ACCOUNTS of chemical research

Article

pubs.acs.org/accounts

Forging the Basis for Developing Protein–Ligand Interaction Scoring Functions

Zhihai Liu,[†] Minyi Su,[†] Li Han,[†] Jie Liu,[†] Qifan Yang,[†] Yan Li,[*,†] and Renxiao Wang[*,†,‡]

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

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

Supporting Information

CONSPECTUS: In structure-based drug design, scoring functions are widely used for fast evaluation of protein–ligand interactions. They are often applied in combination with molecular docking and *de novo* design methods. Since the early 1990s, a whole spectrum of protein–ligand interaction scoring functions have been developed. Regardless of their technical difference, scoring functions all need data sets combining protein–ligand complex structures and binding affinity data for parametrization and validation. However, data sets of this kind used to be rather limited in terms of size and quality. On the other hand, standard metrics for evaluating scoring function used to be ambiguous. Scoring functions are often tested in molecular docking or even virtual screening trials, which do not directly reflect the genuine quality of scoring functions. Collectively, these underlying obstacles have impeded the invention of more advanced scoring functions.

In this Account, we describe our long-lasting efforts to overcome these obstacles, which involve two related projects. On the first project, we have created the PDBbind database. It is the first database that systematically annotates the protein–ligand complexes in the Protein Data Bank (PDB) with experimental binding data. This database has been updated annually since its first public release in 2004. The latest release (version 2016) provides binding data for 16 179 biomolecular complexes in PDB. Data sets provided by PDBbind have been applied to many computational and statistical studies on protein–ligand interaction and various subjects. In particular, it has become a major data resource for scoring function development. On the second project, we have established the Comparative Assessment of Scoring Functions (CASF) benchmark for scoring function evaluation. Our key idea is to decouple the "scoring" process from the "sampling" process, so scoring functions can be tested in a relatively pure context to reflect their quality. In our latest work on this track, i.e. CASF-2013, the performance of a scoring function was quantified in four aspects, including "scoring power", "ranking power", "docking power", and "screening power". All four performance tests were conducted on a test set containing 195 high-quality protein–ligand complexes selected from PDBbind. A panel of 20 standard scoring functions were tested as demonstration. Importantly, CASF is designed to be an open-access benchmark, with which scoring functions developed by different researchers can be compared on the same grounds. Indeed, it has become a popular choice for scoring function validation in recent years.

Despite the considerable progress that has been made so far, the performance of today's scoring functions still does not meet people's expectations in many aspects. There is a constant demand for more advanced scoring functions. Our efforts have helped to overcome some obstacles underlying scoring function development so that the researchers in this field can move forward faster. We will continue to improve the PDBbind database and the CASF benchmark in the future to keep them as useful community resources.

1. INTRODUCTION

Structure-based design has proven to be a successful strategy in modern drug discovery.[1] Among the many computational methods applicable to structure-based drug design, molecular docking is probably the most popular one.[2] Molecular docking explores the possible binding modes of a ligand to a given molecular target (normally a protein) to determine the optimal binding mode. Knowledge of the ligand binding mode can be used in turn to predict ligand binding affinity. Using molecular docking, large compound libraries can be screened computationally first, and only potential binders to the target protein will be validated in experiment. This cost-efficient "virtual screening" technique has become a popular choice for lead discovery in both academia and the pharmaceutical industry.[3]

Received: September 29, 2016
Published: February 9, 2017

DOI: 10.1021/acs.accounts.6b00491
Acc. Chem. Res. 2017, 50, 302–309

During a molecular docking process, numerous ligand binding poses generated during conformational sampling need to be assessed. Ideally, the correct binding pose should be associated with the best binding score. Moreover, a high-affinity ligand should have a better binding score than a low-affinity binder or a nonbinder. Docking programs developed in early years (such as the old versions of DOCK and AutoDock) used to borrow force field energy functions for this purpose. Nevertheless, modern docking programs (such as GOLD, GLIDE, and Surflex-Dock) all rely on a special family of methods called "scoring functions"[4,5] for evaluating protein−ligand interactions. Scoring functions do not attempt to account for protein−ligand interactions with a high-level theory. Instead, they make various approximations to achieve a compromise between speed and accuracy. Thus, they are particularly suitable for high-throughput tasks. Besides molecular docking, automatic *de novo* design methods[6] also employ scoring functions. Such methods construct virtual ligand molecules within the designated binding pocket on a target protein. Typically, thousands of ligand molecules need to be assessed at each round of structural operation during a *de novo* design job. Only protein−ligand interaction scoring functions can accomplish this mission in a reasonable amount of time.

The first wave of scoring functions appeared in the early 1990s. Considerable progress has been made since then. Now scoring functions have grown into a large, diverse family. Our estimation is that over a hundred scoring functions have already been published. Current scoring functions can be classified roughly into four categories, physics-based methods, empirical scoring functions, knowledge-based statistical potentials, and descriptor-based machine-learning scoring functions.[4,5,7] They are based on different assumptions and follow different approaches for model construction. Developing more advanced scoring functions has remained a primary research interest of our group. Previously, we developed an empirical scoring function called X-Score.[8] X-Score performed well in several comparative evaluations conducted by other researchers[9] and ourselves.[10,11] It is still a popular scoring function today.

Nevertheless, perhaps our more recognized contributions so far are supplying the logistics much needed by the scoring function community. To be specific, we have created the PDBbind database, which aims at providing a comprehensive collection of protein−ligand complexes with known structures and binding affinity data. Through continuous efforts over a decade, PDBbind has become a major data resource for scoring function development. In addition, we have developed a whole set of metrics for evaluating the performance of scoring functions, namely, the Comparative Assessment of Scoring Functions (CASF) benchmark. Our benchmark overcomes certain shortcomings in other benchmarks. It produces more objective results regarding the strength and weakness of current scoring functions. CASF has also become a popular choice for scoring function validation among the scoring function community.

In the following sections, we will describe our work on developing the PDBbind database and the CASF benchmark. Relevant work done by other researchers will also be reviewed briefly.

2. CREATION OF THE PDBbind DATABASE

2.1. History of the PDBbind Database

Developing a scoring function needs an adequate amount of data in appropriate form. Scoring functions are designed to characterize protein−ligand interactions at atomic resolution. Thus, three-dimensional structures of protein−ligand complexes resolved by experimental techniques are indispensable for this purpose. Moreover, since scoring functions are often expected to predict ligand binding affinities, binding data are also desired for calibrating a scoring function. In fact, **P**rotein−**L**igand compl**EX**es with known **B**inding **A**ffinity and **S**tructure (PLEXBAS) have been a major form of data required by studies on docking/scoring methods. However, knowledge of PLEXBAS used to be rather limited in early years. It is not uncommon that computational chemists are more enthusiastic about inventing new fancy methods rather than engaging on supportive jobs such as data collection. In nearly ten years (from the early 1990s to the early 2000s), data sets of PLEXBAS used by scoring function developers grew incrementally from several dozens of samples to around three hundred samples (see Table S1 in the Supporting Information). Such a pace of progress is of course not very impressive. Besides, many samples included in those data sets had miscellaneous problems in complex structure or binding data. The true power of the scoring functions calibrated on such data sets was thereby in doubt. Lack of large, reliable data sets was an obvious obstacle in the field of scoring function development at that time.

Some researchers already attempted to build larger collections of PLEXBAS, such as LPDB (~200 complexes, 2001),[12] PLD (~300 complexes, 2003),[13] AffinDB (~490 complexes, 2006),[14] and the collaborative Scoring Function Consortium (~600 complexes, 2008).[15] However, those collections were still modest in size. No regular update was conducted because binding data were essentially curated from literature in a random manner. Those collections also more or less share the same quality problems mentioned above. A more promising solution has yet to be sought.

We were among the first-generation scoring function developers, so we also felt the pain. Back in 2001, we conceived an ambitious project for tackling this problem. Our plan was to identify all of the valid protein−ligand complexes in the Protein Data Bank (PDB)[16] and then collect their binding data from literature. PDB has been the largest open resource of biomolecular structures. Thus, the outcome of this project in principle would be the largest collection of PLEXBAS ever. Besides, if we could establish an efficient workflow, our collection would keep growing through regular updates because PDB itself is in a constant growth. Once the data set became large enough, one could afford the luxury of using only high-quality samples in scoring function development. This seemed to be a promising and sustainable solution to the not-enough-data-to-use problem.

This plan was then carried out by the correspondent author of this Account and co-workers in Prof. Shaomeng Wang's group at the University of Michigan. The outcome was the PDBbind database, which was released to the public in 2004.[17,18] The first release of PDBbind was based on the 19 621 experimental structures available from PDB in January 2003. Among them, 5671 valid protein−ligand complexes were identified, and binding data for 1359 complexes were collected from literature. In addition, 800 complexes with relatively reliable structures and binding data were selected into a so-called "refined set". This data set reset the standard of the PLEXBAS data sets useful for scoring

Table 1. Basic Information for Each Release of PDBbind since 2007

			complexes with binding data				
version	entries in PDB[a]	valid complexes[b]	protein−ligand	nucleic acid−ligand	protein−protein	protein−nucleic acid	total
2007	40 876	11 822	3124	0	0	0	3124
2008	48 092	18 211	3539	40	471	250	4300
2009	55 118	23 284	4277	44	1053	304	5678
2010	62 387	26 434	5075	55	1281	361	6772
2011	70 224	30 259	6051	66	1441	428	7986
2012	78 235	34 180	7121	79	1597	511	9308
2013	87 085	38 918	8302	83	1804	587	10 776
2014	96 592	44 569	10 656	87	1592	660	12 995
2015	105 183	48 821	11 987	109	1807	717	14 260
2016	114 344	53 838	13 308	118	1976	777	16 179

[a] Each release of PDBbind is based on the contents available from PDB in the first week of that year. [b] Defined by the PDBbind classification scheme. Only valid complexes are considered in subsequent binding data collection.

function studies, which was applied by ourselves[11] and other researchers soon afterward. Since 2007, the PDBbind database has been maintained solely by our group at the Shanghai Institute of Organic Chemistry, which is accessible at the PDBbind-CN web server (http://www.pdbbind-cn.org/). We have been able to update this database annually to keep up with the growth of PDB. The basic information on each release since 2007 is summarized in Table 1. The latest release (i.e., version 2016) just went public on December 8, 2016.

2.2. Current Status of the PDBbind Database

Our workflow for compiling PDBbind contains three major steps. The overall workflow has not changed since the very beginning, but the methods employed at each step have improved considerably along the track. Details of our current methods are described in our recent publication.[19] Below they are briefly summarized together with the results from the latest update.

The first step is to classify the valid biomolecular complexes in PDB. The protocols used for this task would better be automated because a large number of entities need to be processed during each update. We have developed a fairly complicated classification scheme (Figure S1 in the Supporting Information). It takes a standard PDB-format file as the input, analyzes the recorded structural information, and then decides if it is a valid complex of our interest. Here, valid complexes include those formed between (i) protein and small-molecule ligand, (ii) nucleic acid and small-molecule ligand, (iii) protein and protein, and (iv) protein and nucleic acid. Early versions of PDBbind only contained protein−ligand complexes. The other three categories of complexes have been added since version 2008 (Table 1). The whole classification scheme is automated by a set of computer programs, which can process roughly one PDB structure per second on a single CPU.

Classification of the entire PDB by PDBbind version 2016 is shown in Figure 1. One can see that half of the PDB structures can be classified as complexes of a certain type. In particular, one-third of the PDB structures are valid protein−ligand complexes, which are most useful for studies on protein−ligand interaction. This list of valid protein−ligand complexes is also available from us upon request.

The second step is to collect binding data for the complexes identified at the first step. The public literature is simply too comprehensive for a systematic screening. Our approach is to focus on the "primary reference" indicated in each PDB structure file. It actually offers a reasonable chance (∼30%) for finding the desired binding data by examining a single reference. Even so, a significant amount of human labor is required for this task because the correct binding data and other relevant information have to be retrieved manually. We record three forms of binding data during this process, including dissociation constant (K_d), inhibition constant (K_i), and concentration at 50% inhibition (IC_{50}). Results obtained from various functional assays (e.g., EC_{50} value) are not considered. Importantly, each reference is examined parallelly by two persons. This quality-control measure is employed to minimize the human error made during this process. It should be emphasized here that every single binding data recorded in PDBbind is retrieved from an original reference rather than copied from other data resources. So far we have examined nearly 30 000 references on this project.

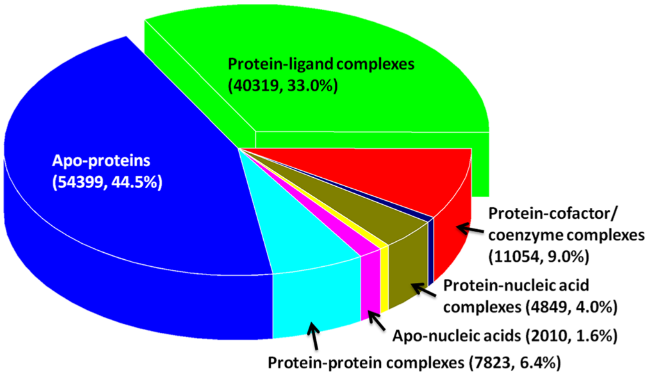

Figure 1. Classification of the entire PDB as in January 2016. Items lower than 1% of the total population are not labeled on this figure.

The latest version of PDBbind (i.e., version 2016) provides binding data for 16 179 molecular complexes in PDB (Table 1). Compared to the first release of PDBbind, the current binding data collection has increased by over 10-fold. Binding data distributions for the four categories of complexes are given in the Supporting Information (Figure S2). One can see that the majority of binding data scatters at the medium range (e.g., log K_a = 5−8). But there are also a considerable number of low-affinity and the high-affinity samples, which may be of interest for certain studies.

At the last step, all collected information is integrated into a web-based database accessible at http://www.pdbbind-cn.org/. On this Web site, users can display protein−ligand complex structures in different modes and search among the binding data and other information recorded in PDBbind (Figure S3 in the Supporting Information). Hyperlinks to several external databases (e.g., PDB and PDBsum) are also provided, so users can reach other information about the complexes under their inspection conveniently.

Raw material will become more useful if it is further processed. As an additional feature, a "refined set" is selected out of the PLEXBAS recorded in PDBbind (namely, the "general set") with concerns on (i) quality of the complex structure, (ii) quality of the binding data, and (iii) biological/chemical nature of the complex. A fairly stringent set of rules are applied to sample selection (Table S2 in the Supporting Information). The refined set is also updated with each version of PDBbind. For the user's convenience, the original PDB structural file of each complex in this data set is processed and saved in standard formats, so it is readily readable by most molecular modeling software. The whole data set can be downloaded as a package from the PDBbind-CN Web site.

The overall quality of the refined set is much better than the general set. Less than one-third of the complexes in the general set qualify for the refined set. For example, the latest refined set (version 2016) contains 4057 protein−ligand complexes, which are selected out of a total of 13 308. The refined set is compiled to provide a more reliable basis for docking/scoring studies. Indeed, many researchers took this data set directly to their studies. However, the refined set should not be considered as an "ideal" or even "high-quality" data set. Instead, it serves as a generally "acceptable" data set by excluding samples with obvious problems. Other researchers are welcome to employ even more stringent criteria to compile their own data sets based on what PDBbind provides.

2.3. Impact of the PDBbind Database

PDBbind is the first open-access database that systematically annotates protein−ligand complexes with binding data on the PDB level. It has been in existence for more than a decade. Our records indicate that researchers from over 70 countries around the world have used this database. Our original motivation for creating PDBbind was to support the studies on scoring function. Indeed, many new scoring functions are calibrated or validated with the data sets from PDBbind. PDBbind has become a dominant data resource for this type of study.[20] We ourselves have also utilized PDBbind in a number of studies, for example, developing the CASF benchmark for scoring function assessment,[21−24] deriving the geometrical preferences of hydrogen bonds,[25] analyzing the binding data of covalent binders,[26] testing the MM-PB/SA method on conformational ensembles of protein−ligand complexes,[27] developing a knowledge-guided scoring strategy,[28] and mining interaction patterns on protein−protein binding interface.[29]

Apart from docking/scoring studies, PDBbind has been applied to various subjects. The protein−ligand complexes included in PDBbind cover a wide range of validated and potential drug targets as well as bioactive small molecules. Such information is utilized to analyze drug−target interaction network, predict side effects and toxicity, build knowledge-based models for discriminating drugs and nondrugs, and so on. According to our survey, nearly two hundred studies that utilize the information provided by PDBbind have already been published. For the readers' reference, a complete list of those works is given in the Supporting Information (Table S3).

Nowadays a number of open-access databases also provide useful information on protein−ligand interactions.[20] A few derivative databases based on PDB, such as Relibase (http://relibase.ccdc.cam.ac.uk/), PDBeMotif (http://www.ebi.ac.uk/pdbe-site/pdbemotif/), and scPDB (http://cheminfo.u-strasbg.fr/scPDB/), assemble protein−ligand complex structures and their relevant structural information (such as binding pocket and ligand-based properties). Several other databases, such as ChEMBL (https://www.ebi.ac.uk/chembl/), PubChem (https://pubchem.ncbi.nlm.nih.gov/), and BindingDB (https://www.bindingdb.org/), provide comprehensive collections of the binding data of bioactive compounds and information about their molecular targets. The rich information recorded in those databases is presumably welcome more by medicinal chemists and chemical biologists. Nevertheless, computational studies on protein−ligand interaction require the connection between structural information and energetic properties. This connection is the primary goal of PDBbind but not any of the databases mentioned above. Some protein−ligand binding data in ChEMBL and Binding DB do have annotations linking to corresponding PDB entries, so such data overlap with the scope of PDBbind. However, our recent analysis revealed that only ~20% of the binding data recorded in PDBbind could find their counterparts in ChEMBL,[30] although the binding data collection hosted by ChEMBL is about 100 times larger than PDBbind. Currently, only the Binding MOAD database (http://www.bindingmoad.org/) stays in the same category as PDBbind. Binding MOAD also aims at collecting binding data for the protein−ligand complexes throughout PDB, but it is developed with a somewhat different focus and workflow.[31] A comparison of the basic features of PDBbind and Binding MOAD is given in the Supporting Information (Table S4).

3. DEVELOPMENT OF THE CASF BENCHMARK

3.1. Why Do We Need a "Scoring Benchmark"?

Since a whole spectrum of scoring functions have already been developed, objective assessment of these methods on some standard benchmarks becomes necessary. The significance of establishing such benchmarks is twofold: End-users of scoring functions may rely on these benchmarks to make smart choices among the available methods, while methodology developers also need these benchmarks to interpret the strength and weakness of current scoring functions. In fact, many comparative studies of docking/scoring methods have been published.[32] Those studies evaluated various combinations of docking and scoring methods by the ability to reproduce the known ligand binding poses and ligand binding affinities or the ability to select active compounds among random molecules. Those studies certainly have their values. However, they are carried out on various data sets with different performance metrics. Controversy does exist in the evaluation results produced by such studies.[33] Moreover, scoring functions are typically tested in context of molecular docking or even virtual screening. The final outcomes of such a test, to its best, reflect the performance of a docking program under a certain configuration but not any of its individual components. Thus, this type of evaluation is not completely appropriate for assessing scoring functions.

Our opinion is that there should be "scoring benchmarks" aside from "docking benchmarks". We have been striving to establish such a benchmark by designing a set of meaningful metrics for measuring the performance of scoring functions. The key idea in our method is to decouple the "scoring" process from the "sampling" process, so scoring functions can be evaluated in a relatively pure context to reflect their genuine quality. In an early attempt published in 2003,[10] we tested 11 scoring functions on 100 protein−ligand complexes. Our evaluation focused on the ability of those scoring functions in reproducing known ligand binding poses and binding affinities. A new standard set in that study was that the possible ligand binding poses for each complex were generated in advance using a standard docking program, and then each scoring function under test was used simply to rank those binding poses. In this way, all scoring functions were evaluated on the same conformational ensemble, and the possible bias originating from the sampling process was eliminated. To the best of our knowledge, this study was the prototype of all scoring benchmarks published afterward.

Our efforts along this track have evolved into the Comparative Assessment of Scoring Functions (CASF) benchmark. Meanwhile, creation of the PDBbind database has helped this project by providing the required data sets. The first published study was CASF-2007,[21] which was named so because the test set used in that study was selected from PDBbind version 2007. Recently, we published another update, CASF-2013.[23,24] Major improvements were made to the evaluation methods. A new test set was compiled through a systematic sampling of PDBbind version 2013, and a larger panel of scoring functions were included in the evaluation. CASF-2013 will be described in the next section. We plan to develop a new update of CASF every few years in the future.

3.2. Main Features of the CASF-2013 Benchmark

3.2.1. Selection of the Test Set. A good benchmark needs a high-quality test set to produce reliable results. To ensure that each selected complex has reliable crystal structure and binding data, the test set used in CASF-2013 is selected from the 2959 protein−ligand complexes included in the PDBbind refined set (version 2013). A systematic sampling on the refined set is conducted in order to select out the nonredundant, representative samples (Figure 2). Basically, all complexes are clustered by protein sequence similarity first, so each resulting cluster consists of complexes formed by one particular protein. Then, the complex with the highest affinity, the one with the lowest affinity, and the one with the median affinity in each cluster are selected as the representatives. In addition, the electron density map of each selected complex is examined visually to ensure that there is no obvious fitting problem on the ligand structure or nearby residues. Thus, quality of the complex structures included in this data set is generally better than those in the refined set. The final outcome, namely, the "core set", is composed of 195 complexes in 65 clusters.

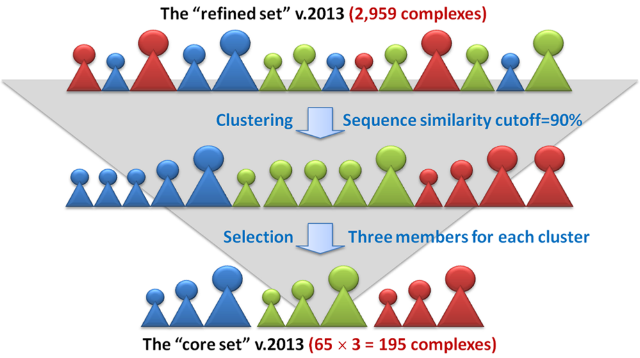
The "refined set" v.2013 (2,959 complexes)
Clustering
Sequence similarity cutoff=90%
Selection
Three members for each cluster
The "core set" v.2013 (65 × 3 = 195 complexes)

Figure 2. Illustration of how the test set in CASF-2013 is selected.

3.2.2. Evaluation Methods. The performance of each scoring function was evaluated in CASF-2007 by three features, "scoring power", "ranking power", and "docking power". A new feature called "screening power" has been added to CASF-2013. This set of metrics is designed to match the typical applications of scoring functions in reality, that is, predicting ligand binding modes and binding affinities as well as retrieving known binders from a random pool in virtual screening.

Scoring power refers to the ability of a scoring function to produce binding scores in a linear correlation with experimental binding data when the three-dimensional structures of the protein−ligand complexes under consideration are given. This feature is evaluated directly on the 195 crystal complex structures in the test set. The scoring power of a scoring function is measured by the Pearson correlation coefficient between its binding scores and experimental binding constants and the standard deviation in regression fitting.

Ranking power refers to the ability of a scoring function to correctly rank the known ligands of a target protein by their binding affinities when the binding poses of these ligands are given. This feature is also evaluated directly on the 195 crystal complex structures in the test set. The ranking power of a scoring function is measured by its success rate of correctly ranking the three complexes in each cluster over the entire test set. Note that ranking power is different from scoring power in two aspects: First, scoring power is evaluated on a set of diverse complexes regardless of ligand type or protein type, whereas ranking power is evaluated on cognate complexes formed by one target protein. Second, the scoring power test requires the scoring function to produce binding scores in linear correlation with experimental binding data, whereas the ranking power test only requires the scoring function to rank the known binders correctly. A scoring function with good ranking power is often adequate for virtual screening jobs.

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. In CASF-2013, docking power is essentially evaluated through a "self-docking" trial. To conduct this test, an ensemble of up to 100 ligand binding poses for each complex are selected among the outcomes produced by three docking programs (GOLD, MOE-Dock, and Surflex-Dock). Then, each scoring function under test is applied to rank all of the ligand binding poses for each complex. Its docking power is measured by the success rate of finding the top-ranked binding pose close enough to the native pose (e.g., RMSD < 2.0 Å).

Screening power refers to the ability of a scoring function to identify the true binders to a target protein among a pool of random molecules. In CASF-2013, screening power is evaluated in a "cross-docking" trial. All 195 ligand molecules in the test set are docked onto each of the 65 target proteins, generating totally 65 × 195 = 12 675 protein−ligand pairs. For each pair, the representative ligand binding poses are also selected from the outcomes produced by the three docking programs. Then, each scoring function is applied to rank all ligand molecules on each target protein to select the optimal binders. Its screening power is measured by enrichment factor or success rate of finding the three true binders for each target protein over the entire test set. One may argue that the self-docking trial employed in our docking power test does not reflect the reality where the conformation of the target protein after ligand binding is unknown in advance. The cross-docking trial in our screening power test actually provides another angle for examining the docking power of a scoring function in a more realistic scenario.

3.2.3. Standard Scoring Functions under Evaluation. A panel of 20 scoring functions are tested in CASF-2013, including 18 scoring functions implemented in several main-stream software suites (Discovery Studio, SYBYL, MOE, Schrödinger, and GOLD) and X-Score. In addition, a single-descriptor scoring function based on the buried solvent-accessible surface area (ΔSAS) of the ligand is introduced as a reference. Descriptions of all 20 scoring functions as well as their results produced in the four performance tests can be found in our recent publication.[24]

In brief, our results indicate that the performance of those scoring functions in the scoring/ranking power tests is generally not very promising. On the other hand, some scoring functions achieve rather impressive success rates in the docking/screening power tests, which explains why molecular docking can be useful. Another observation is that the relatively successful scoring functions in the scoring power test also perform better in the ranking power test; while the relatively successful scoring functions in the docking power test also perform better in the screening power test. A few scoring functions, such as ChemPLP in GOLD and PLP in Discovery Studio, demonstrate a more balanced performance between docking/screening power and scoring/ranking power. Moreover, the top-ranked scoring functions in each performance test in CASF-2013 are roughly the same as those observed in CASF-2007. This indicates the robustness of our evaluation results because the test sets used in these two benchmarks are significantly different.

3.3. Impact of the CASF benchmark

It is our important strategy to design CASF as an open-access benchmark. The data sets used in both CASF-2007 and CASF-2013 have been released to the public. Because there are so many scoring functions, it is simply beyond our capability to collect and test them. If one utilizes our data set and follows our evaluation methods, his/her results can be compared directly to our results. In the same way, other researchers can make comparison among themselves. Compared to organizing a centralized mechanism, this distributed mechanism is more efficient and more flexible to get more scoring functions involved in a fair comparison.

The scoring function community seems to welcome our work warmly. According to our survey, nearly 40 applications of CASF-2007 or CASF-2013, which either fully applied our benchmark to scoring function evaluation or utilized our data sets in certain ways, have already been published since 2010. In particular, a significant portion of the new scoring functions published in recent years have chosen CASF as a major form of validation.[34] With the performance metrics defined by CASF, a notable new trend in this field is that scoring function can be parametrized through multiobjective optimization to achieve a more balanced performance.[35]

It should be noted that our CASF benchmark has its own scope and focus. It is not intended to replace other types of benchmarks for evaluating docking/scoring methods. In fact, there are other high-impact works in this field. A remarkable example is the DUD/DUD-E benchmark designed for validating virtual screening protocols.[36] Some researchers solicit unpublished data sets from pharmaceutical companies to organize blind tests of docking/scoring methods, such as the Community Structure Activity Resource (CSAR) exercise[37] and the more recent Drug Design Data Resource (D3R) project (https://drugdesigndata.org/). A blind test aims at examining the true predictive power of a docking/scoring method and thus has its unique value. Nevertheless, a blind test is usually based on small data sets containing a few particular target proteins. Due to the case-dependent performance of today's docking/scoring methods, conclusions derived from such data sets may not be transferrable to other cases. The readers should be aware of this potential limitation.

4. CONCLUDING REMARKS

It has been over 20 years since the arrival of the first wave of scoring functions. Despite the obvious progress and many successful applications made so far, the performance of today's scoring functions still does not meet people's expectation in many aspects. For example, reliable prediction of ligand binding affinity remains an unsolved problem. Many challenges faced by scoring functions, such as conformation reorganization, interfacial water molecules, and some nonclassical interactions, have been well discussed in literature.[32,33] We believe that the next generation of scoring function should be based on a deeper understanding and a more balanced account of the critical factors in protein−ligand interaction. In addition to the efforts required on the technical aspects of scoring functions, some obstacles underlying scoring function development must be removed as well. Those obstacles often do not receive enough attention, but their impact could be significant in certain circumstances.

In this Account, we have described our work on the development of the PDBbind database and the CASF benchmark. Both projects are intended to overcome the underlying obstacles implied above. PDBbind provides the knowledge basis much needed by scoring function development. With the large data set available from PDBbind, it is now possible to conduct analysis of some specific interactions with relatively low occurrence (e.g., cation−π interaction, halogen bond, and covalent binders). PDBbind is regularly updated, so new possibilities are yet to come. CASF is established as a new type of benchmark particularly for scoring function assessment, that is "scoring benchmark". With the metrics defined by CASF, scoring functions can be examined in a more elaborate manner, and the results obtained thereby provide a valuable guidance for developing more advanced methods. To use a metaphor, PDBbind provides the fuel for scoring function development; while CASF serves as a compass for navigation. Collectively, our works have helped to accelerate the progress in this field.

We are of course aware of the limitations in our current work. For example, the binding data recorded in PDBbind are obtained by different groups using various technical methods and experimental settings. The intrinsic error in such "heterogeneous" binding data[38] actually sets a theoretical upper limit on accuracy if such data are used in scoring function calibration. Moreover, some simplifications are made by us during processing protein−ligand complex structures, such as removing the water molecules bridging protein−ligand interactions, setting the protonation state of the ligand and each residue in a "default" form, and so on. A set of better methods could be employed for this task. As it comes to CASF, our current evaluation methods used in the ranking power and screening power tests are somewhat premature. Besides, a larger test set is preferred in order to obtain more robust statistical results. We are now working on another update of CASF to address these problems. To conclude, both PDBbind and CASF have come a long way in our group. We will continue to improve them in the future to keep them as useful community resources.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.accounts.6b00491.

Additional information about the PDBbind database (PDF)

AUTHOR INFORMATION

Corresponding Authors

*R. Wang. E-mail: wangrx@sioc.ac.cn.
*Y. Li. E-mail: kathyli@sioc.ac.cn.

Notes

The authors declare no competing financial interest.

Biographies

Zhihai Liu received his B.S. in life science (2003) and M.S. in philosophy of science and technology (2007) from the University of Science and Technology of China. He is currently a research associate.

Minyi Su received her B.S. in pharmaceutical science (2014) from Sun Yet-Sen University. She is currently a Ph.D. student.

Li Han received her B.S. in applied chemistry (2007) and Ph.D. in theoretical and computational chemistry (2012) from Shandong University. She is currently a research associate.

Jie Liu received his B.S. in information science (2011) from Nanjing University of Information Science and Technology and Ph.D. in chemical biology (2016) from Shanghai Institute of Organic Chemistry.

Qifan Yang received his B.S. in applied chemistry (2015) from Central South University. He is currently a graduate student.

Yan Li received her B.S. in chemistry (2005) from the University of Science and Technology of China and Ph.D. in organic chemistry (2010) from Shanghai Institute of Organic Chemistry. She is currently a research associate professor. Her research interests focus on understanding protein−ligand and protein−protein interactions with computational methods.

Renxiao Wang received his B.S. in chemistry (1994) and Ph.D. in physical chemistry (1999) from Peking University. He did postdoctoral training at University of California, Los Angeles (1999−2000), and Georgetown University (2000−2001) and worked in Prof. Shaomeng Wang's group at University of Michigan (2001−2005). Then, he joined the faculty of Shanghai Institute of Organic Chemistry, where he is currently a full professor. His research interests focus on developing computational methods for structure-based drug design and applying them to molecular targets with pharmaceutical implications.

ACKNOWLEDGMENTS

We thank all past members of Prof. Wang's group who made contributions to the work described here. We also thank the users of PDBbind and CASF for their valuable feedbacks over these years. Our work is 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. 81430083, 21472227, 21673276, 21102168, and 21402230), Chinese Academy of Sciences (Strategic Priority Research Program, Grant No. XDB20000000), and the Science and Technology Development Foundation of Macao SAR (Grant No. 055/2013/A2).

REFERENCES

(1) Talele, T. T.; et al. Successful Applications of Computer-Aided Drug Discovery: Moving Drugs from Concept to the Clinic. *Curr. Top. Med. Chem.* **2010**, *10*, 127−141.

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

(3) 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.

(4) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of Protein-Ligand Interactions. Docking and Scoring: Successes and Gaps. *J. Med. Chem.* **2006**, *49*, 5851−5855.

(5) Huang, S.-Y.; Grinter, S. Z.; Zou, X. Scoring functions and their evaluation methods for protein−ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys.* **2010**, *12*, 12899−12908.

(6) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649−663.

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

(8) 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.

(9) Englebienne, P.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 4. Are Popular Scoring Functions Accurate for this Class of Proteins? *J. Chem. Inf. Model.* **2009**, *49*, 1568−1580.

(10) Wang, R.; Lu, Y.; Wang, S. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.

(11) Wang, R.; Lu, Y.; Fang, X.; Wang, S. An Extensive Test of 14 Scoring Functions Using the PDBbind Refined Set of 800 Protein-Ligand Complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114−2125.

(12) Roche, O.; Kiyama, R.; Brooks, C. L. III. Ligand-Protein DataBase: Linking Protein-Ligand Complex Structures to Binding Data. *J. Med. Chem.* **2001**, *44*, 3592−3598.

(13) Puvanendrampillai, D.; Mitchell, J. B. O. Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein−ligand complexes. *Bioinformatics* **2003**, *19*, 1856−1857.

(14) Block, P.; Sotriffer, C. A.; Dramburg, I.; Klebe, G. AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB. *Nucleic Acids Res.* **2006**, *34*, D522−D526.

(15) Sotriffer, C. A.; Sanschagrin, P.; Matter, H.; Klebe, G. SFCscore: Scoring functions for affinity prediction of protein−ligand complexes. *Proteins: Struct., Funct., Genet.* **2008**, *73*, 395−419.

(16) Berman, H. M.; Henrick, K.; Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **2003**, *10*, 980.

(17) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.

(18) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.

(19) 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.

(20) Inhester, T.; Rarey, M. Protein−ligand interaction databases: advanced tools to mine activity data and interactions on a structural level. *WIREs Comput. Mol. Sci.* **2014**, *4*, 562−575.

(21) 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.

(22) Li, X.; Li, Y.; Cheng, T.; Liu, Z.; Wang, R. Evaluation of the Performance of Four Molecular Docking Programs on a Diverse Set of Protein-Ligand Complexes. *J. Comput. Chem.* **2010**, *31*, 2109−2125.

(23) Li, Y.; Liu, Z.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; 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.

(24) 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.

(25) Liu, Z. G.; Wang, G. T.; Li, Z. T.; Wang, R. Geometrical Preferences of the Hydrogen Bonds on Protein-Ligand Binding Interface Derived from Statistical Surveys and Quantum Mechanics Calculations. *J. Chem. Theory Comput.* **2008**, *4*, 1959−1973.

(26) Li, X.; Liu, Z.; Li, Y.; Li, J.; Li, J.; Wang, R. A Statistical Survey on the Binding Constants of Covalently Bound Protein-Ligand Complexes. *Mol. Inf.* **2010**, *29*, 87−96.

(27) Li, Y.; Liu, Z.; Wang, R. Test MM-PB/SA on True Conformational Ensembles of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2010**, *50*, 1682−1692.

(28) Cheng, T.; Liu, Z.; Wang, R. A Knowledge-guided Strategy for Improving the Accuracy of Scoring Functions in Binding Affinity Prediction. *BMC Bioinf.* **2010**, *11*, 193−208.

(29) Li, Y.; Liu, Z.; Han, L.; Li, C.; Wang, R. Mining the Characteristic Interaction Patterns on Protein-Protein Binding Interfaces. *J. Chem. Inf. Model.* **2013**, *53*, 2437−2447.

(30) Liu, Z. H.; Li, J.; Liu, J.; Liu, Y. C.; Nie, W.; Han, L.; Li, Y.; Wang, R. Cross-Mapping of Protein-Ligand Binding Data between ChEMBL and PDBbind. *Mol. Inf.* **2015**, *34*, 568−576.

(31) Ahmed, A.; Smith, R. D.; Clark, J. J.; Dunbar, J. B.; Carlson, H. A. Recent improvements to Binding MOAD: a resource for protein−ligand binding affinities and structures. *Nucleic Acids Res.* **2015**, *43*, D465−D469.

(32) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **2008**, *153*, S7−S26.

(33) Waszkowycz, B.; Clark, D. E.; Gancia, E. Outstanding challenges in protein−ligand docking and structure-based virtual screening. *WIREs Comput. Mol. Sci.* **2011**, *1*, 229−259.

(34) Ain, Q. U.; Aleksandrova, A.; Roessler, F. D.; Ballester, P. J. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *WIREs Comput. Mol. Sci.* **2015**, *5*, 405−424.

(35) Yan, Z.; Wang, J. Optimizing the affinity and specificity of ligand binding with the inclusion of solvation effect. *Proteins: Struct., Funct., Genet.* **2015**, *83*, 1632−1642.

(36) 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.

(37) Carlson, H. A. Lessons Learned over Four Benchmark Exercises from the Community Structure−Activity Resource. *J. Chem. Inf. Model.* **2016**, *56*, 951−954.

(38) Kalliokoski, T.; Kramer, C.; Vulpetti, A. Quality issues with public domain chemogenomics data. *Mol. Inf.* **2013**, *32*, 898−905.

DOI: 10.1021/acs.accounts.6b00491