorest complexes in the test set are colored differently.

errors. In the other 100 − 72.3 = 27.7% of cases, it failed to rank the best complex to the top, which we call "hardcore" errors. By examining the gaps between high-level and low-level success rates (Table 6), one can see that most scoring functions under our evaluation made softcore errors in 12−17% cases. However, a few scoring functions tend to make more softcore errors than the others, such as GoldScore@GOLD, G-Score@SYBYL, ASP@GOLD, and ASE@MOE. It seems that these several scoring functions tend to be more problematic when handling medium-affinity and low-affinity complexes. This could be a drawback for virtual screening jobs because the active hits discovered through virtual screening normally have medium-level or low-level affinities to the target protein. Indeed, these several scoring function were not among the top ones in our screening power test.

When the optimized complex structures were applied in the ranking power test, the success rates of most scoring functions became more or less lower. The trend is the same no matter if high-level or low-level success rates are considered. Only a few scoring functions, including X-Score^{HM}, ChemScore@SYBYL, and LigScore@DS, were able to maintain their success rates at approximately the same level as those obtained on crystal structures. It is a common practice in molecular docking to optimize computer-generated ligand binding poses further with a standard force field. Therefore, we recommend those scoring functions that are least affected by minor conformational changes for conducting scoring/ranking tasks.

Here, a special note needs to be made about GlideScore. This scoring function is the least robust one in our scoring and ranking power tests. Both SP and XP modes failed to produce reasonable binding scores for 30−40 complexes in the test set no matter if the crystal structures or optimized structures were used as inputs. The poor performance of GlideScore-XP and -SP in the ranking power test is largely attributed to this technical problem. In some cases, this problem could be resolved if the input complex structure was prepared by using the Schrödinger software instead. We did not attempt that though since it would not be fair to other scoring functions under evaluation if GlideScore received special treatment. We want to emphasize that technical robustness is also a very important quality for scoring functions. Besides, GlideScore-XP (i.e., the extra precision mode) is supposed to be a higher-level method as compared to GlideScore-SP (i.e., the standard precision mode).[5,6,61] Indeed, GlideScore-XP was slower than GlideScore-SP by a few-fold. But GlideScore-XP did not really provide a better performance in both scoring and ranking power tests. This is a good example demonstrating the value of a third-party evaluation.

To summarize, the performance of all scoring functions in our scoring/ranking power tests is not very promising. Even the best-scoring functions achieved only modest success rates. The binding constants in our test set span nearly 10 orders of magnitude (log K_a = 2−12). One can see in Figure 3 that the linear response range of these scoring functions is typically

dx.doi.org/10.1021/ci500081m | J. Chem. Inf. Model. 2014, 54, 1717−1736

Table 6. Performance of All 20 Scoring Functions in the Ranking Power Test

	success rates (%) on crystal structures		success rates (%) on optimized structures	
scoring function[a]	high-level[b]	low-level[c]	high-level	low-level
X-Score^{HM}	58.5	72.3	56.9	73.8
ChemPLP@GOLD	58.5	72.3	46.2	61.5
PLP2@DS	55.4	72.3	47.7	67.7
GoldScore@GOLD	55.4	76.9	43.1	66.2
ChemScore@SYBYL	53.8	67.7	52.3	69.2
Affinity-dG@MOE	53.8	66.2	36.9	50.8
LigScore1@DS	52.3	61.5	50.8	63.1
Alpha-HB@MOE	52.3	66.2	47.7	64.6
G-Score@SYBYL	52.3	72.3	46.2	61.5
LUDI1@DS	52.3	69.2	44.6	66.2
D-Score@SYBYL	49.2	63.1	52.3	63.1
ΔSAS	49.2	67.7	50.8	69.2
PMF@DS	49.2	66.2	46.2	63.1
ASP@GOLD	47.7	72.3	38.5	60.0
ChemScore@GOLD	46.2	63.1	33.8	53.8
London-dG@MOE	43.1	60.0	40.0	60.0
PMF@SYBYL	43.1	61.5	30.8	53.8
GlideScore-SP	43.1	56.9	21.5	38.5
Jain@DS	41.5	58.5	44.6	63.1
ASE@MOE	40.0	64.6	43.1	63.1
GlideScore-XP	35.4	47.7	32.3	46.2

[a]Scoring functions are ranked by their high-level success rates obtained on the original crystal structures. [b]Ranking the three complexes in a cluster as the best > the median > the poorest. [c]Ranking the best complex in a cluster as the top one.

between log K_a = 4 and log K_a = 9. But this binding affinity range is most relevant to pharmaceutical interests. Thus, these scoring functions are not so incompetent in practice as what they look in our tests. The top-ranked scoring functions in the scoring power test and the ranking power test are essentially the same. Most of them are empirical scoring functions. Unlike force-field-based scoring functions or statistical potentials, empirical scoring functions are relatively flexible to incorporate new energies terms. Moreover, they are typically calibrated on data sets of diverse protein−ligand complexes. Thus, empirical scoring functions, if they are well-designed, have certain advantages over other types of scoring functions in terms of scoring/ranking power.

Performance of all 20 scoring functions in our scoring/ranking power tests are basically in a continuous spectrum without an obvious gap (Tables 5 and 6). It is actually subjective to set a line to separate "good" scoring functions from others. This is why we introduced a näive scoring function (ΔSAS) as a reference in comparison. We were very surprised to observe that ΔSAS outperformed most scoring functions in the scoring power test. In fact, only X-Score^{HM} produced slightly better statistical results than ΔSAS. ΔSAS also outperformed half of the scoring functions in the ranking power test. This embarrassing fact indicates that current scoring functions have not advanced too far from very simple accounts of protein−ligand interactions. It also suggests that the major problem in computing ligand binding affinities does not lie in the nonpolar aspect. Nonpolar interactions are conventionally modeled by the nondiscriminative contacting surface area, and many scoring functions actually have such an account explicitly or implicitly. Instead, polar interactions, as well as the desolvation effect associated with them, remain as the real challenge. The observation that current scoring functions encounter a greater problem with high-affinity protein−ligand complexes also supports this statement. It is because hydrophobic contact cannot lead to a very high binding affinity. Instead, a network of specific polar interactions plus a special geometry of the binding pocket are often witnessed among high-affinity protein−ligand complexes.

3.3. Evaluation Results of Docking Power. In our benchmark, the docking power of a scoring function refers to its ability of identifying the native ligand binding pose among

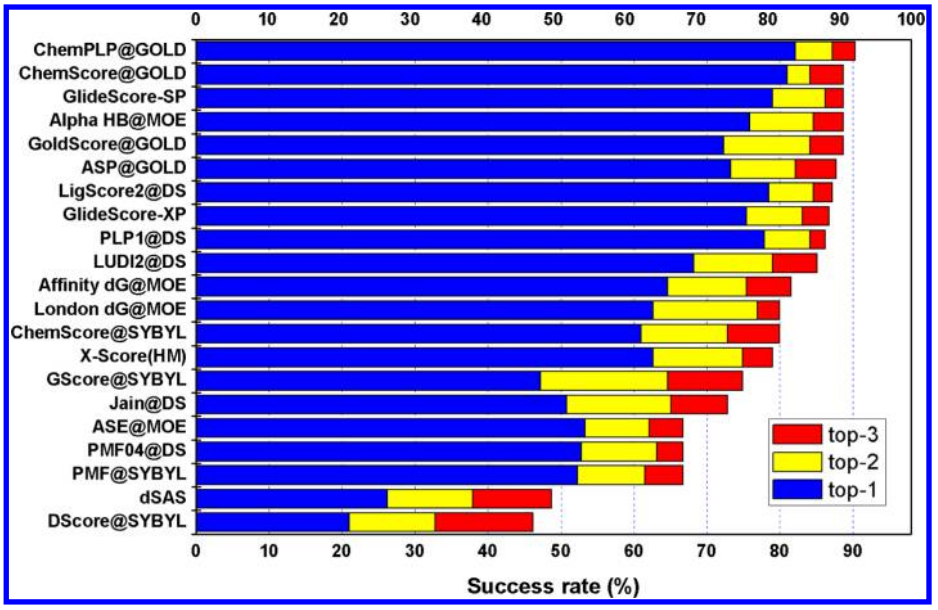


Figure 5. Success rates of 20 scoring functions in the docking power test when the top one (blue bars), the top two (yellow bars), or the top three (red bars) best-scored ligand binding poses are considered to match the native pose. Here, scoring functions are ranked by their success rates when the top three binding poses are considered. The criterion for judging a successful prediction is that the property-matched RMSD value between one best-scored binding pose and the native binding pose is smaller than 2.0 Å.

dx.doi.org/10.1021/ci500081m | J. Chem. Inf. Model. 2014, 54, 1717−1736

computer-generated decoys. The basic method was established in our previous CASF-2007 benchmark.[45] In this study, three new molecular docking programs, i.e., GOLD, Surflex, and MOE, were employed for generating ligand binding poses. The postprocessing workflow for selecting diverse and representative binding poses was also reformed to be more effective. In addition, property-matched RMSD values were introduced to overcome the potential problems made by the standard RMSD algorithm on symmetric molecular structures.

The success rates of all 20 scoring functions in the docking power test are illustrated in Figure 5. This figure presents three sets of results when the top one, two, or three best-scored binding poses were considered in success rate computation. Generally, the scoring functions under our evaluation performed much better in this test. Even if only the best-scored binding pose is considered, a total of 10 scoring functions produced success rates above or really close to 70%. In particular, ChemPLP@GOLD and Chemscore@GOLD achieved success rates above 80%. They are followed closely by GlidScore-SP, LigScore2@DS, PLP1@DS, and GlideScore-XP with success rates between 75−79%. In our previous CASF-2007 benchmark, these several scoring functions, except for the new ChemPLP scoring function, were also top-ranked in the same test. However, their success rates in CASF-2007 ranged from 54% to 69%. The higher success rates produced by the same scoring functions in this study should be largely attributed to the improved quality of the decoy sets. In other words, the current decoy sets contain more binding poses that resemble the native one (but not the native one) so that a scoring function has a better chance to select out one of them. In fact, if we remove the native binding pose from each decoy set and then repeat the docking power evaluation, the success rates of all scoring functions will decrease by only 1−5% (see the Supporting Information, part V). Besides, introduction of the property-matched RMSD algorithm is also helpful. If the standard RMSD values are used in this test instead, the success rates of all scoring functions will decrease by 1−4% (see the Supporting Information, part V). Nevertheless, ranking of scoring functions in this test is almost the same no matter if the property-matched RMSD or the standard RMSD values are used.

In molecular docking practice, one may also want to examine other possible binding poses besides the best-score one. Indeed, our results show that for most scoring functions, considering one more top-ranked binding pose generally increases their success rates by up to 16%. If the top three binding poses are considered, about half of the scoring functions are able to identify the true binding pose with success rates above 80% (Figure 5). However, our results also imply that considering even more top-ranked binding poses will probably result in rather limited improvements and thus is not recommended.

As compared to the scoring/ranking power tests, the performance of all scoring functions is more diverse in the docking power test. Some scoring functions achieved fairly high success rates. For example, all scoring functions implemented in GOLD and GLIDE performed well in this test. On the other hand, a few scoring functions, such as Jain@DS, PMF04@DS, ASE@MOE, PMF@SYBYL, G-Score@SYBYL, and D-Score@SYBYL, achieved success rates only around 50%. We want to emphasize that, in this test, there is no necessary connection between the performance of a scoring function and its category. For example, PMF04@DS, ASE@MOE, and PMF@SYBYL are pairwise statistical potentials; so is ASP@GOLD. But ASP@GOLD is among the best ones in this test. For another example, G-Score@SYBYL is supposed to be an implementation of the original GoldScore scoring function in GOLD. But the performances of these two versions are quite different.

The last comment is on ΔSAS. Its performance in this docking power test is virtually the worst among all scoring functions under evaluation. This finding is not totally unexpected. As mentioned earlier in this work, ΔSAS is a descriptor of the nonpolar factors in protein−ligand interactions, which is nondiscriminative in nature. Therefore, it is not very capable to differentiate the decoys from the true binding pose since they are in fact the same ligand molecule. To achieve a good docking power, a scoring function needs necessary consideration of specific, directional polar interactions, such as hydrogen bonds. Most current scoring functions do have this type of account, which explains their relatively successful performance in the docking power test.

3.4. Evaluation Results of Screening Power. The screening power test is a new development in our CASF benchmark. Screening power of a scoring function refers to its ability of identifying the true binders to a given target protein among random molecules. As the first set of evaluations, performances of all scoring functions are examined by their enhancement factors. A larger enhancement factor corresponds to a higher probability of concentrating true binders among top-ranked candidates in virtual screening. The average enhancement factors computed on the 65 target proteins in our test set for all scoring functions are summarized in Table 7. GlideScore-SP is the best one in this test with an average enhancement factor close to 20 at the top 1% level. It is followed by ChemScore@GOLD, GlideScore-XP, LigScore2@DS, ChemPLP@GOLD, LUDI1@DS, and ASP@GOLD, which all have average enhancement factors over 10 at the

Table 7. Enrichment Factors of All 20 Scoring Functions in the Screening Power Test

	enrichment factor		
scoring function[a]	top 1%	top 5%	top 10%
GlideScore-SP	19.54	6.27	4.14
ChemScore@GOLD	18.90	6.83	4.08
GlideScore-XP	16.81	6.02	4.07
LigScore2@DS	15.90	6.23	3.51
ChemPLP@GOLD	14.28	5.88	4.31
LUDI1@DS	12.53	4.28	2.80
ASP@GOLD	12.36	6.23	3.79
Affinity-dG@MOE	8.21	4.15	3.19
London-dG@MOE	8.08	3.36	2.51
GoldScore@GOLD	7.95	4.52	3.16
PLP1@DS	6.92	4.28	3.04
Jain@DS	5.90	2.51	1.80
PMF@SYBYL	5.38	2.21	1.90
ChemScore@SYBYL	5.26	2.38	2.18
Alpha-HB@MOE	4.87	3.23	1.32
PMF04@DS	4.87	2.87	2.63
ASE@MOE	4.36	2.35	1.59
X-Score^{HM}	2.31	2.14	1.41
D-Score@SYBYL	2.31	1.79	1.46
G-Score@SYBYL	1.92	1.26	1.44
ΔSAS	1.41	1.28	1.12

[a]Scoring functions are ranked by their average enrichment factor obtained at the top 1% level.

top 1% level. It is interesting to observe that GlideScore-SP outperforms GlideScore-XP again in this test, although the advantage is only marginal. At the low end, a few scoring functions produced rather trivial enhancements. Surprisingly, among them is X-Score^{HM}, which is the top scoring function in the scoring/ranking power test. Enhancement factors at the top 5% and top 10% levels are considerably lower for all scoring functions because our test set consists of a rather limited number of true binders (normally three) to each target protein. But the ranking of all scoring functions at these two levels remain basically the same as evaluated at the top 1% level (Table 7).

The screening powers of all scoring functions were also evaluated by their success rates of finding the best ligand molecule for each target protein among top-ranked candidates (Table 8). This criterion only regards the best ligand molecules,

Table 8. Success Rates of Finding the Best Ligand Molecule of All 20 Scoring Functions in the Screening Power Test

	success rates of finding the best ligand molecule among (%)[b]		
scoring function[a]	top 1%	top 5%	top 10%
GlideScore-SP	60.0 (39)	72.3 (47)	76.9 (50)
GlideScore-XP	52.3 (34)	69.2 (45)	73.8 (48)
ChemScore@GOLD	49.2 (32)	78.5 (51)	83.1 (54)
LigScore2@DS	47.7 (31)	75.4 (49)	83.1 (54)
ChemPLP@GOLD	41.5 (27)	70.8 (46)	84.6 (55)
LUDI2@DS	38.5 (25)	53.8 (35)	66.2 (43)
ASP@GOLD	36.9 (24)	75.4 (49)	81.5 (53)
Affinity-dG@MOE	23.1 (15)	50.8 (33)	66.2 (43)
PLP1@DS	21.5 (14)	52.3 (34)	70.8 (46)
GoldScore@GOLD	21.5 (14)	52.3 (34)	66.2 (43)
London-dG@MOE	21.5 (14)	36.9 (24)	49.2 (32)
Jain@DS	16.9 (11)	29.2 (19)	40.0 (26)
ChemScore@SYBYL	15.4 (10)	33.8 (22)	50.8 (33)
Alpha-HB@MOE	13.8 (9)	36.9 (24)	63.1 (41)
PMF@SYBYL	13.8 (9)	23.1 (15)	38.5 (25)
PMF04@DS	12.3 (8)	30.8 (20)	47.7 (31)
ASE@MOE	12.3 (8)	30.8 (20)	38.5 (25)
X-Score^{HM}	9.23 (6)	21.5 (14)	32.3 (21)
D-Score@SYBYL	6.15 (4)	20.0 (13)	24.6 (16)
G-Score@SYBYL	4.62 (3)	16.9 (11)	30.8 (20)
ΔSAS	3.08 (2)	15.4 (10)	24.6 (16)

[a]Scoring functions are ranked by their success rates obtained at the top 1% level. [b]Numbers in brackets are the number of successful cases, for which the upper limit is 65.

which is more straightforward as compared to the enhancement factor. One can see in Table 8 that as expected, ranking of all 20 scoring functions is essentially unaltered as compared to the results by using enhancement factors. If the success rates at the top 1% level are considered, GlideScore-SP leads all scoring functions with a success rate of 60%. This level of success rate is very impressive. Considering that the total number of true binders plus negatives used in our test is 195, it means that GlideScore-SP ranks the best ligand molecule either as the first or the second among all 195 candidates by a chance of 60%. If the success rates at the top 5% or 10% levels are considered, then a few scoring functions, such as ChemScore@GOLD, ASP@GOLD, and LigScore2@DS, provide comparable or even marginally better performance than GlideScore-SP.

A notable observation here is that, without exception, the top-ranked scoring functions in the screening power test are also top-ranked in the docking power test. This result is actually logical. A true binder should have a funnel-shape binding energy landscape, and its correct binding pose locates at the bottom of the funnel. A nonbinder should have a rather flat binding energy landscape on which no particular binding pose is favored over the others. Thus, if a scoring function is able to identify the low-energy binding pose of a true binder among other binding poses of the same molecule (i.e., good docking power), it should be able to discriminate the low-energy binding pose of a true binder from the hypothetical binding poses of other nonbinders as well (i.e., good screening power). The promising docking/screening power of scoring functions explains why they have been successful in virtual screening projects although they are not very capable in computing ligand binding affinities.

The outcomes of our screening power test also suggest that geometrical complementarity to the binding pocket is still a decisive factor for discriminating true binders from random molecules. Taking GlideScore for example, it requires a grid lattice to be generated first to define the binding pocket. This lattice serves as a geometry filter in subsequent scoring steps. Only the molecules with a similar shape to the native ligand can pass this filter. Again, the performance of ΔSAS is virtually the worst in the screening power test. This demonstrates that this nondiscriminative descriptor is not suitable for this task.

3.5. Evaluation Results Obtained on Subsets. In our benchmark, all scoring functions were also evaluated on certain subsets of the primary test set. Comparing the results obtained on these subsets with those obtained on the entire test set provides additional clues to the decisive factors affecting scoring function performances. This strategy was introduced since our previous CASF-2007 benchmark.[45] In this study, three orthogonal descriptors were chosen for subset classification (Table 4), including the number of rotatable bonds on the ligand (subsets A1−A3), the buried percentage of solvent-accessible surface of the ligand (subsets B1−B3), and a hydrophobic scale of the binding pocket on protein (subsets C1−C3). The first descriptor is newly introduced to this study, whereas the other two were also used in CASF-2007.

Results of the docking power test on subsets are illustrated in Figure 6. The complete results can be found in the Supporting Information (part V). Although the performance of each individual scoring function varies considerably here, some general trends can be observed. First, most scoring functions achieved higher success rates on subset A1 or A2 than A3. For example, the success rate gaps between A1/A2 and A3 for the several top-ranked scoring functions can be up to 30%. This indicates that it is still difficult for current scoring functions to identify the correct binding poses of flexible ligand molecules. GlideScore-SP is an exception here, which is able to maintain a relatively robust performance on all three A-subsets. Second, all scoring functions achieved higher success rates on subset B2 or B3 than B1. For example, the success rates of ChemPLP@GOLD on subsets B1, B2, and B3 are 62%, 82%, and 95%, respectively. This trend is also understandable. The ligand molecule will not have many alternative binding poses if it is largely buried inside the binding pocket. Consequently, scoring functions are less likely to miss the correct binding pose in such cases. Third, most scoring functions achieved higher success rates on subset C1 or C2 than C3. Subset C3 consists of complexes with more hydrophobic binding pockets. Hydro-

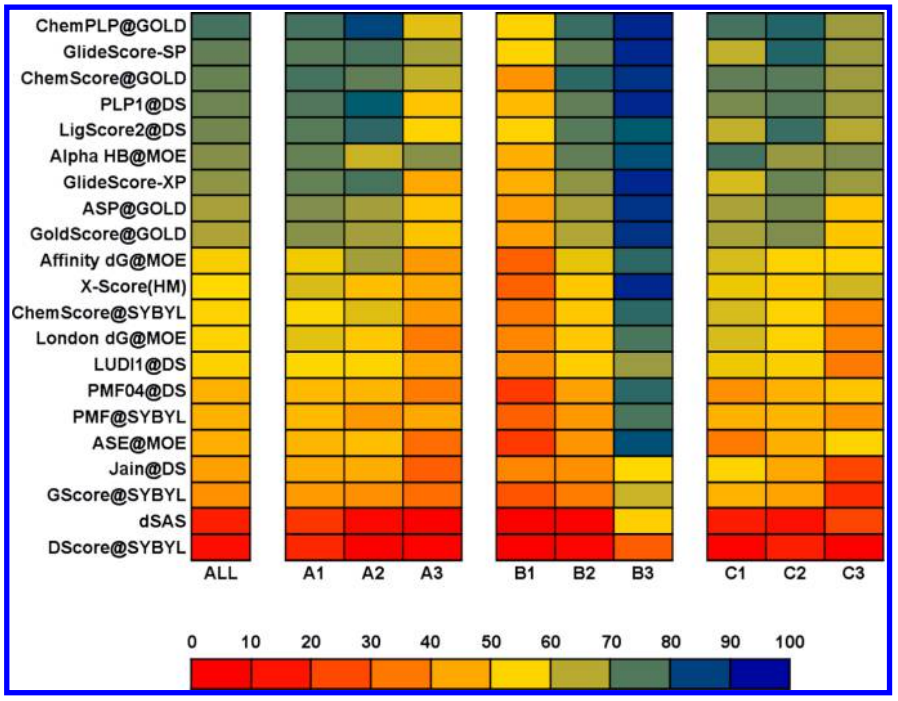


Figure 6. Docking power of all 20 scoring functions on three sets of subsets. Their performance on the entire test set is displayed on the left as reference. Definitions of subsets are given in Table 3.

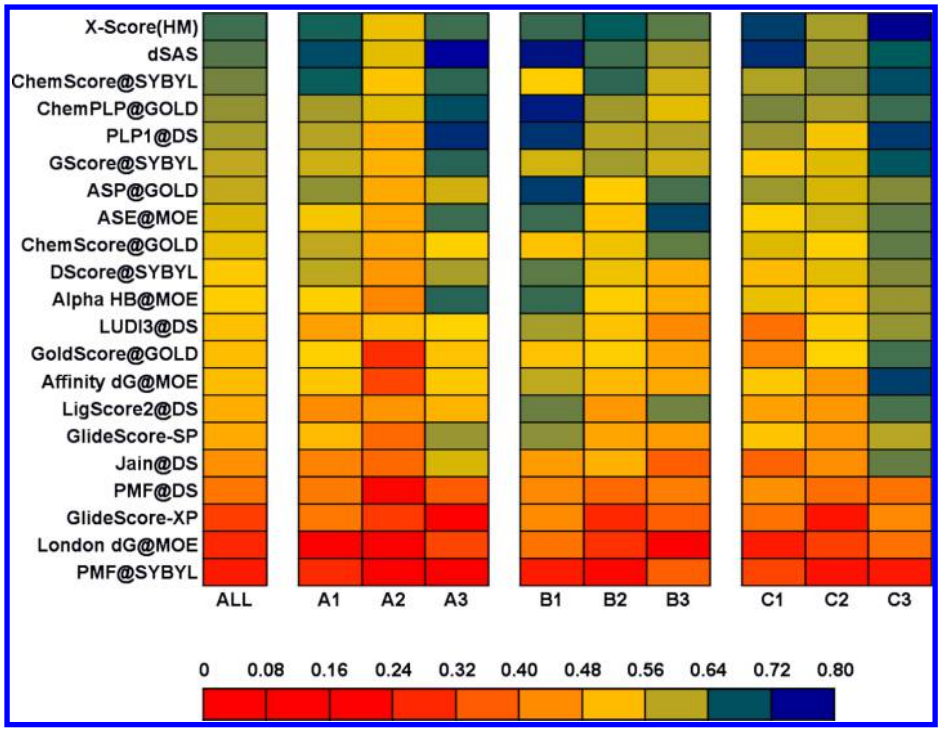


Figure 7. Scoring power of all 20 scoring functions on three sets of subsets. Their performance on the entire test set is displayed on the left as reference. Definitions of subsets are given in Table 3.

phobic contacts are expected to be the dominant factor in protein−ligand binding in such cases. Due to the nonspecific and nondirectional nature of hydrophobic contacts, it is more challenging for scoring functions to distinguish the true ligand binding pose from decoys.

Results of the scoring power test on subsets are illustrated in Figure 7. The complete results can be found in the Supporting Information (part IV). The results obtained on subsets A1−A3 are somewhat unexpected: While subset A3 is more challenging than subset A1 in the docking power test, it is not so in this test. The top-ranked scoring functions in this test all produced better performance on subset A3 than on the entire test set. For example, ΔSAS produced a fairly good correlation coefficient of 0.738 on subset A3. The same trend is also observed in the results obtained on subsets B1−B3. Subset B1 is more challenging than subset B3 in the docking power test, but

dx.doi.org/10.1021/ci500081m | J. Chem. Inf. Model. 2014, 54, 1717−1736

scoring functions tend to perform better on subset B1 in the scoring power test. These observations are a bit confusing at first sight, but they can be understood if combined. A large flexible ligand molecule is of course difficult to enter a highly constrained binding pocket. Instead, it tends to bind to the target protein at a relatively open and flat binding interface. Therefore, subsets A3 and B1 actually have a significant overlap. In such a case, protein−ligand interactions are basically additive, and that is how most current scoring functions are based on. On the contrary, a small-size, relatively rigid ligand molecule normally binds in a highly restrained binding pocket, i.e., subsets A1 or B3. Due to the concave geometry of the binding pocket in such cases, protein−ligand interactions have more nonadditive features, such as a complex hydrogen bond network. Kinetic factors in protein−ligand binding may play an important role as well. Current scoring functions are less capable of handling these types of factors, which explains their less successful performance on subset A1 or B3 in this test. Our preceding analysis also explains why some scoring functions with good docking power do not necessarily have good scoring power.

When discussing the scoring power test on the entire test set, we pointed out that current scoring functions are less capable for handling polar interactions. Results obtained on subsets C1−C3 confirm this statement further. One can see in Figure 7 that most scoring functions performed better on subset C3 than on the entire test set. This subset consists of complexes with more hydrophobic binding pockets, in which the hydrophobic effect is expected to be the dominant factor in protein−ligand binding. Nonspecific hydrophobic effect is relatively easier to model than directional polar interactions. For example, solvent-accessible surface area has been adopted for modeling hydrophobic effect by some scoring functions. We propose that developing more advanced models for characterizing polar interactions is the primary concern for future scoring functions.

3.6. Comparison with the CSAR Exercise. The community structure activity resource exercise[43−46] organized by Prof. Carlson's group is also a high-impact benchmark in this field. It should be clarified to the readers how our CASF benchmark compares to the CSAR exercise. CSAR is created to provide high-quality data sets as well as appropriate metrics for evaluating scoring functions. In this aspect, it shares a similar goal with our CASF benchmark. The first CSAR exercise (CSAR-2010) assessed the performance of a panel of 19 scoring functions on 343 protein−ligand complexes selected from PDB, i.e., the CSAR-NRC HiQ data set.[43,44] The methods used in this exercise and the evaluation results can be compared to the scoring power test in our study. Recently, the CSAR exercise has expanded to chase a more mixed goal. The latest CSAR-2011/2012 exercise[45,46] employed 647 compounds for six protein targets plus 82 selected crystal complex structures in PDB. This exercise was conducted to test a total of 38 docking/scoring schemes in terms of prediction of ligand binding poses, enrichment/discriminating true binders from negatives, and relative ranking of congeneric series of compounds. Thus, the CSAR-2011/2012 exercise was not designed for evaluating scoring functions alone. It should be compared to other comparative studies of docking/scoring schemes but not our study.

In terms of organization, the CSAR exercise is basically a centralized contest. For example, a large portion of the complex structures and binding data used in the CSAR-2011/2012 exercise were collected from the pharmaceutical industry. The participants were required to make blind predictions and submit their results. Then, their results were evaluated by the CSAR team. In contrast, our CASF benchmark adopts a "server−client" structure. At the server end, we are responsible for compiling data sets, developing evaluation methods, and testing some standard scoring functions. Once our job is completed, the whole package is released to the public. Then, the users conduct their own studies at the client end by applying our data sets and evaluation methods. As long as all studies are performed on the same ground, their results can be compared to ours or among themselves. With this strategy, it is more convenient for us to update the server end, while more researchers can get involved at the client end.

This server−client strategy also decides our choice of scoring functions to be tested by the server end. Most scoring functions tested in our CASF-2013 benchmark are implemented in several popular software for drug design, including SYBYL, Discovery Studio, MOE, Schrödinger, and GOLD. Many researchers have access to these software, and thus the outcomes of our benchmark are in their direct interests. But there are certainly more scoring functions in this world, many of which are also publicly available from various resources. It is beyond our capability to collect and test all of them. It is not a good idea either if we consider some of them selectively while neglecting the others. Therefore, we chose to consider only the scoring functions implemented in main-stream software in our own study. No scoring function released by academic groups, except for X-Score, is considered by us any more. Alternatively, we will provide the complete data sets on our PDBbind-CN Web site (http://www.pdbbind-cn.org/) once this study is published so other researchers can assess them. In fact, we have started this practice since CASF-2007.[45] A number of researchers have applied our data sets and methods in their own evaluation of scoring functions since then.[68−92] Their studies have greatly expanded the scope of our benchmark in a collective manner. In turn, the whole community are able to evaluate more scoring functions on an objective basis.

4. CONCLUSION

Our CASF benchmark is created to provide an objective evaluation of scoring functions. Contributions of this benchmark lie in two aspects: First, it explores and demonstrates the appropriate metrics for scoring function evaluation. High-quality data sets are also compiled during this process. Second, it provides practical guidance for scoring function users to make smart choices among available methods. It also elucidates the common weakness in current scoring functions for future improvements. As compared to our previous CASF-2007 study, this study covers more scoring functions in main-stream software. An updated test set with a higher standard of quality is employed. This data set happens to be the same size (195 protein−ligand complexes) as the one used in CASF-2007, but its contents overlap with the previous one by only 13%. A major expansion to our evaluation methods is the introduction of the screening power test in addition to the original scoring power, ranking power, and docking power tests. With this expansion, our evaluation methods now provide a complete coverage of typical applications of scoring functions.

With regard to the evaluation results, the performances of these scoring functions are generally less promising in the scoring/ranking power tests than in the docking/screening power tests. They produced correlation coefficients (R) between 0.22 and 0.61 in the scoring power test and success

rates between 35% and 58% in the ranking power test. However, it should be mentioned that if a small number of significant outliers in the test set were removed, their statistical data would improve considerably. Interestingly, at least half of the scoring functions were outperformed by the näive scoring function ΔSAS in the scoring/ranking tests. In the docking power test, two-thirds of these scoring functions were able to produce success rates over 60% when only the best-scored binding pose was considered. A few scoring functions produced success rates even around 80%. In the screening power test, about one-third of these scoring functions produced enrichment factors over 10 at the top 1% level. The best-scoring function in this test is GlidScore-SP, which produced an enrichment factor of 19.5 or a success rate of 60% for finding the best ligand molecule.

Notably, top-ranked scoring functions in the scoring power test, such as X-Score^{HM}, ChemScore@SYBYL, ChemPLP@GOLD, and PLP@DS, are also top-ranked in the ranking power test. Top-ranked scoring functions in the docking power test, such as ChemPLP@GOLD, Chemscore@GOLD, GlidScore-SP, LigScore@DS, and PLP@DS, are also top-ranked in the screening power test. Scoring functions with good docking/screening power do not necessarily have good scoring/ranking power, such as GlideScore-SP, and vice versa, such as X-Score^{HM}. Thus, it is reasonable to classify current scoring functions as docking functions and scoring functions as some researchers did in the literature. However, some scoring functions, such as ChemPLP@GOLD and PLP@DS, demonstrate balanced docking/screening power and scoring/ranking power. Moreover, the top-ranked scoring functions in the scoring/ranking/docking power tests in CASF-2013 are essentially the same as those identified in CASF-2007. Considering that 170 out of the 195 complexes in the current test set are newly introduced, it indicates that the evaluation results derived from our benchmark are robust.

The results of docking power test obtained on subsets reveal that it is easier for a scoring function to identify the correct ligand binding pose in a highly restrained, more hydrophilic binding pocket than a relatively open, more hydrophobic binding pocket. However, the results of the scoring power test obtained on subsets reveal the opposite trend: better correlations between experimental binding data and computed binding scores are observed for protein−ligand complexes with a relatively open and more hydrophobic binding pocket. Obviously, considerations of binding pocket geometry and specific polar interactions are helpful for predicting ligand binding poses. But the contributions of these factors to ligand binding affinity are difficult to model quantitatively. Based on the results obtained on the entire test set as well as subsets, we conclude that the real challenge in protein−ligand binding affinity prediction does not lie in the nonpolar aspect but in the polar interactions as well as the desolvation effect associated with them. Moreover, the nonadditive features observed among high-affinity protein−ligand complexes should also be addressed properly by future scoring functions.

Our future CASF benchmarks will employ even better data sets and evaluation methods and strive to cover more standard scoring functions. We will also carry on with the server−client strategy so that our benchmark is fully accessible to other researchers. We hope that our efforts, together with the efforts from other researchers using our benchmark, will accelerate the birth of a new generation of scoring functions.

■ ASSOCIATED CONTENT

Supporting Information

Tables listing a summary of other comparative studies of docking/scoring methods, basic information on the test set used in CASF-2013, original and modified coefficients used in X-Score, complete results of scoring, ranking, docking, and screening power tests, correlations between experimental binding constants and binding scores, and enrichment factors in the screening power test, text providing descriptions of the scoring functions under study and accompanying references, and figures showing the definition of the buried solvent-accessible surface area of the ligand molecule and complete results of the docking power test. This material is available free of charge via the Internet at http://pubs.acs.org. All of the data sets used in this study, including experimental binding data, processed protein−ligand complex structures, and decoy ligand binding poses, will be posted on the PDBbind-CN Web site (http://www.pdbbind-cn.org/) for free download.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: wangrx@mail.sioc.ac.cn.

Notes

The authors declare no competing financial interests.

■ ACKNOWLEDGMENTS

This work is supported by the MSD China Postdoctoral Research Fellowship to Dr. Yan Li. We are also grateful for the financial support from the Chinese National Natural Science Foundation (Grant Nos. 81172984, 21072213, 21002117, 21102168, and 21102165), the Chinese Ministry of Science and Technology (863 High-Tech Grant No. 2012AA020308), the Science and Technology Development Fund of Macao SAR (Grant No. 0330), and the Chinese Academy of Sciences.

■ REFERENCES

(1) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.

(2) Ewing, T. J. A.; 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.

(3) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.

(4) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(5) 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* (7), 1739−1749.

(6) 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.

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

(8) 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.

(9) Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11−26.
(10) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(11) Velec, H. F. G.; 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.
(12) Kuntz, I. D. Structure-based strategies for drug design and discovery. *Science* **1992**, *257*, 1078−1082.
(13) Babine, R. E.; Bender, S. L. Molecular recognition of protein−ligand complexes: Applications to drug design. *Chem. Rev.* **1997**, *97*, 1359−1472.
(14) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303*, 1813−1818.
(15) Muegge, I. and Rarey, M. Small molecule docking and scoring. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: Hoboken, NJ, USA, 2001; Vol. *17*, pp 1−60.
(16) Böhm, H. J.; Stahl, M. The use of scoring functions in drug discovery applications. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: Hoboken, NJ, USA, 2002; Vol. *18*, pp 41−88.
(17) Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335−373.
(18) Schulz-Gasch, T.; Stahl, M. Scoring functions for protein-ligand interactions: A critical perspective. *Drug Discovery Today* **2004**, *1*, 231−239.
(19) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935−949.
(20) 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.
(21) Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.
(22) Bursulaya, B.; Totrov, M.; Abagyan, R.; Brooks, C. Comparative Study of Several Algorithms for Flexible Ligand Docking. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 755−763.
(23) Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On evaluating molecular-docking methods for pose prediction and enrichment factors. *J. Chem. Inf. Model.* **2006**, *46*, 401−415.
(24) 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.
(25) 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. Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.
(26) Xing, L.; Hodgkin, E.; Liu, Q.; Sedlock, D. Evaluation and Application of Multiple Scoring Functions for a Virtual Screening Experiment. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 333−344.
(27) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy. *Proteins* **2004**, *57*, 225−242.
(28) Perola, E.; Walters, W. P.; Charifson, P. S. A Detailed Comparison of Current Docking and Scoring Methods on Systems of Pharmaceutical Relevance. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 235−249.
(29) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *J. Med. Chem.* **2004**, *47*, 558−565.
(30) Kontoyianni, M.; Sokol, G. S.; MCclellan, L. M. Evaluation of Library Ranking Efficacy in Virtual Screening. *J. Comput. Chem.* **2005**, *26*, 11−22.
(31) Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. Comparison of Automated Docking Programs as Virtual Screening Tools. *J. Med. Chem.* **2005**, *48*, 962−976.
(32) Zhou, Z.; Felts, A. K.; Friesner, R. A.; Levy, R. M. Comparative Performance of Several Flexible Docking Programs and Scoring Functions: Enrichment Studies for a Diverse Set of Pharmaceutically Relevant Targets. *J. Chem. Inf. Model.* **2007**, *47*, 1599−1608.
(33) Kim, R.; Skolnick, J. Assessment of programs for ligand binding affinity prediction. *J. Comput. Chem.* **2008**, *29*, 1316−1331.
(34) Li, Y.; Liu, Z. H.; Han, L.; Li, J.; Liu, J.; Zhao, Z. X.; Wang, R. X. Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set. *J. Chem. Inf. Model.* **2014**, DOI: 10.021/ci500080q, (companion paper in this issue).
(35) Wang, R.; Lu, Y.; Wang, S. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.
(36) Marsden, P. M.; Puvanendrampillai, D.; Mitchell, J. B. O.; Glen, R. C. Predicting Protein-Ligand Binding Affinities: A Low Scoring Game? *Org. Biomol. Chem.* **2004**, *2*, 3267−3273.
(37) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. Assessing Scoring Functions for Protein−Ligand Interactions. *J. Med. Chem.* **2004**, *47*, 3032−3047.
(38) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47* (12), 2977−2980.
(39) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.
(40) 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.
(41) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.
(42) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582−6594.
(43) 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.
(44) 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.
(45) Damm-Ganamet, K. L.; Smith, R. D.; Dunbar, J. B.; Stuckey, J. A.; Carlson, H. A. CSAR Benchmark Exercise 2011−2012: Evaluation of Results from Docking and Relative Ranking of Blinded Congeneric Series. *J. Chem. Inf. Model.* **2013**, *53*, 1853−1870.
(46) Dunbar, J. B.; 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**, *53*, 1842−1852.
(47) Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M. LigScore: A Novel Scoring Function for Predicting Binding Affinities. *J. Mol. Graphics Modell.* **2005**, *23*, 395−407.
(48) Verkhivker, G.; Appelt, K.; Freer, S. T.; Villafranca, J. E. Empirical free energy calculations of ligand-protein crystallographic complexes. I. Knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Eng.* **1995**, *8*, 677−691.
(49) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Rose, P. W. Deciphering common failures in molecular docking of ligand-protein complexes. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 731−751.

(50) 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.
(51) Muegge, I. A Knowledge-Based Scoring Function for Protein-Ligand Interactions: Probing the Reference State. *Perspect. Drug Discovery Des.* **2000**, *20*, 99−114.
(52) Muegge, I. Effect of Ligand Volume Correction on PMF Scoring. *J. Comput. Chem.* **2001**, *22*, 418−425.
(53) Muegge, I. PMF Scoring Revisited. *J. Med. Chem.* **2006**, *49*, 5895−5902.
(54) Jain, A. N. Scoring Noncovalent Protein-Ligand Interactions: A Continuous Differentiable Function Tuned to Compute Binding Affinities. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 427−440.
(55) Böhm, H.-J. Prediction of Binding Constants of Protein Ligands: A Fast Method for the Prioritization of Hits Obtained from De Novo Design or 3D Database Search Programs. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309−323.
(56) Böhm, H.-J. The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant for a Protein-Ligand Complex of Known Three-Dimensional Structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.
(57) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical Scoring Functions: I. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.
(58) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible Docking Using Tabu Search and an Empirical Estimate of Binding Affinity. *Proteins: Struct., Funct., Genet.* **1998**, *33*, 367−382.
(59) Korb, O.; Stutzle, T.; Exner, T. E. Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49*, 84−96.
(60) Mooij, W. T. M.; Verdonk, M. L. General and Targeted Statistical Potentials for Protein-Ligand Interactions. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 272−287.
(61) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. *J. Med. Chem.* **2006**, *49*, 6177−6196.
(62) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.
(63) Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L. Computation of Octanol−Water Partition Coefficients by Guiding an Additive Model with Knowledge. *J. Chem. Inf. Model.* **2007**, *47*, 2140−2148.
(64) De Matos, P.; Alcantara, R.; Dekker, A.; Ennis, M.; Hastings, J.; Haug, K.; Spiteri, I.; Turner, S.; Steinbeck, C. Chemical entities of biological interest: An update. *Nucleic Acids Res.* **2009**, D249−D254.
(65) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(66) Tao, P.; Wang, R. X.; Lai, L. H. Calculating partition coefficients of peptides by the addition method. *J. Mol. Model.* **1999**, *5*, 189−195.
(67) Irwin, J. J.; Shoichet, B. K. ZINC: A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(68) Shin, W.-H.; Kim, J.-W.; Kim, D.-S.; Seok, C. GalaxyDock2: Protein-Ligand Docking Using Beta-Complex and Global Optimization. *J. Comput. Chem.* **2013**, *34*, 2647−2656.
(69) Liu, Q.; Kwoh, C. K.; Li, J. Binding Affinity Prediction for Protein−Ligand Complexes Based on β Contacts and B Factor. *J. Chem. Inf. Model.* **2013**, *53*, 3076−3085.
(70) Liu, Y.; Xu, Z.; Yang, Z.; Chen, K.; Zhu, W. A knowledge-based halogen bonding scoring function for predicting protein-ligand interactions. *J. Mol. Model.* **2013**, *19*, 5015−5030.
(71) Zilian, D.; Sotriffer, C. A. SFCscore^{RF}: A random forest-based scoring function for improved affinity prediction of protein-ligand complexes. *J. Chem. Inf. Model.* **2013**, *53*, 1923−1933.
(72) Wang, S.-H.; Wu, Y.-T.; Kuo, S.-C.; Yu, J. HotLig: A molecular surface-directed approach to scoring protein−ligand interactions. *J. Chem. Inf. Model.* **2013**, *53*, 2181−2195.
(73) Li, G.-B.; Yang, L.-L.; Wang, W.-J.; Li, L.-L.; Yang, S.-Y. ID-Score: A new empirical scoring function based on a comprehensive set of descriptors related to protein−ligand interactions. *J. Chem. Inf. Model.* **2013**, *53*, 592−600.
(74) Schneider, N.; Lange, G.; Hindle, S.; Klein, R.; Rarey, M. A consistent description of Hydrogen bond and Dehydration energies in protein-ligand complexes: Methods behind the HYDE scoring function. *J. Comput.-Aided. Mol. Des.* **2013**, *27*, 15−29.
(75) Korb, O.; Ten Brink, T.; Victor Paul Raj, F. R.; Keil, M.; Exner, T. E. Are predefined decoy sets of ligand poses able to quantify scoring function accuracy? *J. Comput.-Aided. Mol. Des.* **2012**, *26*, 185−197.
(76) Hsieh, J.; Yin, S.; Wang, X. S.; Liu, S.; Dokholyan, N. V.; Tropsha, A. Cheminformatics Meets Molecular Mechanics: A Combined Application of Knowledge-Based Pose Scoring and Physical Force Field-Based Hit Scoring Functions Improves the Accuracy of Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2012**, *52*, 16−28.
(77) Wang, J.-C.; Lin, J.-H.; Chen, C.-M.; Perryman, A. L.; Olson, A. J. Robust Scoring Functions for Protein−Ligand Interactions with Quantum Chemical Charge Models. *J. Chem. Inf. Model.* **2011**, *51*, 2528−2537.
(78) Neudert, G.; Klebe, G. DSX: A Knowledge-Based Scoring Function for the Assessment of Protein−Ligand Complexes. *J. Chem. Inf. Model.* **2011**, *51*, 2731−2745.
(79) Hsieh, J.-H.; Yin, S.; Liu, S.; Sedykh, A.; Dokholyan, N. V.; Tropsha, A. Combined Application of Cheminformatics- and Physical Force Field-Based Scoring Functions Improves Binding Affinity Prediction for CSAR Data Sets. *J. Chem. Inf. Model.* **2011**, *51*, 2027−2035.
(80) Osolodkin, D. I.; Palyulin, V. A.; Zefirov, N. S. Structure-Based Virtual Screening of Glycogen Synthase Kinase 3 beta Inhibitors: Analysis of Scoring Functions Applied to Large True Actives and Decoy Sets. *Chem. Biol. Drug Des.* **2011**, *78*, 378−390.
(81) Ballester, P. J.; Mitchell, J. B. Comments on "Leave-Cluster-Out Cross-Validation Is Appropriate for Scoring Functions Derived from Diverse Protein Data Sets": Significance for the Validation of Scoring Functions. *J. Chem. Inf. Model.* **2011**, *51*, 1739−1741.
(82) Spitzmueller, A.; Velec, H. F. G.; Klebe, G. MiniMuDS: A New Optimizer using Knowledge-Based Potentials Improves Scoring of Docking Solutions. *J. Chem. Inf. Model.* **2011**, *51*, 1423−1430.
(83) Kramer, C.; Gedeck, P. Global Free Energy Scoring Functions Based on Distance-Dependent Atom-Type Pair Descriptors. *J. Chem. Inf. Model.* **2011**, *51*, 707−720.
(84) Plewczynski, D.; Lazniewski, M.; von Grotthuss, M. VoteDock: Consensus Docking Method for Prediction of Protein−Ligand Interactions. *J. Chem. Inf. Model.* **2011**, *32*, 568−581.
(85) Plewczynski, D.; Lazniewski, M.; Augustyniak, R.; Ginalski, K. Can We Trust Docking Results? Evaluation of Seven Commonly Used Programs on PDBbind Database. *J. Comput. Chem.* **2011**, *32*, 742−755.
(86) Tang, Y. T.; Marshall, G. R. PHOENIX: A Scoring Function for Affinity Prediction Derived Using High-Resolution Crystal Structures and Calorimetry Measurements. *J. Chem. Inf. Model.* **2011**, *51*, 214−228.
(87) Shen, Q.; Xiong, B.; Zheng, M.; Luo, X.; Luo, C.; Liu, X.; Du, Y.; Li, J.; Zhu, W.; Shen, J.; Jiang, H. Knowledge-Based Scoring Functions in Drug Design: 2. Can the Knowledge Base Be Enriched? *J. Chem. Inf. Model.* **2011**, *51*, 386−397.

(88) Kramer, C.; Gedeck, P. Leave-Cluster-Out Cross-Validation Is Appropriate for Scoring Functions Derived from Diverse Protein Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 1961−1969.
(89) Sandor, M.; Kiss, R.; Keseru, G. M. Virtual Fragment Docking by Glide: A Validation Study on 190 Protein-Fragment Complexes. *J. Chem. Inf. Model.* **2010**, *50*, 1165−1172.
(90) Pencheva, T.; Soumana, O. S.; Pajeva, I.; Miteva, M. A. Post-docking virtual screening of diverse binding pockets: Comparative study using DOCK, AMMOS, X-Score and FRED scoring functions. *Eur. J. Med. Chem.* **2010**, *45*, 2622−2628.
(91) Ballester, P. J.; Mitchell, J. B. A machine learning approach to predicting protein−ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26*, 1169−1175.
(92) Das, S. Binding Affinity Prediction with Property-Encoded Shape Distribution Signatures. *J. Chem. Inf. Model.* **2010**, *50*, 298−308.