JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2019, 59, 895−913

pubs.acs.org/jcim

Downloaded via ELI LILLY AND CO on June 19, 2023 at 06:09:18 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

Comparative Assessment of Scoring Functions: The CASF-2016 Update

Minyi Su,[†,‡] Qifan Yang,[†,‡] Yu Du,[†,‡] Guoqin Feng,[†,‡] Zhihai Liu,[†] Yan Li,[*,†] and Renxiao Wang[*,†,‡,§]

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

[‡] University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China

[§] Shanxi Key Laboratory of Innovative Drugs for the Treatment of Serious Diseases Basing on Chronic Inflammation, College of Traditional Chinese Medicines, Shanxi University of Chinese Medicine, Taiyuan, Shanxi 030619, People's Republic of China

Supporting Information

ABSTRACT: In structure-based drug design, scoring functions are often employed to evaluate protein−ligand interactions. A variety of scoring functions have been developed so far, and thus, some objective benchmarks are desired for assessing their strength and weakness. The comparative assessment of scoring functions (CASF) benchmark developed by us provides an answer to this demand. CASF is designed as a "scoring benchmark", where the scoring process is decoupled from the docking process to depict the performance of scoring function more precisely. Here, we describe the latest update of this benchmark, i.e., CASF-2016. Each scoring function is still evaluated by four metrics, including "scoring power", "ranking power", "docking power", and "screening power". Nevertheless, the evaluation methods have been improved considerably in several aspects. A new test set is compiled, which consists of 285 protein−ligand complexes with high-quality crystal structures and reliable binding constants. A panel of 25 scoring functions are tested on CASF-2016 as a demonstration. Our results reveal that the performance of current scoring functions is more promising in terms of docking power than scoring, ranking, and screening power. Scoring power is somewhat correlated with ranking power, so are docking power and screening power. The results obtained on CASF-2016 may provide valuable guidance for the end users to make smart choices among available scoring functions. Moreover, CASF is created as an open-access benchmark so that other researchers can utilize it to test a wider range of scoring functions. The complete CASF-2016 benchmark will be released on the PDBbind-CN web server (http://www.pdbbind-cn.org/casf.asp/) once this article is published.

1. INTRODUCTION

In structure-based drug design, scoring function usually refers to a family of computational methods used for quantifying protein−ligand interaction.[1−4] Scoring functions do not attempt to encode the sophisticated physics underneath protein−ligand interaction with high-level theories. Instead, they aim at achieving a compromise between speed and accuracy by making various approximations, which makes them suitable for high-throughput tasks.[5−8] For example, scoring function can be employed in a molecular docking job to rank hundreds of putative ligand binding poses to select the favorable ones. Binding affinity of the ligand molecule can be estimated in turn based on the selected ligand binding mode. This combined docking/scoring scheme is widely applied to virtual screening for lead discovery as well as structure−activity relationship analysis for lead optimization. Scoring functions are often considered as "fast and dirty". However, some prospective tests of the computational methods for binding affinity prediction have suggested that scoring functions are not always less accurate in practice than the more demanding physics-based methods.[9,10]

Since the early 1990s, development of more advanced scoring functions has remained an active field. New ideas keep emerging along the path. As a result, today's scoring functions have gradually evovled into four main categories, i.e., physics-based methods, empirical scoring functions, knowledge-based statistical potentials, and descriptor-based machine-learning scoring functions.[11] A large number of scoring functions have already been published in the literature, which are often validated on different data sets. It is desired that they can be assessed, preferably in a comparative manner, on some high-quality benchmarks. Apparently, the outcomes of such benchmarks can help the end users make a smart choice among available scoring functions. Scoring function developers also need such benchmarks for testing their new models.

Received: August 14, 2018
Published: November 27, 2018

ACS Publications © 2018 American Chemical Society

In fact, a good number of comparative studies of docking/scoring methods have been published since the early 2000s.[12−20] In those studies, scoring function was usually treated as an integrated component in a molecular docking protocol. Thus, the performance of a docking/scoring protocol was evaluated by how well they could reproduce ligand binding poses and/or rank known ligands to a target protein in molecular docking trials. Sometimes, docking/scoring methods were also tested in virtual screening trials to examine if they could recall known binders from a random pool of molecules. Various data sets were used in those comparative studies, which were mainly obtained from public resources. For example, protein−ligand complex structures are retrieved from the Protein Data Bank (http://www.rcsb.org/).[21] Information about protein−ligand interactions (such as binding data) can be retrieved from ChEMBL,[22] PubChem,[23] and BindingDB.[24] Aside from that, some researchers also attempted to organize blind tests of docking/scoring methods with unpublished data sets solicited from pharmaceutical companies, such as the Community Structure−Activity Resource (CSAR) exercises[25−28] and the D3R Grand Challenge.[9,10]

As implied above, most comparative studies of docking/scoring schemes are conducted in context of molecular docking or even virtual screening. These types of studies can be classified as "docking benchmarks". It is true that scoring functions are applied in couple with molecular docking programs intimately. But, can a "docking benchmark" depict the performance of a scoring function? In our opinion, the answer is NO. It is because the final outcomes of a molecular docking job are affected by many factors, and it is not quite possible to isolate the contribution of scoring function from others. For example, Gathiaka et al. commented on the outcomes of the D3R Grand Challenge 2015 as ⋯ *successful prediction of ligand−protein poses relies not only on the docking program, but also on other steps in the overall protocol, or workflow, which may include a docking program as a central component but also contains key preparatory or procedural steps.*[9] Docking benchmarks are probably welcome more by the end users because such a study may reveal the optimal docking/scoring protocol for certain target protein(s) or ligand set(s) and thus has a practical value. However, docking benchmarks are not the best solution for assessing scoring functions *per se*.

We have been striving to establish a "scoring benchmark" particularly for scoring function assessment. The outcome of our efforts is the Comparative Assessment of Scoring Functions (CASF) benchmark. Key ideas underneath our benchmark include (i) decouple the scoring process from the conformational sampling process in molecular docking, (ii) design a set of performance metrics and provide quantitative evaluation methods, and (3) compile a diverse, high-quality test set for conducting the evaluations. The first published work along this path was CASF-2007.[29] A major update, i.e., CASF-2013, was published later.[30,31] All data sets employed in CASF-2007 and CASF-2013 have been released to the public at the PDBbind web portal (http://www.pdbbind-cn.org/casf.asp). It seems that our CASF benchmark has been well accepted among the scoring function community. According to our literature survey, over 40 applications have already been published by other researchers since 2010, which employed CASF in full or in part for validating new scoring functions or other scoring methods (see the Supporting Information, Part S1).

In this article, we report another major update of our CASF benchmark, namely, CASF-2016. It inherits the basic framework of the previous CASF-2013 benchmark; i.e., the performance of a scoring function is evaluated in terms of scoring power, ranking power, docking power, and screening power. However, significant improvements have been made to the evaluation methods. A new, larger test set is compiled to conduct all evaluations. Besides, a total number of 25 scoring functions are tested as demonstration. We expect that CASF-2016 will remain as a major benchmark for scoring function assessment over the next few years. In the following sections, we describe the evaluation methods used in CASF-2016 and also discuss the evaluation results for the scoring functions being tested.

2. MATERIAL AND METHODS

2.1. Compilation of the Primary Test Set. All performance tests enabled in CASF-2016 were based on a set of protein−ligand complexes with high-quality crystal structures and reliable binding data. This test set was selected from the PDBbind refined set (version 2016). The PDBbind database provides a comprehensive collection of experimental binding data, which are all curated from the original references, for several major categories of biomolecular complexes in the PDB.[32,33] Note that the PDBbind refined set itself was selected from the entire PDBbind database (over 13,300 protein−ligand complexes) by removing the complexes with obvious problems in structure or binding data or other unwanted features. For example, a qualified complex must have a crystal structure with an overall resolution < 2.5 Å and an R-factor < 0.25. Its binding data must be either K_d or Ki, and it must be a noncovalent binary complex, and so on. Detailed descriptions of the rules for selecting the refined set can be found in the Supporting Information (Part S2).

The 4057 protein−ligand complexes included in the PDBbind refined set (v.2016) were then subjected to the following procedure to select the primary test set used in CASF-2016:

(1) All protein−ligand complexes were grouped into clusters by protein sequence similarity. Here, sequence similarity was computed with the CD-hit program (version 4.0) obtained from PDB.[34] The similarity cutoff used in clustering was 90%. Each resulting cluster typically contained the complexes formed by the same protein molecule. This step was the same as CASF-2013.

(2) Only the clusters containing more than five members were considered. In each remaining cluster, five representative complexes were selected by their binding affinities, including the one with the highest binding affinity ($K_{a,\ max}$), the one with the lowest binding affinity ($K_{a,min}$), and three additional complexes. Here, the difference between $K_{a,max}$ and $K_{a,min}$ must be at least 100-fold (i.e., two logarithm units) to create a fairly wide binding affinity range. Binding affinities of the other three complexes should distribute as evenly as possible between the two extremes. The binding constants between any two complexes must differ by at least one-fold (i.e., log 2) because this is close to the estimated intrinsic error of "heterogeneric" binding data.[35] This step was very different from CASF-2013.

(3) The electron density map of each selected complex structure was inspected visually to examine its quality. For this purpose, structural factor data were downloaded from PDB, and the electron density map of each complex structure was displayed and analyzed by using the COOT

software (version 0.7).[36] Our inspection focused on the ligand binding pose and the nearby pocket residues. A complex structure was rejected if it had any of the following defects: (i) Structure factor data for this complex structure were not available from PDB. (ii) Electron density was missing for a major part of the ligand structure or nearby pocket residues or the overall fitting of the ligand to the density map was rather poor. (iii) There were a considerable number of positive and negative density regions around the ligand molecule. (iv) The density map was fit equally well by two alternative ligand binding poses. If a certain protein−ligand complex within a cluster was rejected at this step, it was replaced by another qualified candidate. This step was essentially the same as CASF-2013.

(4) The ligand molecules in all selected complexes were inspected to ensure that no ligand molecule was identical or a stereoisomer (e.g., Z/E or R/S) to another ligand molecule in the test set. This examination guaranteed that every ligand included in the final test set was a unique chemical entity. It also reduced the number of cross-binders (i.e., the same ligand binding to multiple target proteins) included in the test set. This step was a new development in CASF-2016.

Through the above procedure, a set of 285 protein−ligand complexes in 57 clusters was selected. We called it the PDBbind core set (v.2016), which served as the primary test set in CASF-2016. Basic information on this data set is summarized in the Supporting Information (Part S3).

2.2. Preparation of Structural Files. *2.2.1. Processing the Protein−Ligand Complex Structures.* Structures of all 285 complexes in the primary test set were downloaded from the RCSB PDB Web site (http://www.rcsb.org/pdb/). The original PDB structural files were further processed so that they could be utilized by the molecular modeling software applied in this study. In brief, a "biological assembly" of each complex structure was split into a protein molecule and a ligand molecule. Atom types and bond types of the ligand molecule were automatically assigned according to the definitions of the Tripos force field[37] with the I-interpret program developed in our group[38] and then manually corrected if necessary. Hydrogen atoms were added to the protein and the ligand with the SYBYL software (version 8.1, CERTARA Inc.) by following a simple protonation scheme assuming a neutral pH condition; that is, all carboxylic and phosphonate groups were deprotonated, while all aliphatic amine, guanidine, amino groups were protonated. During this process, the AMBER FF99 partial charges were assigned to the protein, while MMFF94 partial charges were assigned to the ligand. All water molecules were removed from the complex structure for the sake of convenience. Nevertheless, metal ions inside the binding pocket, which are often an integrated component of the protein structure and indispensable for its biological function, were retained with the protein structure. Finally, the processed protein structure was saved in a PDB-format file, and the processed ligand structure was saved in a Mol2-format file and a SDF-format file. During the above process, no structural optimization was conducted on either the protein or the ligand to keep their original coordinates from PDB. This step was essentially the same as CASF-2013.

Another set of complex structures was prepared by optimizing the ligand binding pose in each complex structure. This treatment was intended to resolve the severe steric clashes between the ligand and the protein remaining in the crystal structure. This set of optimized complex structures was used in the scoring power and ranking power test in CASF-2016. Structural optimization was conducted by using the AMBER software (version 2012).[39] The AMBER FF12SB force field was applied to the protein, while the AMBER GAFF force field was applied to the ligand. Partial charges on the ligand were computed by the AM1-BCC method. Optimization of the ligand structure was performed through 1000 steepest descent steps, while the protein structure was kept fixed with a restraint force of 100 kcal/mol Å^2. Protein molecules with metal ions needed to be handled properly by loading the special template library *atomic_ions.lib* and the parameter file *frcmod.ionslcm_cm_tip3p*. The final optimized ligand structure was saved in a Mol2-format file.

2.2.2. Preparing the Decoy Ligand Binding Poses. Decoy ligand binding poses are required in our CASF benchmark for a docking power test and screening power test. For a given protein−ligand pair, a set of decoy binding poses of the ligand molecule (which are referred to as the "decoy set" in this article) were generated in prior. To achieve the maximal conformational diversity in a decoy set, three popular molecular docking softwares, including GOLD (version 5.2, Cambridge Crystallographic Data Center), Surflex implemented in the SYBYL software (version 8.1, CERTARA Inc.), and the molecular docking module implemented in the MOE software (version 2015, Chemical Computing Group), were employed, and their outputs were combined to select the final decoy set. Note that the same three docking softwares were also employed for generating the decoy sets used in CASF-2013. But the technical procedure and parameters for running those softwares have been modified in this study. Details for this step are given in the Supporting Information (Part S4).

In brief, in order to select the decoy sets used in docking power test, the outputs of those three docking software were combined to give ∼1000 binding poses. The poses with RMSD values lower than 10 Å were kept. Here, all RMSD values were computed by using the ligand binding pose observed in the crystal complex structure (i.e., the native binding pose) as the reference. The algorithm developed by Allen et al.[40] was adopted for RMSD computation, which was a new feature of this study. This algorithm utilizes the Hungarian algorithm to find the optimal match between two sets of atoms. In theory, it can correct the errors introduced by the standard RMSD algorithm for the molecules with symmetric chemical structures. Then, those binding poses were divided into 10 bins by their RMSD values (0−10 Å) with a bin width of 1 Å. The binding poses in each bin were further grouped into 10 clusters by their conformational similarity. The binding pose with the lowest strain energy in each cluster was selected as the representative of that cluster. This clustering procedure was the same as CASF-2013. However, a new feature here was that all strain energies were computed with the OPLS3 force field[41,42] implemented in the Schrödinger software (version 2016), which was supposed to provide more accurate strain energies for small molecules. The above procedure was desired to select up to 10 × 10 = 100 representative ligand binding poses that evenly distributed over the binding site. However, depending on the shape and size of the binding site (or the ligand molecule), the decoy set actually contained ligand binding poses fewer than 100 for many complexes. The size of the final decoy set for each complex in our test set is given in the Supporting Information (Part S3).

The decoy sets used in the screening power test were generated through a similar procedure. For each of 57 target proteins included in the test set, all 285 ligands were docked into its binding site. Here, the protein structure in the highest-affinity complex in each cluster was chosen for this docking job. Due to the sheer number of possible protein−ligand pairs (57 × 285 = 16,245), the method for generating the decoy set for each protein−ligand pair must be simplified. Thus, the total number of ligand binding poses generated by the three docking software was reduced to 500. Those 500 binding poses were divided into 100 clusters by their conformational similarity (as measured by RMSD values) by the K-means clustering algorithm. The ligand binding pose with the lowest strain energy in each cluster was selected as the representative. As result, the decoy set for each protein−ligand pair contained equally a total of 100 ligand binding poses.

2.3. Basic Evaluation Methods. In CASF-2016, the performance of a scoring function is still evaluated by four metrics, i.e., scoring power, ranking power, docking power, and screening power, which is the same as CASF-2013. However, the evaluation methods have been improved considerably in CASF-2016. All improvements will be indicated explicitly in the following descriptions.

2.3.1. Evaluation of the Scoring Power. "Scoring power" refers to the ability of a scoring function to produce binding scores in a linear correlation with experimental binding data. In CASF-2016, all scoring functions being tested were applied to compute the binding scores for the protein−ligand complexes in the test set based on their original crystal structures as well as the *in situ* optimized structures. Then, the Pearson's correlation coefficient (R) between the computed binding scores and the experimental binding constants (in logarithm units, log K_a) was computed as a quantitative indicator of the scoring power (eq 1). The standard deviation (SD) in linear regression was also recorded as an additional indicator (eq 2).

$$R = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2}\sqrt{\sum_i^n (y_i - \bar{y})^2}} \tag{1}$$

$$SD = \sqrt{\frac{\sum_i^n [y_i - (a + bx_i)]^2}{n - 1}} \tag{2}$$

In eqs 1 and 2, x_i is the binding score computed for 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, respectively. This set of evaluation method is the same as CASF-2013.

2.3.2. Evaluation of the Ranking Power. "Ranking power" refers to the ability of a scoring function to correctly rank the known ligands of a certain target protein by their binding affinities when the precise binding poses of those ligands are given. Apart from scoring power, ranking power does not require a linear correlation between computed binding scores and experimental binding data. In addition, ranking power is applied only to the ligands for the same target protein. In the new test set used in CASF-2016, each complex cluster was increased to contain five members, which allowed more robust statistical methods to be used. Thus, the methods for evaluating ranking power have been completely revised in CASF-2016.

The Spearman's rank correlation coefficient (ρ), Kendall's rank correlation coefficient (τ), and the Predictive Index (PI) were used as the quantitative indicators of ranking power. The Spearman's rank correlation coefficient is computed as[43]

$$\rho = \frac{\sum_i^n (rx_i - \overline{rx})(y_i - \overline{ry})}{\sqrt{\sum_i^n (x_i - \overline{rx})^2}\sqrt{\sum_i^n (y_i - \overline{ry})^2}} \tag{3}$$

Here, rx_i is the rank of the binding score of the ith complex; ry_i is the rank of the experimental binding constant of this complex; n is the total number of samples, which is five in this case.

Kendall's rank correlation coefficient is computed as[44]

$$\tau = \frac{P_{\text{concord}} - P_{\text{discord}}}{\sqrt{(\mathrm{P}_{\text{concord}} + P_{\text{discord}} + T)(\mathrm{P}_{\text{concord}} + P_{\text{discord}} + U)}} \tag{4}$$

Here, P_{concord} and P_{discord} refer to the numbers of concordant pairs and discordant pairs, respectively. Sample i can be represented by (x_i, y_i), where x_i is its computed binding data; y_i is its experimental binding data. When comparing a pair of sample i and j, it is considered as "concordant" if the ranks of x and y are consistent (i.e., $x_i > x_j$ and $y_i > y_j$, or $x_i < x_j$ and $y_i < y_j$); otherwise, it is considered as "discordant" (i.e., $x_i > x_j$ and $y_i < y_j$, or $x_i < x_j$ and $y_i > y_j$). T is the number of ties in x, while U is the number of ties in y. A pair is not counted to T and U if $x_i = x_j$ and $y_i = y_j$ occur simultaneously. The denominator in eq 4 is the total number of pair combination among all given samples; n is the total number of samples, which is five in this case.

The Predictive Index (PI) was proposed by Pearlman et al.[45] It also measures how much a scoring function can correctly rank different ligands but is designed with the concern that the ligands with significantly different binding data should be correctly ranked with a higher priority. This index is computed as

$$PI = \frac{\sum_{j>i}^n \sum_i^n W_{ij}C_{ij}}{\sum_{j>i}^n \sum_i^n W_{ij}} \tag{5}$$

Here, W_{ij} is the gap between the binding data of ligand i and j, i.e., W_{ij} = abs (E_j − E_i). C_{ij} indicates if the rank orders of experimental binding data (E_i and E_j) and predicted binding scores (P_i and P_j) are consistent. That is to say, $C_{ij} = 1$ if $(E_j - E_i)/(P_j - P_i) > 0$, $C_{ij} = -1$ if $(E_j - E_i)/(P_j - P_i) < 0$, and $C_{ij} = 0$ if $(E_j - E_i)/(P_j - P_i) = 0$.

All three indicators described above range from −1 to +1, where +1 indicates a perfect ranking and −1 indicates a total reverse ranking.

2.3.3. Evaluation of the Docking Power. "Docking power" refers to the ability of a scoring function to identify the native ligand binding pose among computer-generated decoys. Ideally, the native binding pose should be identified as the top-ranked one. As described in Section 2.2.2, a decoy set containing up to 100 ligand binding poses was generated for each protein−ligand complex. The native ligand binding pose was added to the corresponding decoy set to ensure that this set contained at least one correct binding pose. Then, the scoring function being tested was applied to score the decoy set of each complex. The RMSD value between the native binding pose and the best-scored binding pose selected by this scoring function was computed with the Hungarian algorithm developed by Allen et al.[40] If the RMSD difference was below a preset cutoff (e.g., 2.0 Å), this complex was marked as a successful case for the given scoring function. This analysis was conducted over the entire test set, and then, an overall success rate was counted as a quantitative indicator of the docking power. The evaluation method above was essentially the same as CASF-2013. But all evaluations were conducted on new decoy sets which were prepared through a more elaborate procedure.

A new development was added to the docking power test in CASF-2016, which we called "binding funnel analysis". The idea is that the native ligand binding pose should correspond to the lowest minimum on a correct binding energy surface. It is desired that the binding energy surface has a funnel-like shape around the lowest minimum, where a ligand binding pose with a smaller RMSD value is associated with a lower binding energy. If so, conformational sampling in a molecular docking process will be guided more efficiently toward the final destination. In CASF-2016, the Spearman's rank correlation coefficient between the RMSD values of the ligand binding poses in a decoy set and their computed binding scores was used as a quantitative indicator of this relationship. For a scoring function being tested, the average value of such Spearman's rank correlation coefficients obtained on all protein−ligand complexes in the test set was recorded. In addition, in order to explore the range of the binding funnel, the Spearman's rank correlation coefficient for each complex was computed by considering the decoy binding poses within a certain RMSD window. A total of nine RMSD windows were considered, including [0−2 Å], [0−3 Å], [0−4 Å], [0−5 Å], [0−6 Å], [0−7 Å], [0−8 Å], [0−9 Å], and [0−10 Å].

2.3.4. Evaluation of the Screening Power. "Screening power" 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 CASF-2016, screening power was evaluated essentially in a cross-docking trial. There were 57 target proteins in the test set. Five complexes were included for each target protein, and thus, five known ligands were available for each target protein. These known ligands were taken as the positives, where the other 285 − 5 = 280 ligands were taken as the negatives. For each target protein, the scoring function being tested was applied to score all 285 ligands. As described in Section 2.2.2, a total of 100 representative ligand binding poses were prepared for each protein−ligand pair. The scoring function being tested was applied to compute all those binding poses, and the best binding score among them was recorded as the predicted binding score for this protein−ligand pair. Then, all 285 ligands were sorted in a descending order by their binding scores. Screening power of the scoring function was evaluated by how well it ranked the true binders to the top.

The first quantitative indicator of screening power was the success rate of identifying the highest-affinity binder among the 1%, 5%, or 10% top-ranked ligands over all 57 target proteins in the test set. The second indicator was the enhancement factor (EF), which is computed as follows:

$$\mathrm{EF}_{\alpha} = \frac{NTB_{\alpha}}{NTB_{\mathrm{total}} \times \alpha} \tag{8}$$

Here, NTB_{α} is the number of true binders observed among the top-ranked candidates (e.g., α = 1%, 5%, or 10%) selected by a given scoring function. NTB_{total} is the total number of true binders for the given target protein, which is typically five. Under each condition, the average enhancement factor obtained across all 57 target proteins was recorded. This set of evaluation methods are essentially the same as CASF-2013.

It should be noted that cross-binders do exist in our test set. The evaluation methods described above are based on the assumption that the other 285 − 5 = 280 ligands in the test set do not bind to the chosen target protein. However, it is possible that the same ligand molecule is able to bind to multiple target proteins in our test set. Thus, such cross-binders must be taken into account properly in the screening power test. In order to identify the cross-binders within the new test set used in CASF-2016, we have examined the binding data recorded in ChEMBL (version 21),[46] which held arguably the largest protein−ligand binding data inventory when we launched this project. A total of 53 cross-binding protein−ligand pairs were found as result, which is summarized in the Supporting Information (Part S5). Information of the binders to each target protein was recorded in a special work list, where cross-binders were combined with "solo-binders". In fact, 21 of the 57 target proteins in the test set have more than five binders due to the existence of cross-binders. In certain cases, one target protein has as many as 10 binders.

The evaluation methods described above were applicable to "forward screening", i.e., identifying potential small-molecule ligands for a chosen target protein. As a matter fact, molecular docking can also be applied to so-called "reverse screening", i.e., identifying potential target proteins for a bioactive small-molecule compound.[47] As a new feature, evaluation of "reverse screening power" is also enabled in CASF-2016. With our benchmark, it is rather straightforward to perform this type of evaluation without the need of additional material or computation. For each ligand included in the test set, all 57 target proteins were sorted in a descending order by the binding scores computed by the scoring function being tested. Then, the reverse screening power of this scoring function was measured by its success rate of identifying the known target protein for each ligand among the 1%, 5%, or 10% top-ranked candidate proteins. Note that cross-binders were also carefully considered during this process.

2.4. Analysis of the Confidence Interval. Analysis of the confidence interval in evaluation results has received more attention in recent years.[48] As a new development in CASF-2016, confidence interval analysis has been introduced into all four performance tests. The bootstrapping method[49] was chosen for this task because this type of method is recommended when only a limited number of samples are considered.[50,51] To be specific, we employed the bias-corrected and accelerated bootstrapping method (BCa)[52] because this method attempts to adjust the bias in an asymmetric bootstrapping distribution.

In brief, the bootstrap sampling in our study was conducted as follows. Random sampling of 10,000 redundant copies with replacements was conducted on the test set. Each copy was in the same size as the original test set. For example, in the scoring or docking power test, the total sample size was 285, while in the ranking or screening power test, the total sample size was 57. Then, for a scoring function being tested, its performance metrics were re-evaluated on each redundant copy of the test set with the methods described in Section 2.3. As result, an ensemble of 10,000 bootstrap samples were obtained for each performance metric, and a certain confidence interval (e.g., 90%) could be derived accordingly. Here, the bootstrap sampling and confidence interval estimation were both performed by using the R software (version 3.4.3).[53] Detailed information about this step can be found in the Supporting Information (Part S6).

In addition to the confidence interval analysis, we also employed the posthoc Friedman test with the Shaffer's method[54,55] to compare the performance metrics of any two given scoring functions to determine if the difference between them was statistically significant. Note that the performance metric distributions obtained from our bootstrap sampling were not in standard normal distributions. The nonparametric Friedman test[56,57] was chosen here because it is suitable for dealing with any type of distribution. If the p-value between the performance metric distributions of two scoring functions derived by the posthoc Friedman test was larger than 0.10, the H_0 hypothesis was considered to be true (i.e., no significance difference between the two distributions). Accordingly, those two scoring functions would receive the same rank in this performance test. All posthoc Friedman tests in our study were also conducted by using the R software.

2.5. Additional Evaluations on Subsets of Protein−Ligand Complexes. Similar to CASF-2013, we also divided the test set in CASF-2016 into some subsets, where each of them was intended to share a common physicochemical property. All scoring functions considered in CASF-2016 were re-evaluated on these three sets of subsets in terms of scoring, ranking, docking, and screening power with the methods described in Section 2.3. The evaluation results so obtained were compared to those obtained on the entire test set, which was expected to provide additional information about the performance of each scoring function on certain groups of protein−ligand complexes.

In CASF-2016, three descriptors depicting the binding pocket on target protein were considered in dividing the entire test set into subsets. The first one was the excluded volume inside the binding pocket upon ligand binding (ΔVOL, in cubic angström). This descriptor reflected the size of the binding pocket that was effectively occupied by the bound ligand. The second one was the percentage of the solvent-accessible surface area of the ligand molecule that was buried upon binding (ΔSAS, in percentage). Unlike the first descriptor, this one reflected the accessibility of the binding pocket. For example, if the ligand molecule located in a deep, highly restrained binding pocket, a high Δ*S*AS percentage would be expected. The third one was what we called the "hydrophobic scale" of the binding pocket (H-scale, in log D units). This descriptor was computed as the sum of the fragmental log D values of the amino acid residues forming the binding pocket divided by the total number of such pocket residues. This descriptor reflected whether the binding pocket was hydrophobic or hydrophilic in nature. All three descriptors were computed with in-house computer programs. The methods used for this task are described in the Supporting Information (Part S7).

The three descriptors mentioned above were computed for all protein−ligand complexes in the test set. The test set was divided into three subsets by each descriptor, respectively. For this purpose, the protein−ligand complex in each complex cluster that had the largest ΔVOL value (i.e., the largest binding pocket) was chosen as the representative of this cluster. All 57 complex clusters were ranked by the chosen descriptor computed with the representative complexes in an ascending order. Then, they were divided into three subsets with a roughly even size (i.e., ideally 19 clusters in each subset). For the sake of clarity, the three subsets divided by ΔVOL were annotated as V1, V2, and V3. The three subsets divided by ΔSAS were annotated as S1, S2, and S3, and the three subsets divided by H-scale were annotated as H1, H2, and H3. Basic information of all three sets of subsets is summarized in Table 1. Note that the descriptors and the method used for defining subsets were significantly different from those employed in CASF-2013.

Table 1. Classification of Test Sets by Three Descriptors

Descriptor used in classification	Mean value[a]	Range[a]	Number of clusters[b]	Subset symbol
Excluded volume inside the binding pocket upon ligand binding (ΔVOL, $Å^3$)	286	[232, 369]	19	V1
	418	[371, 460]	19	V2
	560	[466, 776]	19	V3
Buried percentage of the solvent-accessible area of the ligand upon binding (ΔSAS, %)	0.65	[0.50, 0.71]	19	S1
	0.78	[0.72, 0.82]	21	S2
	0.89	[0.83, 0.97]	17	S3
Hydrophobic scale of the binding pocket (H-scale, in log D units)	−0.51	[−0.80, −0.35]	19	H1
	−0.24	[−0.34, −0.15]	19	H2
	0.12	[−0.14, 0.44]	19	H3

[a] Mean value and range of the descriptor here are calculated among the 57 representative complexes. [b] Number of complex clusters included in this subset. Each cluster consists of five protein−ligand complexes formed by a common target protein.

2.6. Scoring Functions under Assessment. A total of 25 scoring functions were tested in CASF-2016 as demonstration. Basic information of those scoring functions is summarized in Table 2. Additional descriptions can be found in the Supporting Information (Part S8). Among them, 20 scoring functions were from several mainstream molecular modeling softwares, including Discovery Studio, SYBYL, Schrödinger, MOE, and GOLD. The default scoring function implemented in AutoDock Vina (version 1.1.2)[75] was also included. In addition, there were four standalone scoring functions, i.e., X-Score developed by ourselves,[76] $\Delta_{Vina}RF_{20}$ developed by Zhang's group,[77] and DrugScore2018[78] and DrugScore^{CSD}[79] developed by Gohlke's group. Here, GBVI/WSA-dG in MOE, the AutoDock Vina scoring function, $\Delta_{Vina}RF_{20}$, DrugScore2018, and DrugScore^{CSD} were not included in CASF-2013.

Two special issues need to be mentioned here. First, the same as in CASF-2013, a single descriptor (i.e., ΔSAS) was introduced as a reference model in all performance tests. This descriptor refers to the solvent-accessible surface area of the ligand molecule that is buried upon binding. It was shown in our previous study[30] that there was an obvious correlation between ΔSAS and protein−ligand binding constants. Second, some scoring functions considered in CASF-2016 have multiple variations. For example, LUDI in Discovery Studio has three variations, i.e., LUDI1, LUDI2, and LUDI3. If each variation is treated as a separate model, there are a total of 34 models in our test panel (Table 2).

3. RESULTS AND DISCUSSION

In this section, we discuss the new test set used in CASF-2016 first. Then, we describe and discuss the evaluation results of all scoring functions. The same as CASF-2013,[29] all scoring functions are still evaluated in terms of scoring power, ranking power, docking power, and screening power. Nevertheless, evaluation methods have been improved considerably in CASF-2016. In the discussion of each performance test below, the new features in CASF-2016 are emphasized. Note that our discussion focuses on the overall trends observed in the evaluation results rather than try to answer why a particular scoring function succeeded or failed in a certain test. That type of discussion relies on an in-depth analysis of that particular scoring function, which would better be provided by its own developers.

3.1. About the New Test Set. CASF-2016 uses a newly compiled test set consisting of 285 protein−ligand complexes, i.e., the PDBbind core set (version 2016). The size of this data set is ∼50% larger than the one used in CASF-2013 (285 vs

DOI: 10.1021/acs.jcim.8b00545
J. Chem. Inf. Model. 2019, 59, 895−913

Table 2. Summary of Scoring Functions Tested in CASF-2016

Scoring function	Source	Classification	Refs
Jain	Discovery Studio (version 4.1)	empirical scoring function	58
LigScore1/LigScore2		empirical scoring function	59
PMF/PMF04		knowledge-based potential	60,61
LUDI1/LUDI2/LUDI3		empirical scoring function	62,63
PLP1/PLP2		empirical scoring function	64,65
GoldScore	GOLD (version 5.2)	physics-based function	66
ChemScore		empirical scoring function	67,68
ChemPLP		empirical scoring function	69
ASP		knowledge-based potential	70
G-Score	SYBYL (version 8.1)	physics-based function	66
PMF		knowledge-based potential	60,61
D-Score		physics-based function	71
ChemScore		empirical scoring function	67,68
GlideScore-SP	Schrodinger (version 2016)	empirical scoring function	72−74
GlideScore-XP		empirical scoring function	72−74
London-dG	MOE (version 2015)	empirical scoring function	MOE user manual
ASE		empirical scoring function	
Affinity		empirical scoring function	
Alpha-HB		empirical scoring function	
GBVI-WSA-dG		physics-based function	
AutoDock Vina	AutoDock Vina (version 1.1.2)	empirical scoring function	75
X-Score (HP/HM/HS)	X-Score (version 1.3) from the author	empirical scoring function	76
$\Delta_{\mathrm{Vina}}\mathrm{RF}_{20}$	From the author	descriptor-based machine-learning model	77
DrugScore2018 DrugScore$^{\mathrm{CSD}}$	From the author	knowledge-based potential	78 79
ΔSAS	in-house software	single descriptor	N. A.

195). A larger data set is of course expected to produce more robust statistical results. This data set is designed to contain diverse protein and ligand molecules because our CASF benchmark attempts to reflect the "average performance" of scoring function. To investigate the diversity presented in this new test set, we have surveyed some key properties among its members. Binding constants of the protein−ligand complexes in this data set range from 2.07 to 11.82 in log K_a units, spanning nearly 10 orders of magnitude. In order to compare the binding data distributions of the CASF-2016 test set and the CASF-2013 test set (Figure 1), the Shapiro−Wilk normality test[80,81] confirmed that the new data set represented a normal distribution (p-value = 0.054) at the 95% confidence level, whereas the old data set did not (p-value = 0.022). We also computed molecular weight, number of hydrogen bond donor/acceptor atoms, number of rotatable bonds, and 1-octanol/water partition coefficient (log P) for the ligand molecules included in this data set (Figure 2). Generally speaking, the majority of them are "drug-like" organic molecules. A total of 206 ligand molecules (72.3%) comply with Lipinski's "rules of five".[82] Distributions of three other descriptors (ΔVOL, ΔSAS, and H-scale) are also shown in Figure 2. These three descriptors are related to the characteristics of the binding pocket on the target protein (see Section 2.5). As compared to the ligand-based properties mentioned above, these descriptors obviously spread in a wider range. They are thus suitable for subset classification, which is an important aspect of CASF-2016.

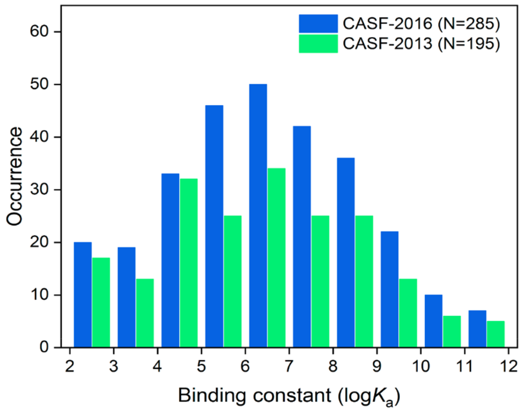

Figure 1. Distribution of the protein−ligand binding constants in the test set of CASF-2016 and the one of CASF-2013.

Reliable evaluation results must be obtained on a high-quality test set. Therefore, the protein−ligand complexes included in the CASF-2016 test set were carefully selected to have reliable crystal structures and experimental binding data. For example, the electron density map of each crystal complex structure was inspected manually to ensure that there was no obvious defect in structure fitting. We also applied the electron density score (EDIAm) proposed by Meyder et al.[83] to evaluate the complex structures. If taking Meyder's threshold of EDIAm value ≥ 0.80 for defining a high-quality electron density map, it turns out that 86% of the complex structures in the CASF-2106 test set meets this standard. Generally speaking, examining the electron density map is a more stringent method than relying on a single structural indicator (e.g., resolution or R-factor) to judge the quality of a protein−ligand complex structure.

Another new feature of the CASF-2016 test set is the increased size of each complex cluster. In the test sets used by the previous CASF benchmarks, each complex cluster was allowed to contain only three representative members. In CASF-2016, each complex cluster consists of five members. Note that larger complex clusters are, in principle, welcome by the ranking power

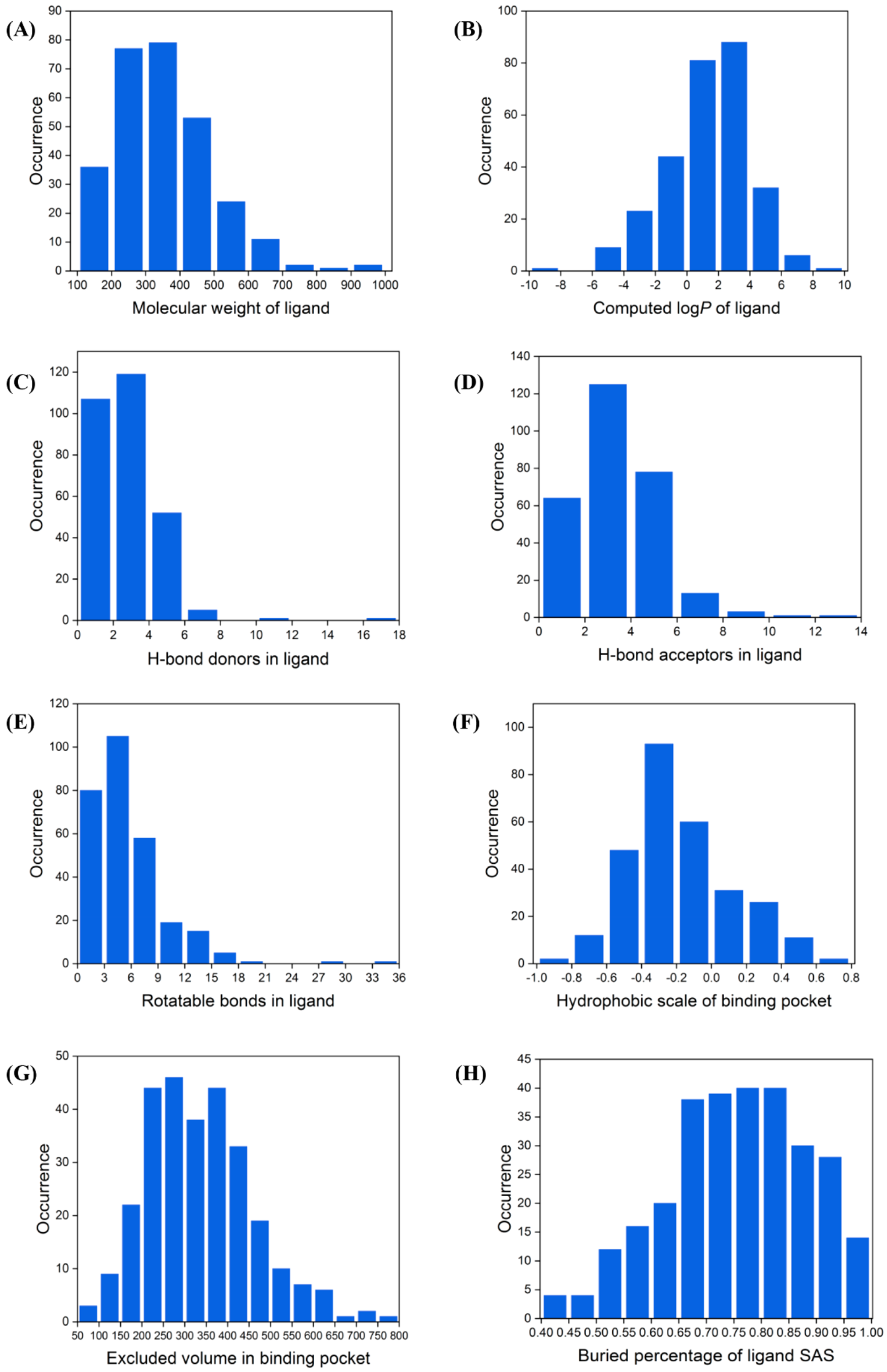

Figure 2. Distribution of some key properties of the protein−ligand complexes in the CASF-2016 test set. (A) Molecular weight of the ligand. (B) Computed log *P* value of the ligand. (C) Number of hydrogen bond acceptors in the ligand structure. (D) Number of hydrogen bond donors in the ligand structure. (E) Number of rotatable single bonds in the ligand structure. (F) Hydrophobic scale of the binding pocket (H-scale). (G) Excluded volume in the binding pocket by ligand binding (ΔVOL). (H) Buried percentage of the solvent-accessible surface of the ligand upon binding (ΔSAS).

and screening power tests. An increase from three to five does not seem to be impressive at the first sight. However, this new development actually breaks the original rigid setting; i.e., only three complexes are allowed for one cluster. Once more binding data become available, the size of each complex cluster can be increased further. Considering that the PDBbind database is undergoing a constant growth, this improvement is certainly expected for future CASF benchmarks.

3.2. Evaluation of Scoring Power. As described in Section 2.3.1, the scoring power of a scoring function is measured by the correlation between the computed binding scores and the experimental binding constants of the given protein−ligand complexes. This test is conducted with the experimentally resolved complex structures. A frequently raised question here is that in reality one normally wants to predict the binding affinity of a protein−ligand complex before its crystal structure is available, so this type of test does not have a practical value. We would like to emphasize that our CASF benchmark is designed to test scoring functions under an idealized setting to reflect their intrinsic quality. If a scoring function cannot even produce reasonable results on known structures, it would stand no chance in a realistic circumstance. In other words, the scoring power test enabled in CASF is an essential test that a useful scoring function has to pass.

Evaluation results obtained on the original crystal complex structures for all scoring functions are given in Figure 3(A), where the Pearson coefficient produced by each scoring function as well as the 90% confidence interval estimated by bootstrapping sampling are shown. The data used for making this figure are summarized in the Supporting Information (Part S9). First, one can see that most of the scoring functions being tested do not have a very promising scoring power, which produce Pearson correlation coefficients at the range of 0.21−0.63. Interestingly, the reference model (ΔSAS) actually produces a fair correlation coefficient of 0.625, while only two scoring functions ($\Delta_{\text{Vina}}\text{RF}_{20}$ and X-Score) outperform it. Here, the highlight is $\Delta_{\text{Vina}}\text{RF}_{20}$. This scoring function combines a machine-learning model with the empirical scoring function implemented in AutoDock Vina.[77] It produces the best Pearson correlation coefficient of 0.816, significantly higher than the other scoring functions. Except for $\Delta_{\text{Vina}}\text{RF}_{20}$, the relatively successful scoring functions in this test are roughly the same as those in CASF-2013, such as X-Score, ChemScore@SYBYL, ChemPLP@GOLD, and PLP1@DS. The worst scoring functions in this test are also the same as those in CASF-2013, such as London-dG@MOE, PMF@SYBYL, and PMF04@DS. Note that only 107 complexes (37%) in the CASF-2016 test set overlap with the CASF-2013 test set. The consistency obtained on two substantially different data sets indicate that our CASF benchmark produces robust evaluation results.

Besides using the original crystal complex structures, the scoring power test is also repeated on the locally optimized complex structures. Some crystal structures included in our test set still contain steric clashes between the protein molecule and the ligand molecule. Certain scoring functions are sensitive to such clashes, and therefore, a structural optimization in prior is necessary for those scoring functions to produce reasonable results. Evaluation results obtained on the locally optimized complex structures for all scoring functions are given in Figure 3(B). The data used for making this figure are summarized in the Supporting Information (Part S9). One can see that most scoring functions under our test are not affected by the minor conformational changes introduced by the structural optimization. However, as indicated by the number of "computable" complexes and the resulting correlation coefficient, the performance of GBVI/WSA-dG@MOE, GoldScore@GOLD, GlideScore-SP, and GlideScore-XP is improved more or less in this test. Indeed, the above three scoring functions all contain a steric repulsion term (see the Supporting Information, Part S8). In particular, GBVI/WSA-dG@MOE exhibits the most significant improvement with a Pearson correlation coefficient of 0.629. Based on the optimized complex structures, it becomes one of the top-ranked scoring functions in this test.

3.3. Evaluation of Ranking Power.

With the new test set in CASF-2016, the ranking power is now evaluated with classical statistical methods instead of the naïve method used in CASF-2013. As described in Section 2.3.2, ranking power of a scoring function is measured by the average Spearman correlation coefficient (ρ), Kendall correlation coefficient (τ), or the Predictive Index (PI) obtained on all 57 complex clusters. The evaluation results for all scoring functions being tested are summarized in the Supporting Information (Part S10). Those three indicators produced by each scoring function are shown in a scatter plot in Figure S2 in the Supporting Information (Part S10). A nice correlation is observed among those three indicators, indicating that they are essentially equivalent for ranking scoring functions in this test. Thus, our discussion below cites Spearman correlation coefficients only.

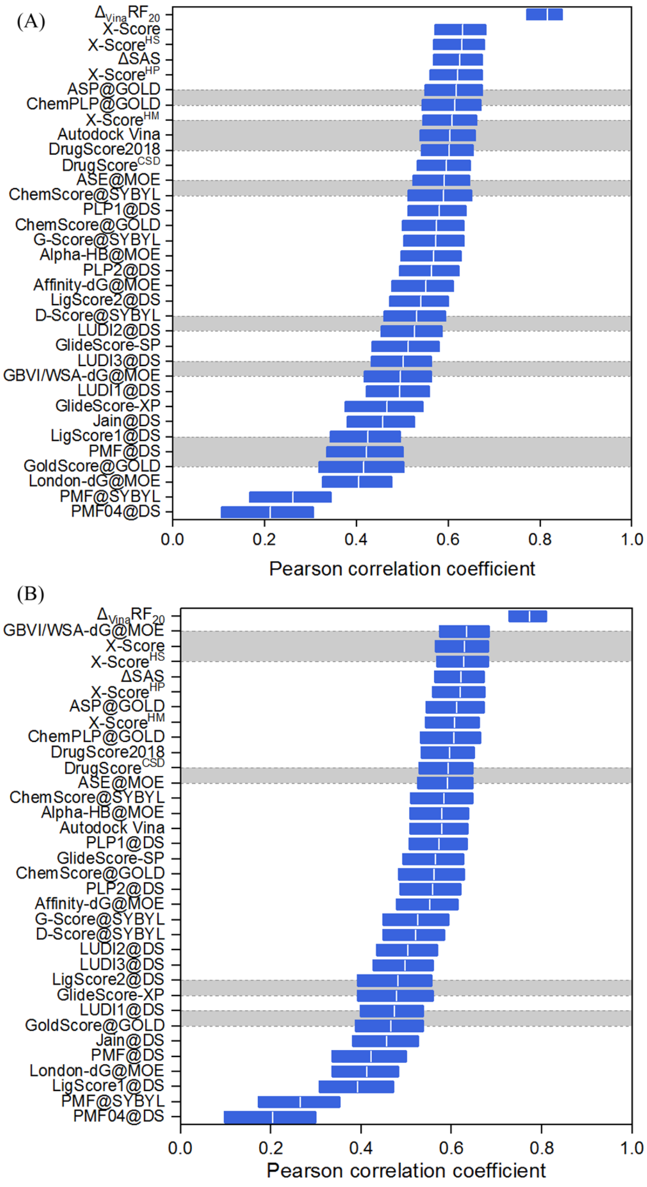

Figure 3. Pearson correlation coefficient (short white line in the middle of each blue bar) as well as the 90% confidence interval (blue bar) given by each scoring function in the scoring power test, which are obtained on (A) the original crystal complex structures and (B) the locally optimized complex structures. In both figures, scoring functions are ranked by the Pearson correlation coefficients in a descending order. The scoring functions riding on the same gray stripe are not differentiable at $\alpha = 0.10$ in the posthoc Friedman test (see Section 2.4).

The Spearman correlation coefficients as well as the 90% confidence intervals produced by each scoring function on the original crystal complex structures are shown in Figure 4(A). Generally speaking, the scoring functions being tested perform slightly better here than in the scoring power test. A notable observation is that a number of scoring functions outperform the reference model (ΔSAS, ranked #10) in this test, suggesting that at least some scoring functions are useful for predicting relative protein−ligand binding affinities. In the ranking power test in CASF-2013, the scoring functions that outperformed the reference model included ChemPLP@GOLD, X-Score, PLP2@DS, GoldScore@GOLD, ChemScore@SYBYL, and Affinity-dG@MOE. Most of them, except for GoldScore@GOLD, still stay on top here. A few more scoring functions are also relatively successful in this test, including DrugScore^{CSD}, DrugScore2018, LUDI1@DS, LUDI2@DS, and LigScore2@DS. The best performance is achieved by $\Delta_{Vina}RF_{20}$ again. It produces a Spearman correlation coefficient of 0.750, taking an obvious lead ahead of all other scoring functions. As discussed earlier, the new CASF-2016 test set include complex clusters with a larger size. Thus, the ranking power results obtained on this new data set are more reliable.

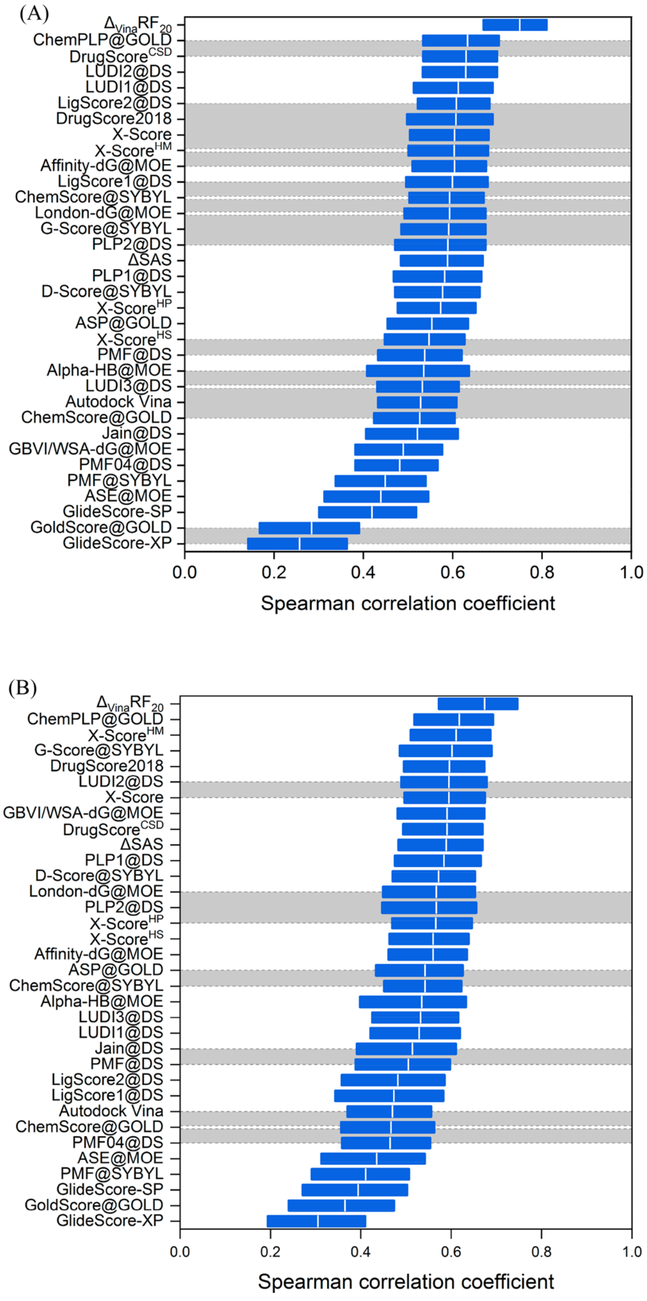

Figure 4. Average Spearman correlation coefficient (ρ) obtained on 57 target proteins by each scoring function in the ranking power test. (A) Results obtained on the original crystal complex structures. (B) Results obtained on the locally optimized complex structures. In both figures, the blue bar indicates the 90% confidence interval of the ρ value. All scoring functions are ranked by their ρ values in a descending order. The scoring functions riding on the same gray stripe are not differentiable at $\alpha = 0.10$ in the posthoc Friedman test.

Ranking power test is also repeated on the locally optimized complex structures. The Spearman correlation coefficients produced by each scoring function in this test are shown in Figure 4(B). Similar to what is observed in the scoring power test, most scoring functions are not affected much, indicating that they are not sensitive to the minor changes in ligand binding poses. Again, GBVI/WSA-dG@MOE is an exception. Its Spearman correlation coefficient increases to 0.609 on the optimized complex structures, making it one of the top-ranked scoring functions in this test. On the other hand, GlidScore-SP, GlideScore-XP, and GoldScore@GOLD, which are also sensitive to steric clashes, are still among the worst players here. They failed to compute a considerable number of complexes in the test set, which in turn lowered their average Spearman correlation coefficients obtained across all complex clusters.

Comparing the scoring power and the ranking power results obtained for the scoring functions being tested, one can see that a scoring function with a relatively good scoring power usually has a relatively good ranking power as well. However, the reverse is not necessarily true; i.e., some scoring functions with a relatively good ranking power do not have a relatively good scoring power. For example, LUDI2@DS and LUDI1@DS are ranked as #3 and #4, respectively, in the ranking power test, but they are ranked as #17 and #20, respectively, in the scoring power test. By our definition, ranking power does not require a linear correlation between the experimental binding data and the computed binding scores. Our evaluation results indicate that it is necessary to keep ranking power as a separate metric from scoring power.

3.4. Evaluation of Docking Power.

Similar to the previous CASF benchmarks, the docking power test in CASF-2016 requires that an ensemble of decoy ligand binding poses must be prepared for each complex in prior. In CASF-2016, the decoy ligand binding poses required by the docking power test have been completely regenerated with refined methods (see Section 2.2.2). For example, at the postprocessing step, the RMSD value of each ligand binding pose was computed with the newly introduced Hungarian algorithm,[40] which was able to handle symmetric molecular structures more properly than the standard RMSD algorithm. Besides, the OPLS3 force field,[41,42] which is arguably the state-of-the-art force field for dealing with small organic molecules, was employed to compute the internal strain energy of each ligand binding pose. The results of these methods have guided the selection of the final decoy ligand binding poses.

The success rates of identifying the native ligand binding poses by all scoring functions are illustrated in Figure 5. The data used for making this figure are given in the Supporting Information (Part S11). In this test, over one dozen of scoring functions produce success rates over 70% even under the

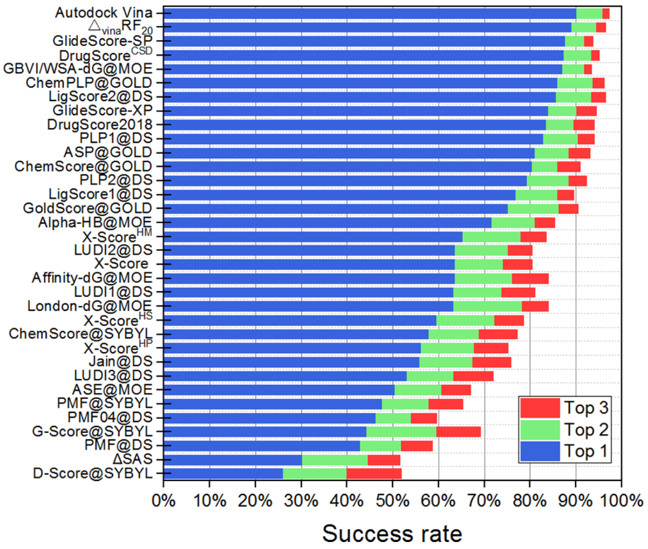

Figure 5. Success rate given by each scoring function for detecting the native ligand binding pose (RMSD < 2.0 Å) in the docking power test. Blue, green, and red bars indicate the success rates obtained by considering the top one, top two, and top three ligand binding poses, respectively. All scoring functions are ranked in a descending order by their success rates obtained at the first scenario.

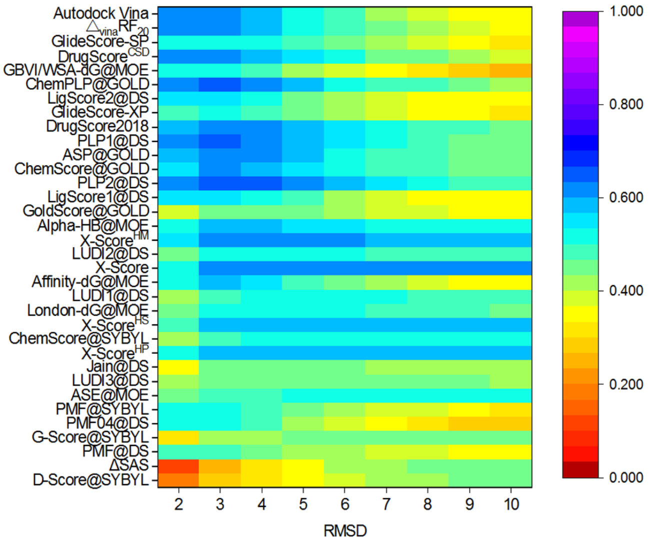

Figure 6. Results of the binding funnel analysis for each scoring function being tested. The x-axis indicates the RMSD range (e.g., [0−2 Å], [0−3 Å], and so on) where the Spearman correlation coefficients between the RMSD values and the computed binding scores are computed. Spearman correlation coefficients are indicated with a heat map ranging between 0 and 1. For the convenience of making a comparison, all scoring functions are ranked in the same order as in Figure 5, i.e., by their success rates of detecting the native ligand binding pose.

strictest criterion (i.e., only considering the best-scored binding pose). AutoDock Vina and $\Delta_{Vina}RF_{20}$ even produce success rates close to 90%. If considering the top two or top three binding poses, the success rates are generally higher by a few percent or even up to 10% for certain scoring functions. On the other hand, a few scoring functions produce relatively low success rates (e.g., < 50%). Notably, the reference model (ΔSAS) is almost the worst player here with a success rate of 30%. Thus, it is not wise to predict ligand binding pose based on protein−ligand contact surface area only.

In this test, we have also examined the success rate of each scoring function if the native ligand binding pose is not mixed into the decoy set (see the Supporting Information, Part S11). As result, the success rates of most scoring functions are lower by 1%−5% under this setting, and some scoring functions are ranked differently. Nevertheless, the overall trends discussed above remain essentially the same.

Our results demonstrate that the docking power of at least some scoring functions is promising. The success rates obtained in our test seem to be too optimistic if compared to what one would experience in reality. Note that scoring functions are tested under an idealized setting with our CASF benchmark. Here, the decoy set for each protein−ligand complex is prepared by combining the outcomes of three different docking programs, providing a fairly complete coverage of the possible ligand binding poses. This is often not possible for a molecular docking job in reality especially when a limited CPU time is assigned to the conformational sampling process. In other words, the success rate of a scoring function derived from our docking power test may be interpreted as the upper limit of its real performance in molecular docking.

In the previous CASF-2013 benchmark, the top five scoring functions in this test were ChemPLP@GOLD, GlideScore-SP, ChemScore@GOLD, PLP1@DS, and LigScore2@DS, and the worst five were D-Score@SYBYL, ΔSAS, PMF@SYBYL, PMF04@DS, and ASE@MOE. As one can see in Figure 5, all those scoring functions remain roughly the same ranks here. This again verifies that our CASF benchmark produces robust evaluation results. Among the several top-ranked scoring functions in this test, there are a physics-based model (i.e., GBVI/WSA-dG@MOE), empirical scoring function (i.e., AutoDock Vina), statistical potential (DrugScore^{CSD}), and machine-learning model ($\Delta_{vina}RF_{20}$). This interesting observation suggests that if carefully designed and calibrated, any type of scoring function can be useful at least in terms of docking power.

If comparing the evaluation results obtained from the scoring/ranking power tests and the docking power test, one can see that some relatively successful scoring functions in the scoring/ranking power tests only exhibit a modest performance in the docking power test, such as X-Score, LUDI1@DS, and LUDI2@DS. On the other hand, some not-so-good scoring functions in the scoring/ranking power tests exhibit good performance in the docking power test, such as AutoDock Vina, GlideScore-SP, and GlideScore-XP. Some scoring functions are successful in both the scoring/ranking power tests and the docking power test, such as $\Delta_{Vina}RF_{20}$, ChemPLP@GOLD, and DrugScore^{CSD}. Thus, the connection between the docking power and the scoring/ranking power is not so straightforward. It still depends on how the scoring function is designed.

If using the concept of binding energy surface to depict a protein−ligand binding process, the native ligand binding pose is assumed to locate at the lowest minimum. The success rate of identifying the native ligand binding pose discussed above reflects the "accuracy" of a scoring function for locating this particular spot on the energy binding surface. Nevertheless, it is also desired that the binding energy surface surrounding the lowest minimum forms a funnel-like shape. If so, when approaching to this region, conformational sampling will be directed to the final destination more quickly by the descending binding energies. On the contrary, a rugged binding energy surface will make conformational sampling essentially random and thus much less efficient. To address this issue, we have introduced the so-called "binding funnel analysis" into the docking power test (see Section 2.3.3), which is one of the new developments in CASF-2016. This additional analysis is intended to reflect the "efficiency" of a scoring function in molecular docking.

Results of the binding funnel analysis for all scoring functions are illustrated in Figure 6. The data used for making this figure are given in the Supporting Information (Part S11, Table S10). In order to make a convenient comparison, all scoring functions in Figure 6 are ranked in the same order as in Figure 5, i.e., by their success rates of identifying the native ligand binding pose. In Figure 6, one should see a blue region at the short RMSD range (e.g., RMSD < 5 Å) for those scoring functions creating an obvious funnel on their binding energy surface. An immediate observation here is that some scoring functions with good "accuracy" (i.e., those staying at the top of this figure) do not necessarily have good "efficiency", such as GBVI/WSA-dG@MOE, GlideScore-SP, GlideScore-XP, and ChemPLP@GOLD. On the other hand, some scoring functions with good "accuracy" also have good "efficiency", such as AutoDock Vina, $\Delta_{\mathrm{Vina}}\mathrm{RF}_{20}$, DrugScore2018, DrugScore$^{\mathrm{CSD}}$, PLP1@DS, and PLP2@DS. The latter type of scoring functions are more useful for molecular docking programs because in theory they are more efficient to reach the correct solution.

It is also interesting to notice that some scoring functions produce significant Spearman correlation coefficients at a wide RMSD range (i.e., RMSD < 10 Å), e.g., X-Score and its variations. Scoring functions with such a feature can be useful in a blind docking job where the ligand binding site on the target protein is not annotated in prior. However, those scoring functions do not always achieve a high accuracy (i.e., identifying the correct ligand binding pose) at the same time. Those scoring functions probably need refinements to produce a more sensitive conformation-energy response especially near the destination of conformational sampling. This is just one example to demonstrate how the multidimensional information offered by our CASF-2016 benchmark can be used to guide the development of better scoring functions.

3.5. Evaluation of Screening Power. As compared to CASF-2013, the screening power test in CASF-2016 is conducted on a more solid basis. First, the decoy ligand binding poses for each protein−ligand pair have been reprepared through a more elaborate workflow (see Section 2.2.2). The maximal number of decoy ligand binding poses for each protein−ligand pair used to be 50 in CASF-2013. Now, it has been increased to 100 to provide a more complete coverage of possible ligand binding poses. Second, each target protein in the test set used to have only three known ligands in CASF-2013. Now, it has been increased to five, which makes it possible to measure the screening power with classical statistical indices.

The average enhancement factors obtained on 57 target proteins by all scoring functions in the forward screening test are illustrated in Figure 7. The data used for making this figure are given in the Supporting Information (Part S12, Table S11). As described in Section 2.3.4, the forward screening trial examines how well a scoring function ranks the known ligands of a given target protein on top of other random molecules. In this test, the top five scoring functions, including ChemPLP@GOLD, $\Delta_{\mathrm{Vina}}\mathrm{RF}_{20}$, GlideScore-SP, GlideScore-XP, and ChemScore@GOLD, produce enhancement factors above eight when the top 1% candidates in screening (i.e., 285 × 1% ≈ 3) are considered. The enhancement factors by considering the top 5% and top 10% candidates are summarized in the Supporting Information (Part S12, Table S11). The enhancement factors computed at the top 5% or 10% level are generally lower than those computed at the top 1%. Nevertheless, the top-ranked scoring functions are exactly the same.

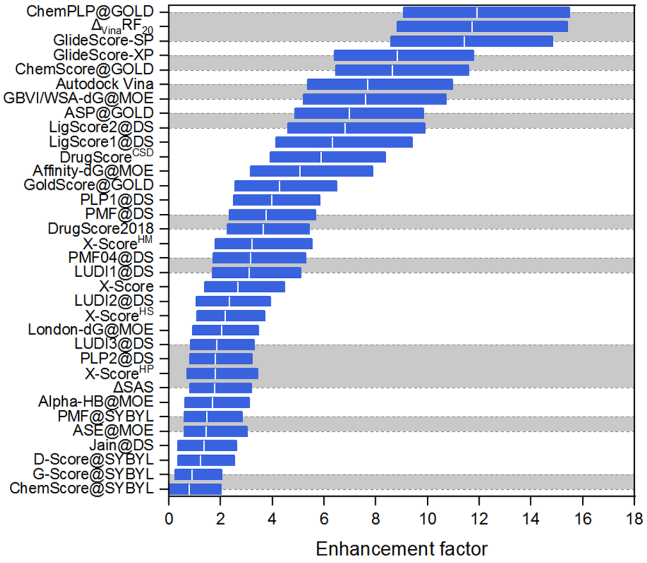

Figure 7. Average enhancement factor obtained on 57 target proteins given by each scoring function in the forward screening test. The enhancement factor is computed at the top 1% level. The blue bar indicates its 90% confidence interval. All scoring functions are ranked by their enhancement factors in a descending order. The scoring functions riding on the same gray stripe are not differentiable at $\alpha = 0.10$ in the posthoc Friedman test.

In CASF-2016, the forward screening power of a scoring function is also measured by its success rate of identifying the highest-affinity ligand for a given target protein. The success rates by considering the top 1%, 5%, and 10% candidates in screening for all scoring functions are illustrated in Figure 8. The data for making this figure are given in the Supporting Information (Part S12, Table S12). Here, exactly the same five scoring functions, i.e., $\Delta_{\mathrm{Vina}}\mathrm{RF}_{20}$, GlideScore-SP, ChemPLP@GOLD, Autodock Vina, and ChemScore@GOLD, are top ranked, where their success rates are above 28% at the top 1% level. Here, $\Delta_{\mathrm{Vina}}\mathrm{RF}_{20}$ produces the best success rate of 42.1%. Considering the size of our test set, it means that this scoring function has a chance of 42.1% to rank the highest-affinity ligand among the top three candidates. In CASF-2013, the top five scoring functions in the screening power test were GlideScore-SP, ChemScore@GOLD, GlideScore-XP, LigScore2@DS, and ChemPLP@GOLD. Most of these scoring functions are also ranked on the top in CASF-2016.

In CASF-2016, the screening power is evaluated in reverse screening trial as well. Evaluation of the "reverse screening power" will hopefully provide guidance to reverse docking studies. Here, the reverse screening power of a scoring function is measured by its success rate of identifying the true target protein for a given ligand molecule. The success rates by considering the top 1%, 5%, and 10% candidates in screening for all scoring functions are illustrated in Figure 9. The data for making this figure are given in the Supporting Information (Part S12, Table S13). In this test, the top five scoring functions are ChemPLP@GOLD, GlideScore-SP, DrugScore$^{\mathrm{CSD}}$, $\Delta_{\mathrm{Vina}}\mathrm{RF}_{20}$, and DrugScore2018. Their success rates of selecting the true target protein as the best-scored candidate (i.e., 57 × 1% ≈ 1) are over 15%. Except for DrugScore$^{\mathrm{CSD}}$ and DrugScore2018,

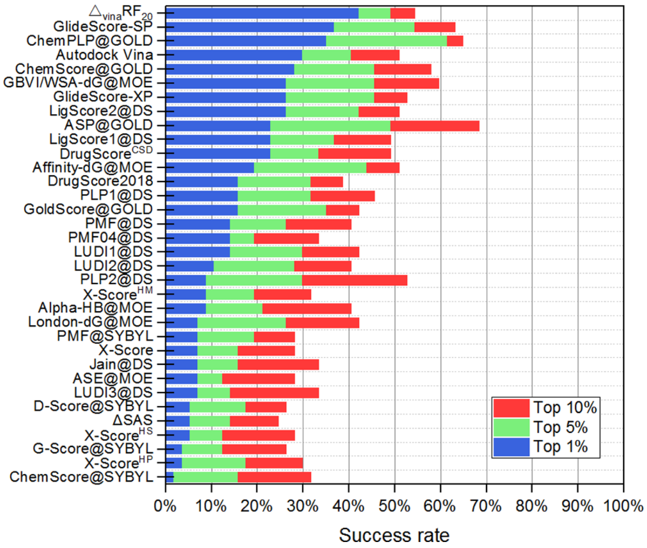

Figure 8. Success rate for detecting the highest-affinity ligand for a given target protein by each scoring function in the forward screening test. Blue, green, and red bars indicate the success rates obtained by considering the top 1%, 5%, and 10% candidates in screening, respectively. All scoring functions are ranked in a descending order by their success rates obtained at the top 1% level.

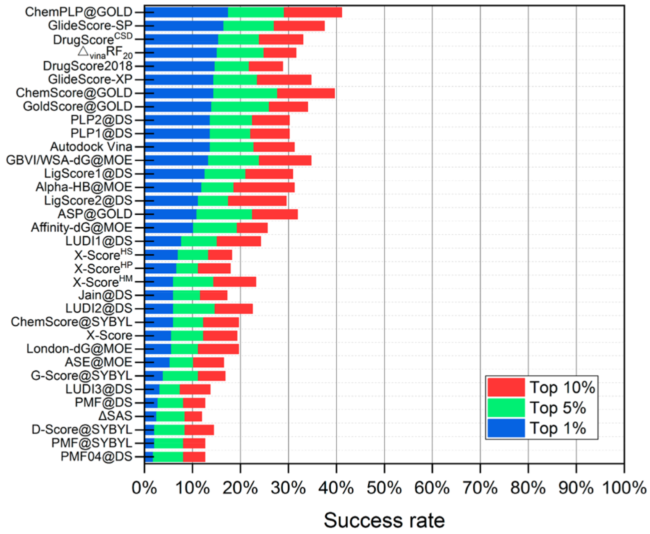

Figure 9. Success rate for detecting the best target protein for a given ligand molecule by each scoring function in the reverse screening test. Blue, green, and red bars indicate the success rates obtained by considering the top 1%, 5%, and 10% candidates in screening, respectively. All scoring functions are ranked in a descending order by their success rates obtained at the top 1% level.

those top-ranked scoring functions are generally top ranked in the forward screening trial as well. However, their success rates in the reverse screening trial are only half (or even lower) compared to their performance in the forward screening trial. For example, the success rates of $\Delta_{Vina}RF_{20}$ are 42.1% and 15.1% in the forward screening and reverse screening trials, respectively. Our results indicate that reverse screening is indeed a more challenging task than forward screening.

To summarize, our results indicate that the screening power of the scoring functions being tested is only modest as compared to their docking power. In concept, screening power may be considered as a combination of ranking power and docking power. Given the fact that the ranking power of the scoring functions under our test is generally not very promising, the observed modest screening power for them is not surprising. A very clear trend here is that the relatively successful scoring functions in the screening power test are also relatively successful in the docking power test. In particular, the scoring functions implemented in several popular molecular docking programs, such as AutoDock Vina, Glide, and GOLD, are relatively successful in the screening power test. It explains why numerous successful applications of those molecular docking programs to virtual screening have been described in literature. Finally, our results show that the reverse screening power of a scoring function is generally correlated to its forward screening power. This is also understandable because the dominant factors in protein−ligand interaction are not different between forward docking and reserve docking. Thus, the docking/scoring methods that are good at forward docking can be applied to reverse docking jobs as well.

3.6. Evaluation on Subsets of Protein−Ligand Complexes.

The primary test set in CASF-2016 is a mixture of protein−ligand complexes formed by various target proteins and ligand molecules. The purpose of using such a test set is to reflect the "average performance" of a scoring function. Nevertheless, testing a scoring function on certain types of protein−ligand complexes can provide additional information of its performance. As a conventional practice of our CASF benchmarks, the primary test set is further divided into several subsets, and then, the evaluations are repeated on those subsets. In CASF-2016, the methods used for classifying subsets have been reformed significantly. Three descriptors, including ΔVOL, ΔSAS, and H-scale, are chosen for subset classification. They all reflect the characteristics of the binding pocket on a target protein in a certain aspect (Table 1), and thus, the complexes formed by the same target protein always stay in the same subset. Consequently, all four types of performance tests are possible on every subset now, whereas only scoring power test and docking power test were possible in CASF-2013. Moreover, the primary test set is divided into three subsets in roughly the same size by any of those three descriptors (Table 1). Therefore, statistical results derived from any subset may be compared to others without the influence of uneven data size.

Evaluation results derived on three series of subsets (i.e., labeled as H1/H2/H3, S1/S2/S3, and V1/V2/V3) are illustrated in Figure 10. The results of all four major performance tests (i.e., scoring, ranking, docking, and screening) are summarized in Figure 10D, respectively. The data used for making these figures are given in the Supporting Information (Part S13, Tables S14−S17). For the convenience of making comparison, the performance of each scoring function on the entire test set is also included in Figure 10. Very rich information is embedded in this figure. In fact, each scoring function has its unique "performance profile" on those subsets. An in-depth analysis of the performance profile of each scoring function is beyond the scope of our study. We narrow our discussion to the general trend observed in each test with focus on the top-ranked scoring functions.

Evaluation results of scoring power obtained on three series of subsets are illustrated in Figure 10A. The top-ranked scoring functions here obviously perform better on target proteins with a large- or medium-sized binding pocket (i.e., subset V2 and V3). This signals an alert that future scoring functions should be more capable at handling target proteins with a small binding pocket. As on the S-series subset, our results suggest that the top-ranked

DOI: 10.1021/acs.jcim.8b00545

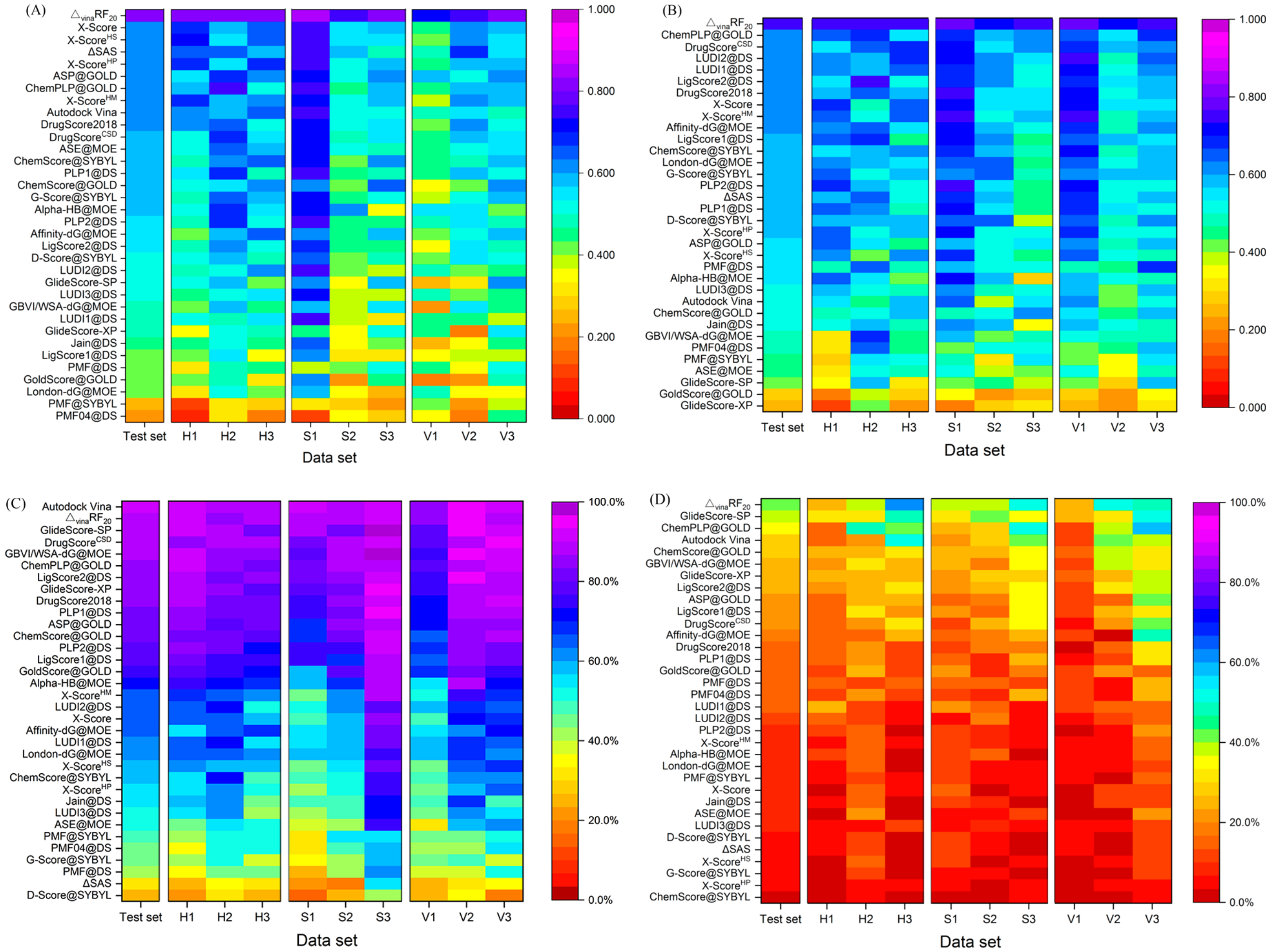

Figure 10. Evaluation results of all scoring functions obtained on the three sets of subsets. (A) Pearson correlation coefficients obtained in the scoring power test. (B) Spearman correlation coefficients obtained in the ranking power test. Evaluation results of all scoring functions obtained on the three sets of subsets. (C) Success rates of detecting the native ligand binding pose in the docking power test. (D) Success rates of detecting the highest-affinity ligand among top 1% candidates for each target protein in the forward screening power test. In all four panels, scoring functions are ranked in a descending order by their performance on the entire test set.

scoring functions here perform better when the bound ligand molecules are more exposed to the solvent (i.e., subset S1). We speculate that if a bound ligand molecule is largely buried inside the binding pocket, the strain energy change will become significant. Unfortunately, no current scoring function is able to account for such a factor sufficiently. As on the H-series subset, one can see that the top-ranked scoring functions tend to be less successful on subset H2. This suggests that it is still challenging for current scoring functions to reconcile the contributions of polar interaction and hydrophobic effect in a mixed environment.

Different patterns in the ranking power of the scoring functions being tested are observed on the three series of subsets (Figure 10B). One can see that the top-ranked scoring functions here perform better on target proteins with a small binding pocket (i.e., subset V1) or when the bound ligands molecule are more exposed to the solvent (i.e., subset S1). In either case, the protein−ligand contact interface is relatively restricted, where one particular factor, such as a hydrogen bond or a hydrophobic contact, may dominate the ranking of different ligand molecules. Such a case is relatively easy for scoring functions to deal with. On the H-series subsets, however, no clear preference can be observed on either H1, H2, or H3.

Evaluation results of docking power obtained on three series of subsets are illustrated in Figure 10C. A clear trend here is that not only the top-ranked scoring functions but also other scoring functions perform generally better on target proteins with a larger binding pocket (i.e., subsets V2 and V3) or when the bound ligand molecules are largely buried inside the binding pocket (i.e., subsets S2 and S3). This is understandable. It is expected that the native ligand binding pose forms optimal interactions with the target protein than other decoy binding poses. Therefore, when the ligand molecule contacts with the target protein more sufficiently, it is easier for current scoring functions to discriminate the native ligand binding pose from others. As on the H-series subset, performance of the top-ranked scoring functions has a marginal preference on the target proteins with a polar binding pocket (i.e., subset H1). Unlike the nondirectional hydrophobic contacts, polar interactions are usually directional, which are helpful for detecting the correct ligand binding pose. Note that very different performance profiles are observed for the top-ranked scoring functions in the docking power test as compared to the scoring/ranking power test. This again verifies that docking power is not correlated to scoring/ranking power.

Evaluation results of forward screening power obtained on three series of subsets are illustrated in Figure 10D. Our results indicate that screening power of the scoring functions being tested is generally the weakest as compared to their scoring, ranking, and docking power. Performance profiles observed here are similar to those observed in the docking power test. The top-ranked scoring functions perform generally better on target proteins with a more spacious binding pocket (i.e., subsets V2 and V3) or when the ligand molecules are buried more deeply inside the binding pocket (i.e., subsets S2 and S3). Besides, they perform better on target proteins with a more hydrophobic binding pocket (i.e., subsets H2 and H3). Here, our evaluation results again verify the correlation between the docking power and the screening power of a scoring function.

3.7. Known Limitations in CASF-2016. It is important to be aware of the limitations in CASF-2016 in order to correctly interpret the evaluation results obtained on this benchmark. Below we discuss three limitations that probably will receive the most attention from the users.

The first limitation lies in how the protein structures and ligand structures were prepared. As explained in Section 2.2, we adopted a rather simple, straightforward scheme for this job. For example, all water molecules contained in the crystal structures were removed. As a matter of fact, all 25 scoring functions under our test cannot consider water molecules explicitly. Therefore, removing all water molecules is an acceptable option for the sake of convenience. However, it is well known that water molecules bridging protein−ligand interaction may play an important role.[84,85] Thus, it would be useful to retain such water molecules in the complex structures. Another issue is how to set the protonation states of the ionizable groups on proteins and ligands. In this study, we applied a naïve protonation scheme to both the protein side and the ligand side, which may not be correct in some cases. Some advanced algorithms may be helpful for solving this issue.

Various computational methods have been described in literature to deal with water molecules, protonation states, and other issues in structure preparation. We have experience with some of them, such as the Ligand Preparation and Protein Preparation protocols implemented in the Schrödinger software. So, why have not we adopted those methods in our study? The primary difficulty faced by us is that for every task in structure preparation, we need an efficient enough method to deal with a very large number of real and computer-generated complex structures. If such a method is not well validated on a wide range of structures, it probably will introduce noise rather than be helpful. So far, we have not decided what methods are fully suitable for our purpose. We will attempt to make improvements in this aspect for future CASF benchmarks. On the other hand, keeping the simple scheme for structure preparation also has its rationale. In reality, most end users will employ a simple scheme like ours for preparing protein−ligand complex structures. If this is the case, the evaluation results provided by our CASF benchmark will match with what one can get in reality more closely.

The second limitation is that all scoring functions are tested in our study as is. The evaluation results described here may be biased if a scoring function under our test is calibrated on a data set that overlaps with the CASF-2016 test set. It is possible because the data sets derived from our PDBbind database have been very popular in scoring function development (see the Supporting Information, Part S1), while the test set used in CASF is also selected from PDBbind. Here, a valid question here is as follows: will the CASF evaluation results be biased if the given scoring function is calibrated on an overlapping training set? The answer actually depends on the nature of that scoring function. Some scoring functions (e.g., certain regression-based scoring functions) will not benefit from an overlapping training set, whereas some scoring functions will (e.g., certain machine-learning scoring functions). For example, $\Delta_{\mathrm{Vina}}\mathrm{RF}_{20}$ was calibrated on over 3300 protein−ligand complexes selected from the PDBbind v.2017 data set, which actually included 140 complexes (~50%) in the CASF-2016 test set. We speculate that this overlap contributes to the outstanding performance of $\Delta_{\mathrm{Vina}}\mathrm{RF}_{20}$ in our scoring power and ranking power tests (Figures 3 and 4). Nevertheless, this is a complicate issue. In order to get a fair, in-depth viewpoint, the readers are recommended to check the recent study by Yang[86] and other contradictory studies.[87]

The third limitation lies in the screening power test. Virtual screening is perhaps the most important application of docking/scoring methods. Some public benchmarks have been established for validating the performance of docking/scoring methods in virtual screening trials. Among them, DUD-E is arguably the most popular one.[88] DUD-E has its own concerns on selecting decoy molecules. For example, DUD-E requires that decoy molecules should not be structurally similar to but share similar physicochemical properties with the known binders. This is certainly a reasonable option, but it should not be the only one. In the test set of CASF-2016, each target protein has five known binders, and the other ligand molecules (after filling out cross-binders) are taken as decoys. A technical advantage of our method is that no additional decoy molecules outside the CASF-2016 test set are needed. Besides, the decoy molecules chosen by our method are structurally diverse, which mimic real circumstances. Nevertheless, the total number of known binders for each target protein (i.e., typically five) is still limited in the CASF-2016 test set as compared to DUD-E. This aspect can be improved though. Future CASF benchmarks will certainly employ new test sets containing larger complex clusters.

4. CONCLUSIONS

Our CASF benchmark is created particularly for validating scoring functions, i.e., a "scoring benchmark". In this sense, it is different from many other benchmarks developed for validating combined docking/scoring schemes. The core value of CASF is the definition of a set of performance metrics along with the quantitative evaluation methods. In this article, we have described the latest update of our CASF benchmark, i.e., CASF-2016. CASF-2016 is not an incremental update but has major improvements in several aspects. First, all performance tests in CASF-2016 are based on a new data with a larger size and better quality. More importantly, new evaluation methods are introduced. For example, ranking power is now evaluated by classical statistical indices, and binding funnel analysis is added to the docking power test. Also, a reverse screening trial is added to the screening power test, and confidence interval analysis is conducted in all major tests. As result, our CASF benchmark has evolved into a fairly sophisticated system, which provides multidimensional information about the performance of scoring functions. It should be emphasized that our CASF benchmark does not mean to replace other established benchmarks for validating docking/scoring methods. Instead, it should be considered as a new option with some unique features.

As a demonstration, a panel of 25 scoring functions have been evaluated in CASF-2016. Those scoring functions come from several main-stream molecular modeling softwares plus a few scoring functions released by academic groups. Our results show that the docking power of the scoring functions being tested is most promising, where a number of scoring functions are able to achieve success rates over 70%. Nevertheless, even the top-ranked scoring functions (except for $\Delta_{\mathrm{Vina}}\mathrm{RF}_{20}$) produce only modest correlation coefficients around 0.60 in the scoring power and ranking power tests. Screening power is the weakest aspect of the scoring function, where the top-ranked scoring functions produce success rates around 40%. There is no obvious correlation between docking power and scoring/ranking power. However, scoring power and ranking power are somewhat correlated, and so are docking power and screening power. Also, where most scoring functions are relatively successful in only one or two aspects, certain scoring functions exhibit a more balanced performance in all aspects, such as $\Delta_{\mathrm{Vina}}\mathrm{RF}_{20}$ and ChemPLP@GOLD. Moreover, similar trends are observed in the outcomes of CASF-2016 and CASF-2013, indicating that our benchmark produces robust evaluation results.

Although only a limited number of scoring functions have been tested in this study, it still can be observed that scoring functions developed in recent years tend to outperform the old ones. For example, the top-ranked scoring functions in the docking power and screening power test include $\Delta_{\mathrm{Vina}}\mathrm{RF}_{20}$, AutoDock Vina, ChemPLP@GOLD, GlideScore-SP, and GlideScore-XP, and all of them are relatively new. This verifies that the field of scoring function development has been moving forward as a whole. In particular, scoring functions developed with the aid of machine-learning techniques represent a new tide. Some machine-learning models have already been tested on the CASF-2013 benchmark by other researchers, which demonstrated superior performance over conventional scoring functions (see the Supporting Information, Part S1). As the chosen representative of this type of scoring function in CASF-2016, $\Delta_{\mathrm{Vina}}\mathrm{RF}_{20}$ takes an obvious lead in the scoring power and ranking power tests. It is also among the top three in the docking power and screening power tests. Although the results produced by $\Delta_{\mathrm{Vina}}\mathrm{RF}_{20}$ in our study should be interpreted with care, we are optimistic that machine-learning techniques will accelerate scoring function development.

Our CASF benchmark is designed as an open-access benchmark from the very beginning. The complete CASF-2016 benchmark, including all raw data and computer scripts for conducting the performance tests, will be released to the public at the PDBbind-CN Web site (http://www.pdbbind-cn.org/casf.asp) once this article is published. This offers the possibility that scoring functions developed by different research groups can be compared on the same ground. We are grateful for the many valuable feedbacks from the CASF users in the past. It has been a driving force for us to keep improving CASF as a useful community resource.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00545.

> Part S1: Summary of the published applications of previous CASF benchmarks. Part S2: Current rules used for selecting the PDBbind refined set. Part S3: Basic information of the PDBbind core set v.2016. Part S4: Technical methods for generating the decoy ligand binding pose.s Part S5: Cross-binding protein−ligand pairs in the PDBbind core set v.2016. Part S6: Methods used for conducting statistical tests with the R software. Part S7: Methods used for computing the three descriptors in compiling subsets. Part S8: Descriptions of the scoring functions tested in CASF-2016. Part S9: Performance of all scoring functions in the scoring power test. Part S10: Performance of all scoring functions in the ranking power test. Part S11: Performance of all scoring functions in the docking power test. Part S12: Performance of all scoring functions in the screening power test. Part S13: Performance of all scoring functions in four major performance tests conducted on subsets of protein−ligand complexes. (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

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

ORCID

Zhihai Liu: 0000-0002-4527-8829
Renxiao Wang: 0000-0003-0485-0259

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This study was financially supported by the Ministry of Science and Technology of China (National Key Research Program, Grant No. 2016YFA0502302), the National Natural Science Foundation of China (Grant Nos. 81725022, 81430083, 21472227, 21673276, and 21402230), and Chinese Academy of Sciences (Strategic Priority Research Program, Grant No. XDB20000000). We are also grateful to Prof. Yingkai Zhang at New York University and Prof. Holger Gohlke at Heinrich Heine University Düsseldorf for providing access to their scoring functions.

■ REFERENCES

(1) 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, 2002; Vol. *18*, pp 41−88.

(2) Schulz-Gasch, T.; Stahl, M. Scoring Functions for Protein-ligand Interactions: A Critical Perspective. *Drug Discovery Today: Technol.* **2004**, *1*, 231−239.

(3) 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.

(4) Rajamani, R.; Good, A. C. Ranking Poses in Structure-based Lead Discovery and Optimization: Current Trends in Scoring Function Development. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 308−315.

(5) Muegge, I.; Rarey, M. Small Molecule Docking and Scoring. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: Hoboken, NJ, 2001; Vol. *17*, pp 1−60.

(6) 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.

(7) Schneider, G.; Fechner, U. Computer-based De Novo Design of Drug-like Molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649−663.

(8) Kutchukian, P. S.; Shakhnovich, E. I. De Novo Design: Balancing Novelty and Confined Chemical Space. *Expert Opin. Drug Discovery* **2010**, *5*, 789−812.

(9) Gathiaka, S.; Liu, S.; Chiu, M.; Yang, H.; Stuckey, J. A.; Kang, Y. N.; Delproposto, J.; Kubish, G.; Dunbar, J. B., Jr; Carlson, H. A.; Burley,

S. K.; Walters, W. P.; Amaro, R. E.; Feher, V. A.; Gilson, M. K. D3R grand challenge 2015: Evaluation of Protein−ligand Pose and Affinity Predictions. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 651−668.

(10) Gaieb, Z.; Liu, S.; Gathiaka, S.; Chiu, M.; Yang, H.; Shao, C.; Feher, V. A.; Walters, W. P.; Kuhn, B.; Rudolph, M. G.; Burley, S. K.; Gilson, M. K.; Amaro, R. E. D3R Grand Challenge 2: Blind Prediction of Protein−ligand Poses, Affinity Rankings, and Relative Binding Free Energies. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 1−20.

(11) Liu, J.; Wang, R. Classification of Current Scoring Functions. *J. Chem. Inf. Model.* **2015**, *55*, 475−482.

(12) 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.

(13) Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.

(14) 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.

(15) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *J. Med. Chem.* **2004**, *47*, 558−565.

(16) 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.

(17) 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.

(18) 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.

(19) Tuccinardi, T.; Poli, G.; Romboli, V.; Giordano, A.; Martinelli, A. Extensive Consensus Docking Evaluation for Ligand Pose Prediction and Virtual Screening Studies. *J. Chem. Inf. Model.* **2014**, *54*, 2980−2986.

(20) Xu, W.; Lucke, A. J.; Fairlie, D. P. Comparing Sixteen Scoring Functions for Predicting Biological Activities of Ligands for Protein Targets. *J. Mol. Graphics Modell.* **2015**, *57*, 76−88.

(21) Rose, P. W.; Prlić, A.; Altunkaya, A.; Bi, C.; Bradley, A. R.; Christie, C. H.; Costanzo, L. D.; Duarte, J. M.; Dutta, S.; Feng, Z.; Green, R. K.; Goodsell, D. S.; Hudson, B.; Kalro, T.; Lowe, R.; Peisach, E.; Randle, C.; Rose, A. S.; Shao, C.; Tao, Y.-P.; Valasatava, Y.; Voigt, M.; Westbrook, J. D.; Woo, J.; Yang, H.; Young, J. Y.; Zardecki, C.; Berman, H. M.; Burley, S. K. The RCSB Protein Data Bank: Integrative View of Protein, Gene and 3D Structural Information. *Nucleic Acids Res.* **2017**, *45*, D271−D281.

(22) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrian-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magarinos, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945−D954.

(23) Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 Update. *Nucleic Acids Res.* **2017**, *45*, D955−D963.

(24) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. *Nucleic Acids Res.* **2016**, *44*, D1045−D1063.

(25) 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.

(26) 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.

(27) 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.

(28) Smith, R. D.; Damm-Ganamet, K. L.; Stuckey, J. A.; Ahmed, A.; Convery, M. A.; Somers, D. O.; Kranz, M.; Elkins, P. A.; Cui, G.; Peishoff, C. E.; Lambert, M. H.; Dunbar, J. B., Jr; Carlson, H. A. CSAR **2014**: A Benchmark Exercise Using Unpublished Data from Pharma. *J. Chem. Inf. Model.* **2016**, *56*, 1063−1077.

(29) 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.

(30) Li, Y.; Liu, Z. H.; Han, L.; Li, J.; Liu, J.; Zhao, Z. X.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: I. Compilation of the Test Set. *J. Chem. Inf. Model.* **2014**, *54*, 1700−1716.

(31) Li, Y.; Han, L.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: II. Evaluation Methods and General Results. *J. Chem. Inf. Model.* **2014**, *54*, 1717−1736.

(32) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-wide Collection of Binding Data: Current Status of the PDBbind Database. *Bioinformatics* **2015**, *31*, 405−412.

(33) Liu, Z.; Su, M.; Han, L.; Liu, J.; Yang, Q.; Li, Y.; Wang, R. Forging the Basis for Developing Protein−Ligand Interaction Scoring Functions. *Acc. Chem. Res.* **2017**, *50*, 302−309.

(34) Huang, Y.; Niu, B.; Gao, Y.; Fu, L.; Li, W. CD-HIT Suite: A Web Server for Clustering and Comparing Biological Sequences. *Bioinformatics* **2010**, *26*, 680−682.

(35) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public Ki Data. *J. Med. Chem.* **2012**, *55*, 5165−5173.

(36) Emsley, P.; Lohkamp, B.; Scott, W.; Cowtan, K. Features and Development of Coot. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, *66*, 486−501.

(37) Clark, M.; Cramer, R. D.; Vanopdenbosch, N. Validation of the General-Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, *10* (8), 982−1012.

(38) Zhao, Y.; Cheng, T.; Wang, R. Automatic Perception of Organic Molecules Based on Essential Structural Information. *J. Chem. Inf. Model.* **2007**, *47*, 1379−1385.

(39) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Goetz, A. W.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Salomon-Ferrer, R.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER 12*; University of California, San Francisco, 2012.

(40) Allen, W. J.; Rizzo, R. C. Implementation of the Hungarian Algorithm to Account for Ligand Symmetry and Similarity in Structure-Based Design. *J. Chem. Inf. Model.* **2014**, *54*, 518−529.

(41) Shivakumar, D.; Harder, E.; Damm, W.; Friesner, R. A.; Sherman, W. Improving the Prediction of Absolute Solvation Free Energies Using the Next Generation OPLS Force Field. *J. Chem. Theory Comput.* **2012**, *8*, 2553−2558.

(42) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 281−296.

(43) Spearman, C. The Proof and Measurement of Association Between Two Things. *Am. J. Psychol.* **1904**, *15*, 72−101.

(44) Knight, W.R. A. Computer Method for Calculating Kendall's Tau with Ungrouped Data. *J. Am. Stat. Assoc.* **1966**, *61*, 436−439.

(45) Pearlman, D. A.; Charifson, P. S. Are Free Energy Calculations Useful in Practice? A Comparison with Rapid Scoring Functions for the p38 MAP Kinase Protein System. *J. Med. Chem.* **2001**, *44*, 3417−3423.

DOI: 10.1021/acs.jcim.8b00545

(46) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, 1083−1090.

(47) Lee, A.; Lee, K.; Kim, D. Using Reverse Docking for Target Identification and Its Applications for Drug Discovery. *Expert Opin. Drug Discovery* **2016**, *11*, 707−715.

(48) Nicholls, A. Confidence Limits, Error Bars and Method Comparison in Molecular Modeling. Part 1: The Calculation of Confidence Intervals. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 887−918.

(49) Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics* **1979**, *7*, 1−26.

(50) Wortmann, J. H.; Park, C. L.; Edmondson, D. Trauma and Ptsd Symptoms: Does Spiritual Struggle Mediate the Link? *Psychological Trauma* **2011**, *3*, 442−452.

(51) Adèr, H. J.; Mellenbergh, G. J.; Hand, D. J. *Advising on Research Methods: A Consultant's Companion*; Johannes van Kessel Publishing: Huizen, The Netherlands, 2008.

(52) Efron, B. Better Bootstrap Confidence Intervals (with Discussion). *J. Am. Stat. Assoc.* **1987**, *82*, 171−200.

(53) The R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017. https://www.R-project.org/.

(54) Shaffer, J. Modified Sequentially Rejective Multiple Test Procedures. *J. Am. Stat. Assoc.* **1986**, *81* (395), 826−831.

(55) Demsar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mac. Learn. Res.* **2006**, *7*, 1−30.

(56) Friedman, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675−701.

(57) Friedman, M. A Comparison of Alternative Tests of Significance for the Problem of M Rankings. *Ann. Math. Stat.* **1940**, *11* (1), 86−92.

(58) 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.

(59) 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.

(60) 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.

(61) Muegge, I. PMF Scoring Revisited. *J. Med. Chem.* **2006**, *49*, 5895−5902.

(62) 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.

(63) 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.

(64) 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., Des. Sel.* **1995**, *8*, 677−691.

(65) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Rose, P. W.; et al. Deciphering Common Failures in Molecular Docking of Ligand-protein Complexes. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 731−751.

(66) 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.

(67) 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.

(68) 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.

(69) 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.

(70) Mooij, W. T. M.; Verdonk, M. L. General and Targeted Statistical Potentials for Protein-Ligand Interactions. *Proteins: Struct., Funct., Genet.* **2005**, *61*, 272−287.

(71) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated Docking with Grid-based Energy Evaluation. *J. Comput. Chem.* **1992**, *13*, 505−524.

(72) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(73) 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.

(74) 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.

(75) Trott, O.; Olson, J. A. Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.

(76) 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.

(77) Wang, C.; Zhang, Y. K. Improving Scoring-Docking-Screening Powers of Protein−Ligand Scoring Functions using Random Forest. *J. Comput. Chem.* **2017**, *38*, 169−177.

(78) Dittrich, J.; Schmidt, D.; Pfleger, C.; Gohlke, H. Converging a Knowledge-Based Scoring Function: DrugScore2018. *J. Chem. Inf. Model.*, in revision.

(79) 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.

(80) Royston, J. P. An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. *Applied Statistics* **1982**, *31*, 115−124.

(81) Royston, J. P. Algorithm AS 181: The W Test for Normality. *Applied Statistics* **1982**, *31*, 176−180.

(82) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(83) Meyder, A.; Nittinger, E.; Lange, G.; Klein, R.; Rarey, M. Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-ray Structures. *J. Chem. Inf. Model.* **2017**, *57*, 2437−2447.

(84) Spyrakis, F.; Ahmed, M. H.; Bayden, A. S.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E. The Roles of Water in the Protein Matrix: A Largely Untapped Resource for Drug Discovery. *J. Med. Chem.* **2017**, *60*, 6781−6827.

(85) Garcia-Sosa, A. T. Hydration Properties of Ligands and Drugs in Protein Binding Sites: Tightly-Bound, Bridging Water Molecules and Their Effects and Consequences on Molecular Design Strategies. *J. Chem. Inf. Model.* **2013**, *53*, 1388−1405.

(86) Li, Y.; Yang, J. Structural and Sequence Similarity Makes a Significant Impact on Machine-Learning-Based Scoring Functions for Protein−Ligand Interactions. *J. Chem. Inf. Model.* **2017**, *57*, 1007−1012.

(87) Li, H.; Peng, J.; Leung, Y.; Leung, K.-S.; Wong, M.-H.; Lu, G.; Ballester, P. J. The Impact of Protein Structure and Sequence Similarity on the Accuracy of Machine-Learning Scoring Functions for Binding Affinity Prediction. *Biomolecules* **2018**, *8*, 12−19.

(88) 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.