Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set

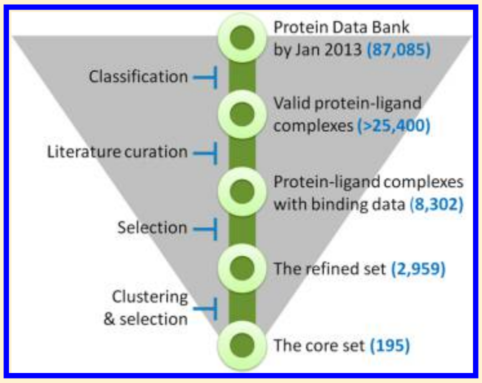
Yan Li,[†] Zhihai Liu,[†] Jie Li,[†] Li Han,[†] Jie Liu,[†] Zhixiong Zhao,[†] and Renxiao Wang*,[†,‡]

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

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

S Supporting Information

ABSTRACT: Scoring functions are often applied in combination with molecular docking methods to predict ligand binding poses and ligand binding affinities or to identify active compounds through virtual screening. An objective benchmark for assessing the performance of current scoring functions is expected to provide practical guidance for the users to make smart choices among available methods. It can also elucidate the common weakness in current methods for future improvements. The primary goal of our comparative assessment of scoring functions (CASF) project is to provide a high-standard, publicly accessible benchmark of this type. Our latest study, i.e., CASF-2013, evaluated 20 popular scoring functions on an updated set of protein–ligand complexes. This data set was selected out of 8302 protein–ligand complexes recorded in the PDBbind database (version 2013) through a fairly complicated process. Sample selection was made by considering the quality of complex structures as well as binding data. Finally, qualified complexes were clustered by 90% similarity in protein sequences. Three representative complexes were chosen from each cluster to control sample redundancy. The final outcome, namely, the PDBbind core set (version 2013), consists of 195 protein–ligand complexes in 65 clusters with binding constants spanning nearly 10 orders of magnitude. In this data set, 82% of the ligand molecules are "druglike" and 78% of the protein molecules are validated or potential drug targets. Correlation between binding constants and several key properties of ligands are discussed. Methods and results of the scoring function evaluation will be described in a companion work in this issue (doi: 10.1021/ci500081m).



1. INTRODUCTION

Simulation of protein–ligand interactions is the basis of structure-based drug design.[1−4] A variety of computational methods, such as molecular docking programs and scoring functions, have been developed for modeling protein–ligand interactions in the past.[5−11] These methods are routinely applied in both academia and the pharmaceutical industry nowadays, and many successful applications have been publicly reported. Nevertheless, it is also well-known that the performance of current docking/scoring methods is still below one's expectation in many aspects.[8−11] Although the pioneering works in this field were published almost 3 decades ago,[12] a strong motivation for improvement makes development of docking/scoring methods continue to be an active field.

In order to develop better docking/scoring methods, a clear understanding of the limitations in the current methods is indispensable. Well-designed validations can provide useful clues. In the past, many comparative assessments of docking/scoring methods have been conducted with various standards and data sets.[13−25] They typically evaluated the performance of docking/scoring methods in ligand binding pose and binding affinity prediction, or in virtual screening trials. One of the earliest comparative assessments of docking methods was conducted by Bissantz et al., who evaluated three docking programs (DOCK, FlexX, and GOLD) in combination with seven scoring functions on 10 thymidine kinase complexes and 10 estrogen receptor complexes.[13] Bursulaya et al. evaluated five docking programs (DOCK, FlexX, AutoDock, GOLD, and ICM) on 37 complexes formed by 11 different proteins.[14] There are also comparative assessments of scoring functions alone, such as the study by Wang et al.,[26] who evaluated 11 scoring functions on 100 selected protein–ligand complexes regarding their ability of identifying correct ligand binding poses and reproducing experimental binding data. After all, such comparative studies are also greatly welcome by the users of docking/scoring methods since they can make a reasonable choice among available methods accordingly.

However, those comparative studies on docking/scoring methods often have some obvious problems. A major problem is that the data sets employed by those studies are not always in high quality. It is well-known that many structures deposited in

Received: February 10, 2014
Published: April 9, 2014

the Protein Data Bank (PDB)[27] have minor or even major defects.[28−30] Structures from PDB thus must be used with care. In addition, studies on docking/scoring methods often need experimental protein−ligand binding data as well. According to our past experience, protein−ligand binding data sets quoted in public literature have miscellaneous problems too. For example, some binding data do not really have original references. Sometimes IC_{50} or EC_{50} values are not differentiated from equilibrium constants (K_d or K_i). There are even typographical errors that occurred during data transcription, e.g., "nM" mistyped as "mM". These problems in complex structures or binding data of course introduce a certain level of noise to docking/scoring evaluation results.

Another major problem in those data sets is perhaps less noticed. Those data sets are composed of either a mixture of diverse protein−ligand complexes or congeneric sets of ligands binding to a handful of target proteins. In either way, they are normally assembled in a random manner at the researchers' own convenience. Performance of current docking/scoring methods is often case-dependent; i.e., they are more successful on certain types of protein−ligand complexes but not on others. Thus, evaluation results obtained on randomly compiled data sets could be more or less biased due to the compositions of those data sets. In fact, it happens repeatedly in literature that one research group reports "docking program A is better than program B″ while another research group reports the opposite. Such confusing conclusions are often simply because different data sets were employed in evaluation.

The problems in data sets mentioned above can be largely overcome if some high-quality, representative data sets are available in this field as standard benchmarks. Several research groups have made efforts toward this direction along different approaches. One example is the Astex diverse set,[28] which consists of 85 diverse protein−ligand complex structures selected from PDB. In order to compile this data set, high-resolution protein−ligand complexes were selected from PDB that can be assessed by reconstructing the electron density for the ligand binding pose using the deposited structure factor data. Besides, the protein molecule in each complex has to be of direct interest for the pharmaceutical or agrochemical industry, and the ligand molecule has to be "druglike". Then, one complex from each remaining cluster was selected into the final data set. Later, another data set was provided by Prof. Carlson's group as part of their community structure−activity resource (CSAR) exercise.[31−34] This data set, namely, the CSAR-NRC HiQ data set, consists of 343 diverse protein−ligand complexes.[31] These complexes were selected mostly from the binding MOAD database[35,36] through multiple quality filters regarding crystal structures, binding data, and other features. Sample redundancy, however, was not addressed in the CSAR-NRC HiQ data set. A new data set, called "Iridium", was recently reported by Warren et al. at OpenEye.[29] It began as the combination of several previously described data sets, a total of 728 protein−ligand complexes in PDB. Then, it was reduced to 233 complexes by requiring available structural factor data. Among these, 121 were classified as "highly trustworthy", 104 as "mildly trustworthy", and the remaining eight as "untrustworthy" with a set of stringent criteria regarding the quality of crystal structures. Compilation of the Iridium data set employed perhaps the most complete set of criteria for evaluating crystal structures reported in the literature so far. The quality of the binding data was not seriously addressed in this study. But this defect does not impair their primary goal because this data set is prepared for validating docking methods only.

In our opinion, a good data set for evaluating docking/scoring methods should have two essential qualities: (1) The protein−ligand complexes included in this data set should have both high-quality three-dimensional structures and reliable binding data. (2) In order to assess the general performance of docking/scoring methods, this data set should consist of diverse protein−ligand complexes rather than a handful of congeneric families of complexes. Here, diversity should be fulfilled at the protein side as well as the ligand side. It will provide additional value if the protein molecules included in this data set are validated or potential drug targets. At the same time, sample redundancy should be carefully controlled to ensure that statistical results derived from this data set are not biased toward any certain types of complexes.

Creation of the PDBbind database in early 2000s was at least partly motivated by the need for such a data set. The PDBbind database aims at collecting experimental binding data for the molecular complexes deposited in PDB. It thus provides the connection between three-dimensional structures and binding data, which is essential for establishing benchmarks of docking/scoring methods. This database was originally developed in Prof. Shaomeng Wang's group at the University of Michigan and was released to the public in 2004.[37,38] In recent years, it has been maintained and further developed in our group at the Shanghai Institute of Organic Chemistry. This database is now updated on an annual base. The latest release is version 2013, which provides binding data for over 10,700 protein−ligand, protein−protein, protein−nucleic acid, and nucleic acid−ligand complexes in PDB.

The PDBbind database has served as a solid basis for us to develop a better benchmark for scoring function evaluation. We call this project comparative assessment of scoring functions (CASF). The first completed study employed several data sets selected from PDBbind (version 2007), which is referred to as CASF-2007 throughout this work. In CASF-2007,[39] a special data set, namely, the PDBbind core set, was used as the primary test set. This data set was compiled to fulfill the two desired qualities mentioned above. It consisted of 195 diverse protein−ligand complexes with high-resolution crystal structures and reliable binding constants. These complexes were selected through a systematic nonredundant sampling of the PDBbind database. Each type of protein included in this data set has the same number of samples. Based on this data set, a total of 16 popular scoring functions were evaluated in CASF-2007 in three aspects, i.e., "docking power", "scoring power", and "ranking power".[39] We have released all of the data sets used in CASF-2007 on the PDBbind-CN Web site (http://www.pdbbind-cn.org/). Thus, other researchers can utilize these data sets and follow the same metrics described by us to test the scoring functions in their interests.

As implied above, our CASF benchmark consists of three cornerstones, i.e., a set of protein−ligand complexes as the test set, a panel of popular scoring functions, and corresponding evaluation methods. Here, we report an updated study, i.e., CASF-2013. This study has the same framework as CASF-2007, but substantial improvements have been made in all three aspects. This work describes the compilation of the primary test set used in CASF-2013 because it is a relatively independent task. Basic features of this data set are analyzed. Evaluation methods and evaluation results of a panel of 20 scoring functions will be described in a companion work in this issue.[40]

Table 1. Rules for Selecting Qualified Protein−Ligand Complexes into the PDBbind Refined Set (Version 2013)

description[a]	rationale
Category I: Concerns on the Quality of Complex Structures[b]	
Only complexes with crystal structures are accepted; while complexes with NMR-resolved structures are not.	Quality of crystal structures is generally better than NMR structures.
Resolution of the complex structure must be better than 2.5 Å, and *R*-factor must be lower than 0.250.	To use common crystallography parameters to control the overall quality of the complex structure.
If any fragment on the ligand molecule is missing in the crystal structure, the complex is not accepted.	To ensure structural integrity of the ligand molecule.
If any backbone or side chain fragment is missing at the protein binding site (within 8 Å from the ligand), the complex is not accepted.	To ensure structural integrity of the binding pocket on the protein molecule.
Covalent complexes are not accepted.	Covalent complexes are usually not in pharmaceutical interests.
A noncovalent complex is not accepted if any significant steric clash (distance < 2.0 Å) exists between a pair of heavy atoms on the protein and the ligand.	To remove the complexes with obvious steric clashes.
Category II: Concerns on the Quality of Binding Data[c]	
Complexes with known dissociation constants (K_d) or inhibition constants (K_i) are accepted. Complexes with only half-inhibition or half-effect concentrations (IC_{50} or EC_{50}) values are not.	K_d and K_i are equilibrium constants, which are in theory comparable if they are derived from different binding assays. In contrast, IC_{50} or EC_{50} values are dependent on experimental settings.
Complexes with extremely low (K_d or K_i > 10 mM) or extremely high (K_d or K_i < 1 pM) binding affinities are not accepted.	Extremely low or extremely high binding data are often estimated values due to the physical limit in binding assay methods. Besides, such complexes are usually not in pharmaceutical interests.
Estimated binding data, e.g., $K_d \sim 1$ nM or $K_i > 10$ μM, are not accepted.	Estimated binding data are not good enough.
If both K_d or K_i values are available for the same complex, the K_d value is chosen as the preferred data. If binding data are obtained under multiple experimental settings for the same complex, the binding data measured at room temperature and neutral pH (or a condition closest to this) is chosen as the preferred data.	To select the preferred binding data for a given complex when multiple data are available.
If the protein molecule has multiple binding sites which are associated with significantly different binding constants (>10 folds), this complex is not accepted.	To avoid complicated cases.
The protein molecule used in binding assay has to match the one used in crystal growth, i.e., the same species, subtype, and mutation. Similarly, the ligand molecule used in the binding assay has to match the one used in crystal growth.	To match the binding data with the crystal structure.
Category III. Concerns on the Nature of the Complex[b]	
Molecular weight of the ligand molecule must be lower than 1000 if it is a regular organic molecule. It must not contain more than 10 residues if it is a polypeptide molecule, or it must not contain more than three residues if it is a polynucleotide molecule.	To control that the ligand molecule is a low-weight, “regular” organic molecule.
The ligand molecule must not consist of atoms other than carbon, nitrogen, oxygen, phosphorus, sulfur, halogen, and hydrogen atoms. The binding site on the protein molecule must not contain any nonstandard amino acid residues in direct contact with the bound ligand (distance < 5 Å).	To ensure that the complex structure can be readily processed by most molecular modeling software without encountering the “missing-parameter” problem.
Only binary complexes are accepted; i.e., the complex must be formed distinctly between one protein molecule and one ligand molecule. Ternary complexes are not accepted, e.g., a cofactor and a substrate binding closely (distance < 5 Å) at the same site on the protein.	To avoid complicated cases. Binding data can be defined in multiple ways in the case of a ternary complex. Such details are not always clearly indicated in the literature.
If the buried surface area of the ligand molecule is below 15% of its total surface area, this complex was not accepted.	To remove the complexes in which the ligand has little physical contact with the protein, i.e., “floating” ligand.

[a] New rules introduced since version 2007 are indicated in underlined text. [b] These examinations were conducted with computer programs to ensure accuracy and completeness. [c] These examinations were conducted manually by referring to the original references from which the binding data were curated.

2. METHODS

The basis of our CASF benchmark is a set of protein−ligand complexes with known three-dimensional structures and binding data. A new version of the PDBbind core set, i.e., version 2013, was employed as this data set. Similar to all previous versions, this data set was compiled through a two-step process. But the methods used in this process have been updated in many aspects since CASF-2007. It is thus necessary to describe this process here in sufficient details.

2.1. Selection of Good-Quality Protein−Ligand Complexes from the PDBbind Database: The Refined Set. The primary test set used in our benchmark was selected out of the PDBbind database. This database now collects experimentally measured binding data for all major categories of biomolecular complexes deposited in the Protein Data Bank, including protein−small-ligand complexes, protein−protein complexes, protein−nucleic acid complexes, and nucleic acid−small-ligand complexes. The latest release, i.e., version 2013, provides binding data for over 10,700 complexes. Among them, a total of 8,302 complexes are formed between protein molecules and small-molecule ligands. Here, a valid small-molecule ligand must contain at least six non-hydrogen atoms. It cannot be an organic solvent, a buffer component, or an inorganic molecule. Polypeptides with fewer than 10 residues and polynucleotides with fewer than four residues were considered as valid small-molecule ligands. Besides, four classes of cofactors/coenzymes, including coenzyme A (CoA), nicotinamide adenine (NAD), flavin adenine (FAD), and heme as well as their derivatives were not considered as valid ligand molecules.

Apparently, not every protein−ligand complex in PDBbind has the desired quality to be included in our benchmark. We therefore applied a set of rules to select the qualified candidates. These rules reflect our concerns on the quality of the complex structures, the quality of the binding data, and the nature of the complex, which are summarized in Table 1. The protein−ligand complexes that passed all of these filters, 2,959 in total, were referred to as the PDBbind "refined set". They form the basis for selecting the final test set used in our benchmark.

2.2. Sampling of Representative Protein−Ligand Complexes: The Core Set. The PDBbind refined set, however, cannot be used directly in our benchmark. First, its size is too large for this purpose. More importantly, there is considerable redundancy in its contents because some types of proteins are simply more popular than the others in PDB. For example, nearly 10% of the protein−ligand complexes in the refined set are formed by HIV-1 protease. In order to remove sample redundancy, a systematic sampling on the PDBbind refined set was performed to select the representative ones into the PDBbind core set. Here, the new rules introduced since CASF-2007 are indicated in underlined text.

At the first step, all protein−ligand complexes in the refined set were grouped into clusters by protein sequence similarity. For this task, sequence similarity was computed with the CD-hit program (version 4.0, http://www.cd-hit.org).[41] This program is also used by PDB for computing sequence similarity. A similarity cutoff of 90% was used in clustering. As a result, each cluster was typically composed of complexes formed by the same type of protein. The clusters with fewer than five members were ignored because they did not provide enough samples for subsequent selection.

At the second step, in each remaining cluster, the complex with the highest binding constant (log $K_{a,max}$), the one with the lowest binding constant (log $K_{a,min}$), and the one with a binding constant equal to or close to the mean values of log $K_{a,max}$ and log $K_{a,min}$ were selected as the representatives of this cluster. For the sake of convenience, these three complexes will be referred to as "the best", "the median", and "the poorest" in this work. Here, the binding constants of the best complex and the median complex must differ by at least 10-fold, and the binding constants of the median complex and the poorest complex must differ by at least 10-fold. Consequently, the binding constants of the best complex and the poorest complex in each cluster differ by at least 100-fold. Selection of these three complexes aims at maximizing the binding affinity range in each cluster.

At the third step, electron density maps of the candidate complexes selected at the previous step were examined visually as an additional quality control on crystal structures. Structure factors of all candidate complexes were downloaded from PDB. Electron density maps were displayed with the COOT program (version 0.7.1, http://www.biop.ox.ac.uk/coot/). Our visual examination focused on the ligand binding pose as well as the nearby residues. If any of the following situations were observed, the complex was rejected: (i) Electron density is missing for a major part of the ligand structure, or the overall fitting of the ligand structure to the density map is rather poor. (ii) There are a considerable number of "positive" and "negative" density regions around the ligand structure. (iii) Two alternative ligand binding poses fit to the density map equally well. (iv) Structure factor data are not available for the given complex structure. Note that most structures resolved in early years were deposited into PDB without structure factor data. Although they are not necessarily problematic, we did not consider them to keep a high standard.

If a complex structure failed at the third step, another appropriate candidate, e.g., the one with the second highest binding constant, was selected from the same cluster for examination. This loop was repeated until three suitable complexes meeting all requirements were selected out for each cluster. The final data set, namely, the PDBbind core set, consists of 195 protein−ligand complexes in 65 clusters.

2.3. Process of Protein−Ligand Complex Structures. Coordinates of the 195 protein−ligand complexes in the test set were all downloaded from the PDB official site (http://www.rcsb.org/pdb/). The original structural files from PDB were processed so that they could be readily utilized by the software considered in our study. Briefly, each complex in a complete "biological unit" was split into a ligand molecule and a protein molecule. Definition of the biological unit of each complex was also taken from PDB. Atomic types and bond types of the ligand molecule were automatically assigned by using the I-interpret program,[42] and then were visually examined and corrected. Hydrogen atoms were added to the protein and the ligand by using the SYBYL software. Both the protein and the ligand were set in a simple protonation scheme assuming a neutral pH: all carboxylic acid and phosphate groups were deprotonated, while all aliphatic amine, guanidino, and amidino groups were protonated. The protein was assigned the AMBER FF99 charges; while the ligand was assigned the Gasteiger−Hückel partial charges. Metal ions, if they reside inside the binding pocket and bridge the binding of ligand and protein, were treated as part of the protein molecule. Water molecules included in the original crystal structure were kept with the protein molecule. Finally, the processed protein structure was saved in a PDB-format file, and the processed ligand structure was saved in a Mol2-format and a SD-format file. In the above process, no structural optimization was conducted on either the protein molecule or the ligand molecule to retain their original coordinates from PDB.

2.4. Computation of Key Properties. A few key properties related to protein−ligand binding were computed to characterize the protein−ligand complexes in the test set.

The first property is the buried solvent-accessible surface area of the ligand molecule upon binding. The Richards−Lee solvent-accessible surface of each ligand molecule was computed using an in-house computer program. This type of surface is the complete trajectory of a probe rolling on the van der Waals surface of a given molecule.[43] In our computation, a probe radius of 1.0 Å and a dot density of 4 dots/Å^2 were used in surface generation. A surface dot on the ligand molecule was considered to be buried upon binding if it is enclosed in the solvent-accessible surface of the protein molecule (Figure 1). Integration of all such dots gave the buried area of the ligand molecule. The atomic radii used in our computation were as follows (which were cited from the classical work of Bondi[44]): C, 1.70 Å; N, 1.55 Å; O, 1.52 Å; P, 1.80 Å; S, 1.80 Å; F, 1.47 Å; Cl, 1.75 Å; Br, 1.85 Å; *I*, 1.98 Å; H, 1.20 Å.

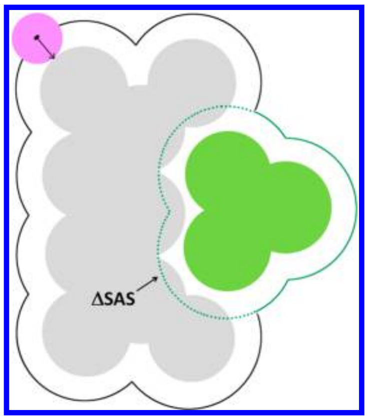


Figure 1. Definition of the solvent-accessible surface area of a ligand molecule buried upon binding to the target protein (the sector in dashed line). The radius of the solvent probe is set to 1.0 Å. The volume of the ligand molecule is computed as the space enclosed in the solvent-accessible surface.

The second property is the buried volume of the ligand molecule upon binding. In order to compute this property, a cubic lattice with an even spacing of 0.5 Å along all three dimensions was generated to enclose the whole ligand molecule. The total volume of the ligand molecule was computed by integrating all of the grids enclosed within its solvent-accessible surface. An "emitting-vector" algorithm is implemented in our program to judge if a certain grid is inside the binding pocket or not. This algorithm examines a set of 12 vectors emitting from the given point. These vectors are oriented according to the positions of the 12 vertices on an icosagon to ensure that they are evenly scattered in the three-dimensional space. Each vector is then examined if it is blocked by any protein atom within 8.0 Å. If more than half of the vectors (i.e., six) are blocked by the protein, the given gird is considered to be inside the binding pocket; otherwise it is considered to be outside the binding pocket. Integration of all such grids gave the buried volume of the ligand molecule. Note that this property equals to the volume inside the binding pocket that is excluded upon ligand binding.

In addition, five key chemical descriptors of the ligand molecules, including molecular weight, the number of hydrogen bond donor atoms, the number of hydrogen bond acceptor atoms, number of rotatable bonds, and the octanol−water partition coefficient (log *P*), were also computed. The first three descriptors can be computed straightforwardly based on the chemical structure of each ligand molecule. Rotatable bond was defined as an acyclic sp3−sp3 or sp3−sp2 single bond between two non-hydrogen atoms. Single bonds connecting terminal groups, such as -CH_3, -NH_2, -OH, and -X (X = halogen atoms), whose rotation does not produce new rearrangement of heavy atoms are not counted as rotors. log *P* values were computed using the XLOGP3 program.[45]

3. RESULTS AND DISCUSSION

3.1. Basic Features of the Data Set. The flowchart for selection of the PDBbind core set is illustrated in Figure 2. We started with over 87,000 structures released by PDB in the first week of 2013. These structures were analyzed by a set of computer programs to identify the molecular complexes among them. Experimentally measured binding data for over 8,300 complexes formed between protein and small-molecule ligand were collected from the literature. Among them, 2,959 complexes were selected into the refined set by a set of filters regarding the quality of complex structures and binding data. The entire refined set was then clustered by the sequence of the protein molecule in each complex. Finally, a total of 195 protein−ligand complexes were selected out from 65 clusters, where each cluster is represented by three members. In brief, this data set is the outcome of a systematic, nonredundant sampling of the protein−ligand complexes in the PDBbind database.

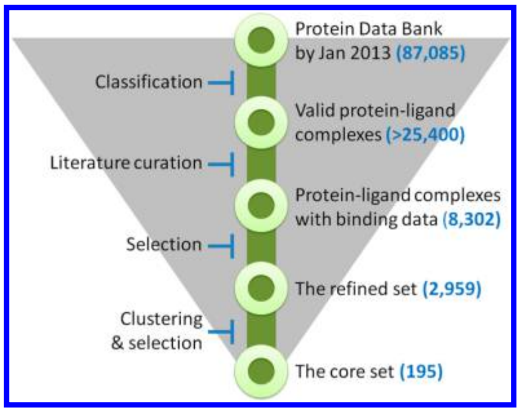

Protein Data Bank by Jan 2013 (87,085)
Classification
Valid protein-ligand complexes (>25,400)
Literature curation
Protein-ligand complexes with binding data (8,302)
Selection
The refined set (2,959)
Clustering & selection
The core set (195)

Figure 2. Flowchart of how the PDBbind core set is compiled.

Basic information on the protein−ligand complexes in this data set, including PDB codes, experimental binding data, protein names, and EC numbers, are listed in Table 2. Binding constants (K_a) of the protein−ligand complexes in this data set range from 2.00 to 11.85 (in log units), spanning over nearly 10 magnitudes (Figure 3). To make a comparison, distribution of the binding constants in the refined set is also shown in Figure 3. One can see that the core set, as a subset of the refined set, covers the entire binding constant range of the refined set. Mean values of the binding constants in the core set and the refined set are very close. However, sample distribution in the core set is relatively flat between log K_a = 4−9, while the samples in the refined set exhibit roughly a normal distribution with a peak at log K_a = 6−7;. Moreover, the percentage of low-affinity (log K_a < 4) and high-affinity (log K_a > 9) complexes in the core set is obviously higher than the refined set. This difference reflects our intention of compiling a more diverse data set with a flatter

Table 2. Summary of the Protein−Ligand Complexes Included in the PDBbind Core Set (Version 2013)

PDB code	log K_a	EC no.	protein name	druggability[a]
1PS3	2.28	E.C.3.2.1.114	α-mannosidase II	II
3D4Z	4.89	E.C.3.2.1.114	α-mannosidase II	
3EJR	8.57	E.C.3.2.1.114	α-mannosidase II	
2QMJ	4.21	E.C.3.2.1.20	maltase-glucoamylase, intestinal	I
3L4W	6.00	E.C.3.2.1.20	maltase-glucoamylase, intestinal	
3L4U	7.52	E.C.3.2.1.20	maltase-glucoamylase, intestinal	
3L7B	2.40	E.C.2.4.1.1	glycogen phosphorylase, muscle form	II
3G2N	4.09	E.C.2.4.1.1	glycogen phosphorylase, muscle form	
3EBP	5.91	E.C.2.4.1.1	glycogen phosphorylase, muscle form	
2W66	4.05	E.C.3.2.1.169	*O*-glcnacase BT_4395	I
2WCA	5.60	E.C.3.2.1.169	*O*-glcnacase BT_4395	
2VVN	7.30	E.C.3.2.1.169	*O*-glcnacase BT_4395	
2X97	5.66	E.C.3.4.15.1	angiotensin converting enzyme	I
2XHM	6.80	E.C.3.4.15.1	angiotensin converting enzyme	
2X8Z	7.96	E.C.3.4.15.1	angiotensin converting enzyme	
2X0Y	4.60	E.C.3.2.1.169	*O*-glcnacase NAGJ	II
2CBJ	8.27	E.C.3.2.1.169	*O*-glcnacase NAGJ	
2J62	11.34	E.C.3.2.1.169	*O*-glcnacase NAGJ	
3BKK	6.08	E.C.3.4.15.1	angiotensin converting enzyme	I
3L3N	8.18	E.C.3.4.15.1	angiotensin converting enzyme	
2XY9	9.19	E.C.3.4.15.1	angiotensin converting enzyme	
1GPK	5.37	E.C.3.1.1.7	acetylcholinesterase	I
1H23	8.35	E.C.3.1.1.7	acetylcholinesterase	
1E66	9.89	E.C.3.1.1.7	acetylcholinesterase	
3CJ2	4.85	E.C.2.7.7.48	RNA-dependent RNA polymerase	I
2D3U	6.92	E.C.2.7.7.48	RNA-dependent RNA polymerase	
3GNW	9.10	E.C.2.7.7.48	RNA-dependent RNA polymerase	
3F3A	4.19	NA	transporter	I
3F3C	6.02	NA	transporter	
3F3E	7.70	NA	transporter	
4GQQ	2.89	E.C.3.2.1.1	α-amylase	I
1U33	4.60	E.C.3.2.1.1	α-amylase	
1XD0	7.12	E.C.3.2.1.1	α-amylase	
2WBG	4.45	E.C.3.2.1.21	β-glucosidase A	II
2J78	6.42	E.C.3.2.1.21	β-glucosidase A	
2CET	8.02	E.C.3.2.1.21	β-glucosidase A	
2ZXD	5.22	E.C.3.2.1.51	α-L-fucosidase	II
2ZWZ	7.79	E.C.3.2.1.51	α-L-fucosidase	
2ZX6	10.60	E.C.3.2.1.51	α-L-fucosidase	
3UDH	2.85	E.C.3.4.23.46	β-secretase 1	II
4DJV	6.72	E.C.3.4.23.46	β-secretase 1	
4GID	10.77	E.C.3.4.23.46	β-secretase 1	
3FK1	2.62	E.C.2.5.1.19	3-phosphoshikimate 1-carboxyvinyltransferase	II
2QFT	5.26	E.C.2.5.1.19	3-phosphoshikimate 1-carboxyvinyltransferase	
2PQ9	8.11	E.C.2.5.1.19	3-phosphoshikimate 1-carboxyvinyltransferase	
1F8D	3.40	E.C.3.2.1.18	neuraminidase	I
1F8B	5.40	E.C.3.2.1.18	neuraminidase	
1F8C	7.40	E.C.3.2.1.18	neuraminidase	
1N2V	4.08	E.C.2.4.2.29	queuine tRNA-ribosyltransferase	II
1R5Y	6.46	E.C.2.4.2.29	queuine tRNA-ribosyltransferase	
3GE7	8.70	E.C.2.4.2.29	queuine tRNA-ribosyltransferase	
3HUC	5.99	E.C.2.7.11.24	mitogen-activated protein kinase 14	II
3GCS	7.25	E.C.2.7.11.24	mitogen-activated protein kinase 14	
3E93	8.85	E.C.2.7.11.24	mitogen-activated protein kinase 14	
1Q8T	4.76	E.C.2.7.11.11	cAMP-dependent protein kinase	II
1Q8U	5.96	E.C.2.7.11.11	cAMP-dependent protein kinase	
3AG9	8.05	E.C.2.7.11.11	cAMP-dependent protein kinase	
3OWJ	6.07	E.C.2.7.11.1	casein kinase II, α subunit	II
2ZJW	7.70	E.C.2.7.11.1	casein kinase II, α subunit	
3PE2	9.76	E.C.2.7.11.1	casein kinase II, α subunit	
2V00	3.66	E.C.3.4.23.22	endothiapepsin	III

Table 2. continued

PDB code	log K_a	EC no.	protein name	druggability[a]
3PWW	7.32	E.C.3.4.23.22	endothiapepsin	
3URI	9.00	E.C.3.4.23.22	endothiapepsin	
3MFV	2.52	E.C.3.5.3.1	arginase-1	I
3F80	4.22	E.C.3.5.3.1	arginase-1	
3KV2	7.32	E.C.3.5.3.1	arginase-1	
2HB1	3.80	E.C.3.1.3.48	protein-tyrosine phosphatase 1b	I
2QBR	6.33	E.C.3.1.3.48	protein-tyrosine phosphatase 1b	
2QBP	8.40	E.C.3.1.3.48	protein-tyrosine phosphatase 1b	
3FCQ	2.77	E.C.3.4.24.27	thermolysin	II
1OS0	6.03	E.C.3.4.24.27	thermolysin	
4TMN	10.17	E.C.3.4.24.27	thermolysin	
3PXF	4.43	E.C.2.7.11.22	cell division protein kinase 2	I
2XNB	6.83	E.C.2.7.11.22	cell division protein kinase 2	
2FVD	8.52	E.C.2.7.11.22	cell division protein kinase 2	
1QI0	2.35	E.C.3.2.1.4	endoglucanase B	II
1W3K	4.30	E.C.3.2.1.4	endoglucanase 5A	
1W3L	6.28	E.C.3.2.1.4	endoglucanase 5A	
3IMC	2.96	E.C.6.3.2.1	pantothenate synthetase	III
3IVG	4.30	E.C.6.3.2.1	pantothenate synthetase	
3COY	6.02	E.C.6.3.2.1	pantothenate synthetase	
3B3S	2.55	E.C.3.4.11.10	leucyl aminopeptidase	II
3B3W	4.19	E.C.3.4.11.10	leucyl aminopeptidase	
3VH9	6.24	E.C.3.4.11.10	leucyl aminopeptidase	
3MSS	4.66	E.C.2.7.10.2	tyrosine-protein kinase ABL1	I
3K5 V	6.30	E.C.2.7.10.2	tyrosine-protein kinase ABL1	
2V7A	8.30	E.C.2.7.10.2	tyrosine-protein kinase ABL1	
2BRB	4.86	E.C.2.7.11.1	serine/threonine-protein kinase Chk1	II
3JVS	6.54	E.C.2.7.11.1	serine/threonine-protein kinase Chk1	
1NVQ	8.25	E.C.2.7.11.1	serine/threonine-protein kinase Chk1	
3ACW	4.76	E.C.2.5.1.96	dehydrosqualene synthase	II
2ZCR	6.87	E.C.2.5.1.96	dehydrosqualene synthase	
2ZCQ	8.82	E.C.2.5.1.96	dehydrosqualene synthase	
1BCU	3.28	E.C.3.4.21.5	thrombin	I
1OYT	7.24	E.C.3.4.21.5	thrombin	
3UTU	10.92	E.C.3.4.21.5	thrombin	
3U9Q	4.38	NA	peroxisome proliferator-activated receptor γ	I
2YFE	6.63	NA	peroxisome proliferator-activated receptor γ	
2P4Y	9.00	NA	peroxisome proliferator-activated receptor γ	
3UO4	6.52	E.C.2.7.11.1	serine/threonine-protein kinase 6	III
2WTV	8.74	E.C.2.7.11.1	serine/threonine-protein kinase 6	
3MYG	10.70	E.C.2.7.11.1	serine/threonine-protein kinase 6	
3KGP	2.57	E.C.3.4.21.73	urokinase-type plasminogen activator	I
1O5B	5.77	E.C.3.4.21.73	urokinase-type plasminogen activator	
1SQA	9.21	E.C.3.4.21.73	urokinase-type plasminogen activator	
3KWA	4.08	E.C.4.2.1.1	carbonic anhydrase II	I
2WEG	6.50	E.C.4.2.1.1	carbonic anhydrase II	
3DD0	9.00	E.C.4.2.1.1	carbonic anhydrase II	
2XDL	3.10	NA	heat shock protein Hsp90-α	I
1YC1	6.17	NA	heat shock protein Hsp90-α	
2YKI	9.46	NA	heat shock protein Hsp90-α	
1P1Q	4.89	NA	glutamate receptor 2	I
3BFU	6.27	NA	glutamate receptor 2	
4G8M	7.89	NA	glutamate receptor 2	
3G2Z	2.36	E.C.3.5.2.6	β-lactamase	I
4DE2	4.12	E.C.3.5.2.6	β-lactamase	
4DE1	5.96	E.C.3.5.2.6	β-lactamase	
1VSO	4.72	NA	glutamate receptor, ionotropic kainate 1	I
3GBB	6.90	NA	glutamate receptor, ionotropic kainate 1	
3FV1	9.30	NA	glutamate receptor, ionotropic kainate 1	
2Y5H	5.79	E.C.3.4.21.6	coagulation factor XA	I
2XBV	8.43	E.C.3.4.21.6	coagulation factor XA	

Table 2. continued

PDB code	log K_a	EC no.	protein name	druggability[a]
1MQ6	11.15	E.C.3.4.21.6	coagulation factor XA	
1LOQ	3.70	E.C.4.1.1.23	orotidine 5′-monophosphate decarboxylase	III
1LOL	6.39	E.C.4.1.1.23	orotidine 5′-monophosphate decarboxylase	
1LOR	11.06	E.C.4.1.1.23	orotidine 5′-monophosphate decarboxylase	
1UTO	2.27	E.C.3.4.21.4	trypsin *β*	III
3GY4	5.10	E.C.3.4.21.4	trypsin *β*	
1O3F	7.96	E.C.3.4.21.4	trypsin *β*	
2YGE	5.06	NA	heat shock protein Hsp82	III
2IWX	6.68	NA	heat shock protein Hsp82	
2VW5	8.52	NA	heat shock protein Hsp82	
2YMD	3.16	NA	acetylcholine receptor	I
2XYS	7.42	NA	acetylcholine receptor	
2X00	11.33	NA	acetylcholine receptor	
2R23	3.72	NA	antibody FAB fragment	III
3BPC	4.80	NA	antibody FAB fragment	
1KEL	7.28	NA	antibody FAB fragment	
3OZT	4.13	E.C.2.1.1.6	catechol *O*-methyltransferase	I
3OE5	6.88	E.C.2.1.1.6	catechol *O*-methyltransferase	
3NW9	9.00	E.C.2.1.1.6	catechol *O*-methyltransferase	
1ZEA	5.22	NA	antibody FAB fragment	III
2PCP	8.70	NA	antibody FAB fragment	
1IGJ	10.00	NA	antibody FAB fragment	
1LBK	3.18	E.C.2.5.1.18	glutathione *S*-transferase P1-1	I
2GSS	4.94	E.C.2.5.1.18	glutathione *S*-transferase P1-1	
10GS	6.40	E.C.2.5.1.18	glutathione *S*-transferase P1-1	
3SU5	5.58	E.C.3.4.21.98	NS3/4A protease	I
3SU2	7.35	E.C.3.4.21.98	NS3/4A protease	
3SU3	9.13	E.C.3.4.21.98	NS3/4A protease	
3N7A	3.70	E.C.4.2.1.10	3-dehydroquinate dehydratase	III
3N86	5.64	E.C.4.2.1.10	3-dehydroquinate dehydratase	
2XB8	7.59	E.C.4.2.1.10	3-dehydroquinate dehydratase	
3AO4	2.07	E.C.2.7.7.0	HIV-1 integrase	III
3ZSX	3.28	E.C.2.7.7.0	HIV-1 integrase	
3ZSO	5.12	E.C.2.7.7.0	HIV-1 integrase	
3NQ3	3.78	NA	*β*-lactoglobulin	III
3UEU	5.24	NA	*β*-lactoglobulin	
3UEX	6.92	NA	*β*-lactoglobulin	
3LKA	2.82	E.C.3.4.24.65	macrophage metalloelastase (MMP-12)	I
3EHY	5.85	E.C.3.4.24.65	macrophage metalloelastase (MMP-12)	
3F17	8.63	E.C.3.4.24.65	macrophage metalloelastase (MMP-12)	
3CFT	4.19	NA	transthyretin	III
4DES	5.85	NA	transthyretin	
4DEW	7.00	NA	transthyretin	
3DXG	2.40	E.C.3.1.27.5	ribonuclease A	I
1W4O	5.22	E.C.3.1.27.5	ribonuclease A	
1U1B	7.80	E.C.3.1.27.5	ribonuclease A	
3OV1	5.20	NA	growth factor receptor-bound protein 2	I
3S8O	6.85	NA	growth factor receptor-bound protein 2	
1JYQ	8.70	NA	growth factor receptor-bound protein 2	
1A30	4.30	E.C.3.4.23.16	HIV-1 protease	I
3CYX	8.00	E.C.3.4.23.16	HIV-1 protease	
4DJR	11.52	E.C.3.4.23.16	HIV-1 protease	
3I3B	2.23	E.C.3.2.1.23	*β*-galactosidase	II
3MUZ	3.46	E.C.3.2.1.23	*β*-galactosidase	
3VD4	4.82	E.C.3.2.1.23	*β*-galactosidase	
2VO5	4.89	E.C.3.2.1.25	*β*-mannosidase	I
2VL4	6.01	E.C.3.2.1.25	*β*-mannosidase	
2VOT	7.14	E.C.3.2.1.25	*β*-mannosidase	
1N1M	5.70	E.C.3.4.14.5	dipeptidyl peptidase 4	I
2OLE	7.25	E.C.3.4.14.5	dipeptidyl peptidase 4	
3NOX	8.66	E.C.3.4.14.5	dipeptidyl peptidase 4	

Table 2. continued

PDB code	log K_a	EC no.	protein name	druggability[a]
1HNN	6.24	E.C.2.1.1.28	phenylethanolamine *N*-methyltransferase	III
2G70	7.77	E.C.2.1.1.28	phenylethanolamine *N*-methyltransferase	
2OBF	8.85	E.C.2.1.1.28	phenylethanolamine *N*-methyltransferase	
1Z95	7.12	NA	androgen receptor	I
3B68	8.40	NA	androgen receptor	
3G0W	9.52	NA	androgen receptor	
1SLN	6.64	E.C.3.4.24.17	stromelysin-1	III
2D1O	7.70	E.C.3.4.24.17	stromelysin-1	
1HFS	8.70	E.C.3.4.24.17	stromelysin-1	
2JDY	4.37	NA	fucose-binding lectin PA-IIL	II
2JDM	5.40	NA	fucose-binding lectin PA-IIL	
2JDU	6.72	NA	fucose-binding lectin PA-IIL	

[a] Potential as drug target: class *I* = targets of marketed drugs; class II = targets of drug candidates currently in clinical trials; class III = proteins without known compounds in clinical trials.

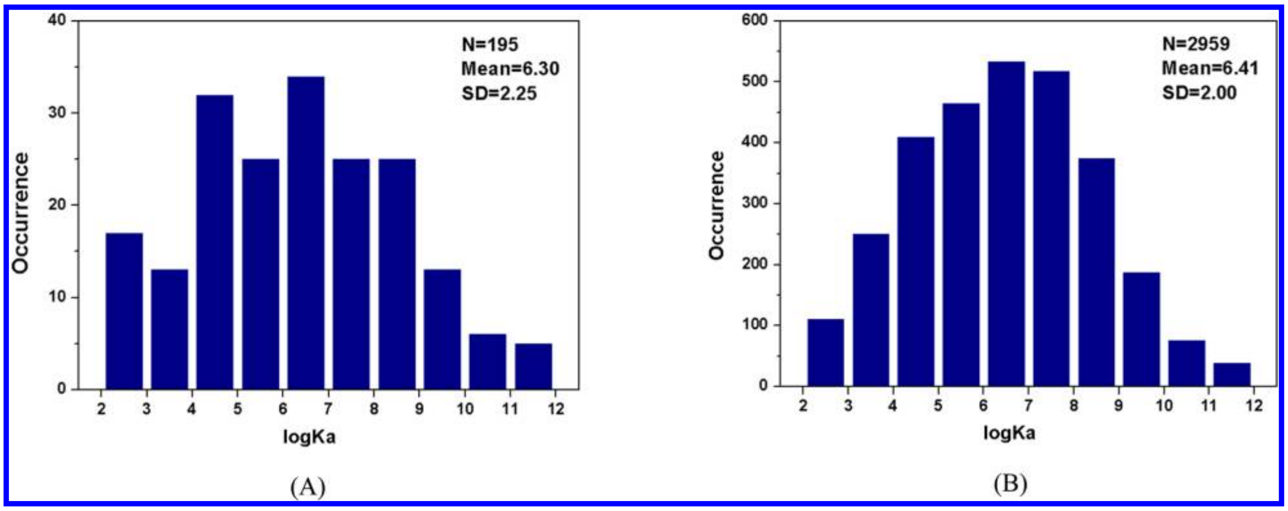


Figure 3. Distributions of binding constants of the protein–ligand complexes in (A) the PDBbind core set and (B) the PDBbind refined set.

sample distribution rather than a typical normal distribution around a certain point.

Distributions of five key chemical descriptors of the ligand molecules in the PDBbind core set, including molecular weight, the number of rotatable bonds, the number of hydrogen bond donor atoms, the number of hydrogen bond acceptor atoms, and the computed log *P* value, are also given in Figure 4. One can see from the distribution of molecular weights that the majority of the ligand molecules included in our data set have molecular weights lower than 500. A number of larger ligand molecules are also included, many of which are peptide molecules. Distributions of the other four properties also indicate that most ligand molecules in this data set are "regular" small-molecule compounds. Notably, although "drug-likeness" or "druggability" was not our primary concern, this data set turns out to be largely so. In fact, 82% of the ligand molecules in this data set (160 in 195) are in accordance with Lipinski's "rules of five".[46] To make a comparison, we retrieved the information on 1489 approved small-molecule drugs from DrugBank,[47] and then computed the same set of properties of them. Our results indicate that distributions of these five properties in two data sets are very close (see the Supporting Information, Figure S1)

With regard to the protein molecules included in our data set, the 65 proteins in this data set include 51 enzymes, six receptor or receptor-binding proteins, one sugar-binding protein, three transport proteins, two chaperon proteins, and two antibodies. By examining the information from DrugBank, we found that 33 proteins (51%) are known drug targets, for which some small-molecule drugs are already on the market, such as the angiotensin converting enzyme (ACE), tyrosine-protein kinase ABL1, carbonic anhydrase II, and HIV-1 protease (Table 2). Another 18 proteins (27%) are potential drug targets, for which some small-molecule compounds are being tested as drug candidates in clinical trials. No record was found by us for the remaining 14 proteins (22%) to indicate that they have potential pharmaceutical applications.

3.2. Ligand Efficiency. Medicinal chemists often notice that larger molecules tend to have higher binding affinities to target protein, especially when van der Waals interaction or hydrophobic effect is the dominant factor in drug–target interactions. But larger molecules may have problems in terms of solubility or absorption/distribution properties, which are also critical for the success of a drug candidate. In order to reveal the true value of an active compound by offsetting the size–affinity correlation, the concept of "ligand efficiency"[48−50] has been very popular in recent years.

Ligand efficiency is normally defined as the ligand binding affinity per number of non-hydrogen atoms. Considering the physical basis of protein–ligand interactions, it is assumed that there is an upper limit of ligand efficiency. This assumption was first described by Kuntz et al. in the late 1990s through a statistical survey on 159 protein–ligand complexes with known

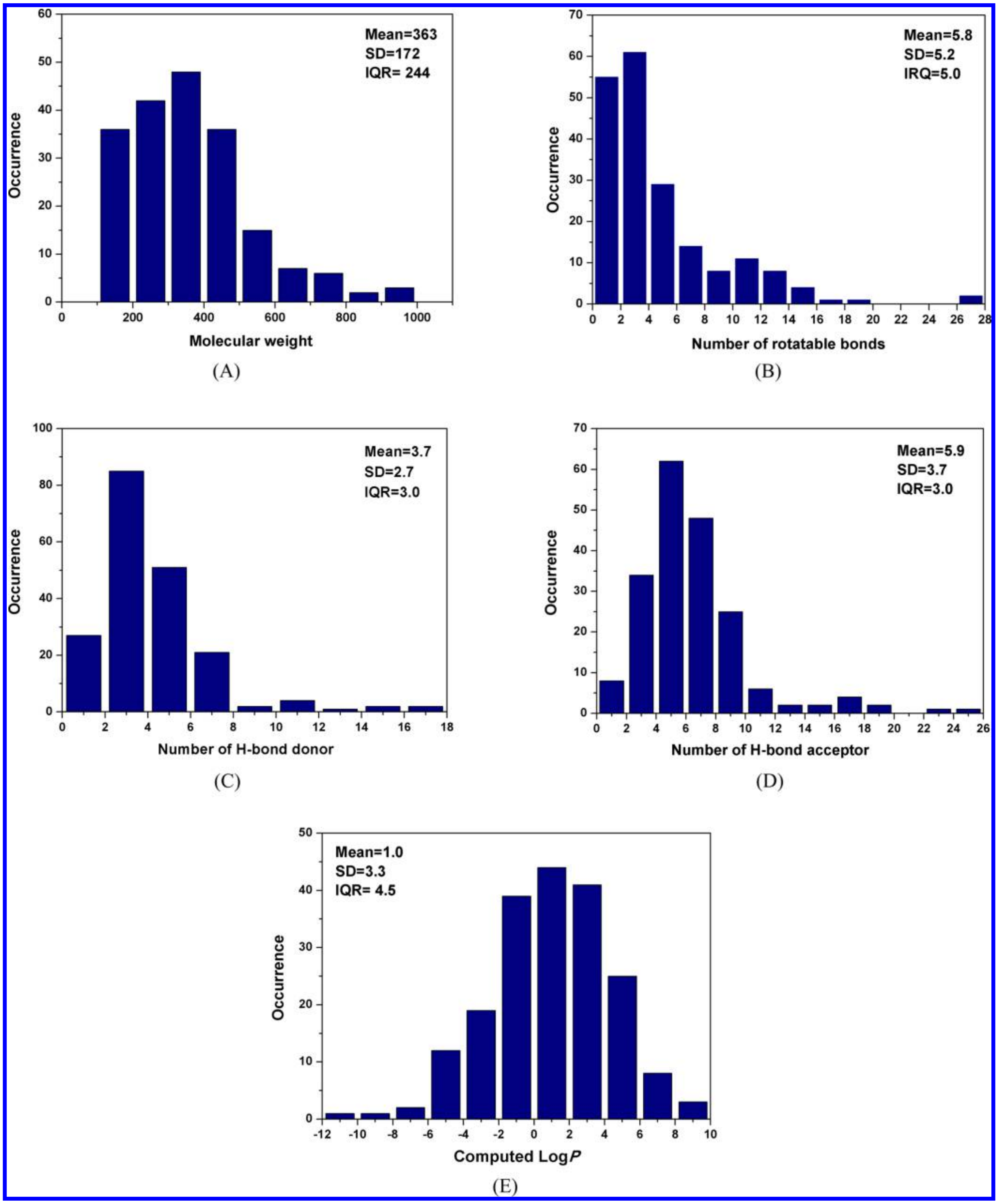


Figure 4. Distributions of five key properties of the ligand molecules in the PDBbind core set: (A) molecular weight; (B) number of rotatable bonds; (C) number of hydrogen bond donors; (D) number of hydrogen bond acceptors; (E) computed log *P* values. The mean value, the standard deviation (SD), and the interquartile range (IQR) of each property are given.

binding data.[51] They reported that each non-hydrogen atom on the ligand molecule can contribute up to −1.5 kcal/mol of binding energies. Later, Carlson et al. conducted an extended analysis on 2298 protein−ligand complexes in their binding MOAD databases.[52,53] They reported that the upper limit of ligand efficiency is −1.75 kcal/mol per atom, and 95% of their

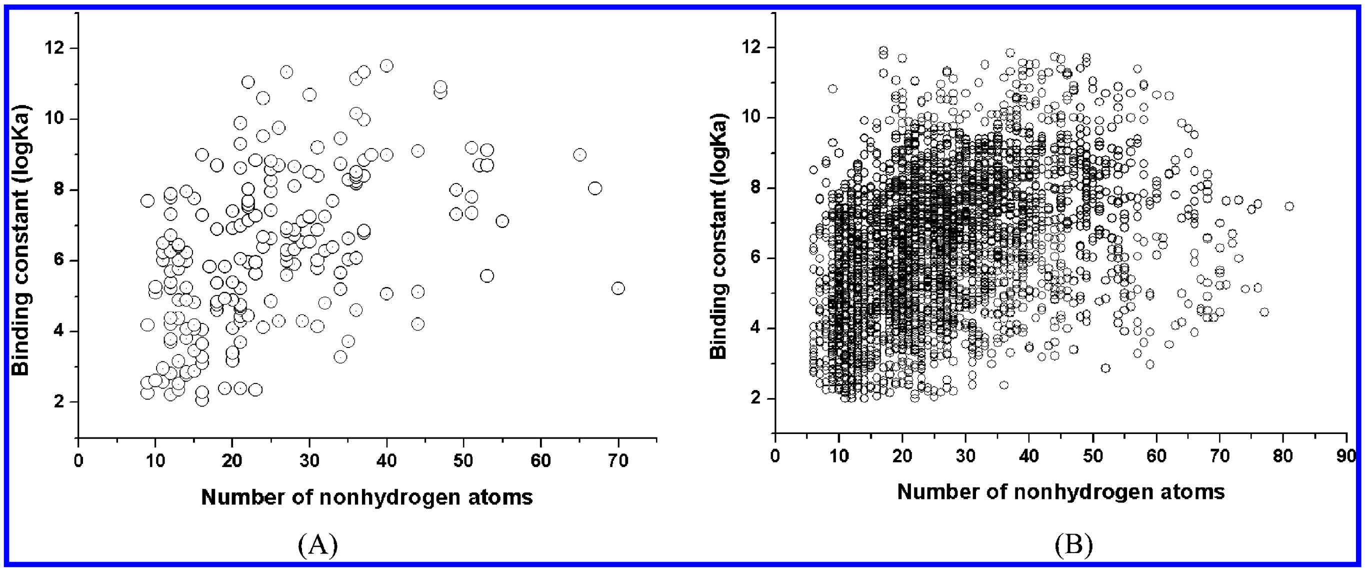


Figure 5. Scatter plot of binding constant versus non-hydrogen atoms on the ligand molecule for the protein−ligand complexes included in (A) the PDBbind core set (N = 195, R^2= 0.229) and (B) the PDBbind refined set (N = 2959, R^2= 0.126).

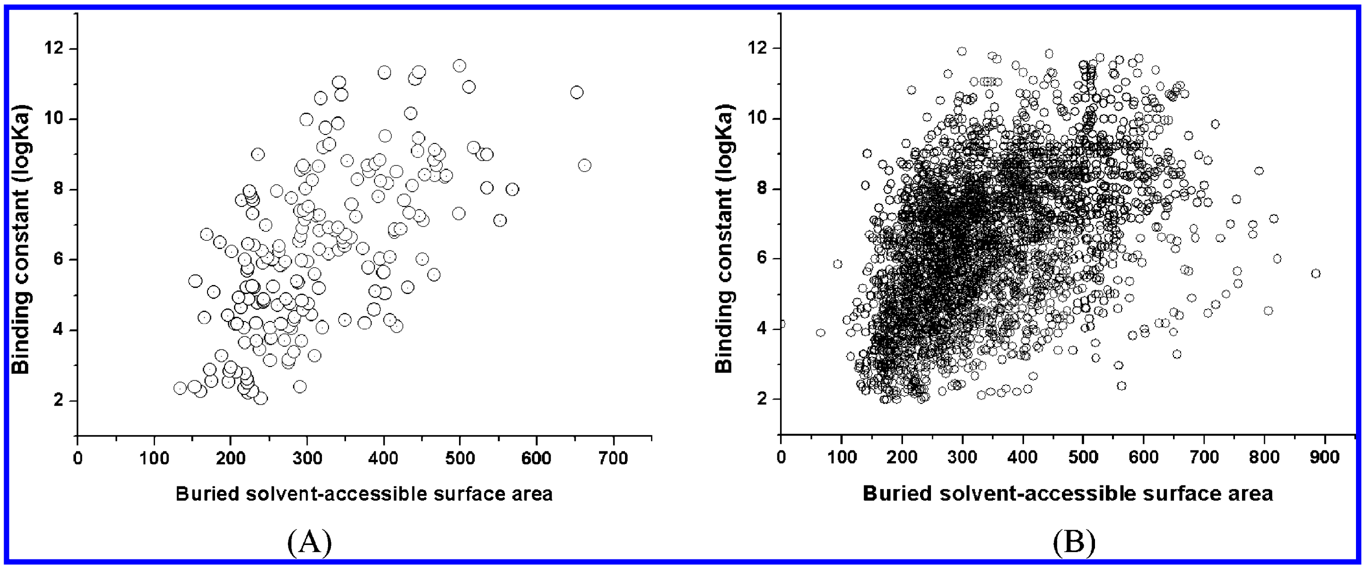


Figure 6. Scatter plot of binding constant versus buried solvent-accessible surface area (in Å^2) of the ligand molecule for the protein−ligand complexes included in (A) the PDBbind core set (N = 195, R^2=0.371) and (B) the PDBbind refined set (N = 2959, R^2= 0.224).

data set have efficiencies below a “soft limit” of −0.83 kcal/mol per atom.

In order to verify the results derived by other researchers, we also investigated ligand efficiency on our data sets. Our analyses were conducted on the PDBbind core set (N = 195) as well as the PDBbind refined set (N = 2959). We considered three descriptors of the ligand in computing ligand efficiency, including the number of non-hydrogen atoms (NHA), buried solvent-accessible surface area (ΔSAS), and buried molecular volume (ΔVOL). The scatter plots of binding constants versus these three descriptors on the two data sets are given in Figures 5−7. In each case, the upper limits of ligand efficiency at the 100%, 99%, and 95% levels are computed and summarized in Table 3. Here, the ″100% level” is the maximal ligand efficiency observed in the given data set. The ″99% level” means only 1% of samples have higher ligand efficiency than this standard, while the ″95% level” means 5% of samples have higher ligand efficiency than this standard.

Scatter plots of binding constants versus NHA of ligand molecules are given in Figure 5. One can see that there is a rough correlation between these two properties on the core set with a squared correlation coefficient (R^2) of 0.227 (Figure 5A). But this correlation is less obvious on the more comprehensive refined set (Figure 5B). As for the ligand efficiency values computed by $\Delta G_{\text{binding}}$/NHA, the maximal value on the refined set is −1.94 kcal/mol per non-hydrogen atom, which is higher than what were previously reported by Kuntz (−1.5 kcal/mol) and Carlson (−1.75 kcal/mol). This record-breaking complex is formed between putrescine, a small-size, positively charged molecule, and SpuD, a polyamine transport protein (Figure 8A). The dominant interactions between putrescine and SpuD are apparently the salt bridges as well as additional hydrogen bonds formed between two amine groups on putrescine and several Asp and Ser residues on SpuD. The dissociation constant (K_d) of this complex is reported to be 3.0 nM as measured by isothermal titration calorimetry.[54]

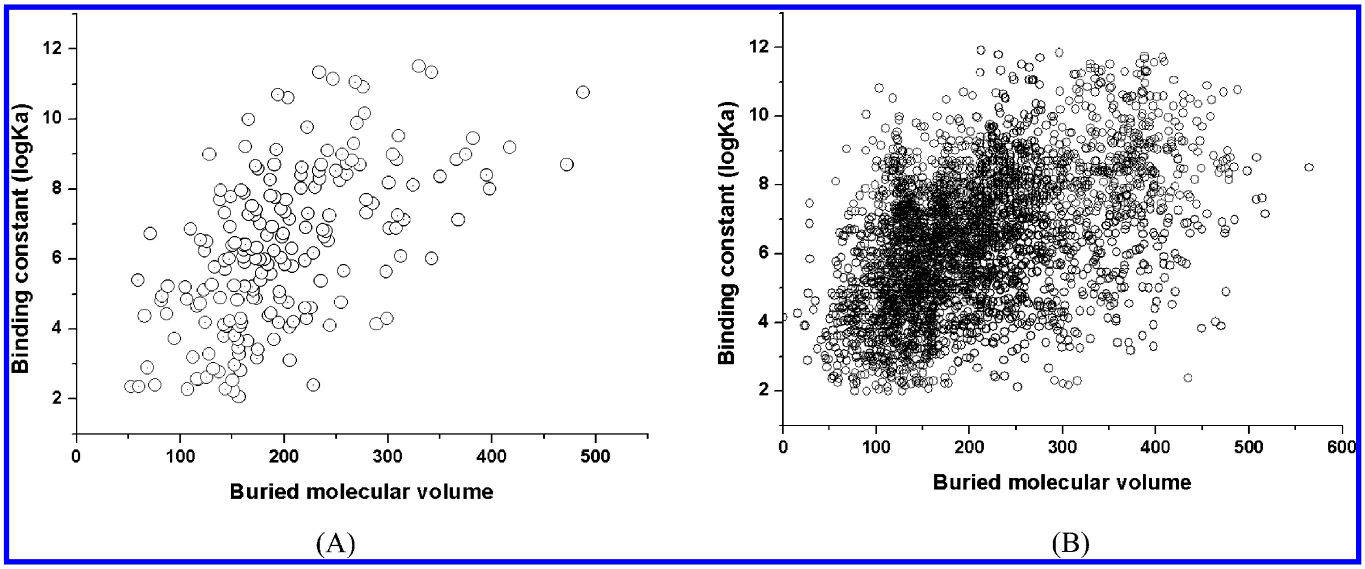


Figure 7. Scatter plot of binding constant versus buried molecular volume (in Å^3) of the ligand molecule for the protein−ligand complexes included in (A) the PDBbind core set (N = 195, R^2 = 0.325) and (B) the PDBbind refined set (N = 2959, R^2 = 0.223).

Table 3. Upper Limits of Ligand Efficiency Computed by Three Different Descriptors

	on the refined set (N = 2959)			on the core set (N = 195)		
ligand efficiency (kcal/mol)[a]	100% level	99% level	95% level	100% level	99% level	95% level
$\Delta G_{\text{binding}}$/NHA (kcal/(mol·at.))	−1.94 (3TTM)[b]	−1.09	−0.79	−1.17 (3F3E)[b]	−0.89	−0.76
$\Delta G_{\text{binding}}$/ΔSAS (cal/(mol·Å^2))	−82 (1AVN)[b]	−55	−46	−54 (2JDU)[b]	−49	−45
$\Delta G_{\text{binding}}$/ΔVOL (cal/(mol·Å^3))	−330 (3QIN)[b]	−127	−89	−128 (2JDU)[b]	−95	−79

[a] Standard binding free energies are computed from binding constants as $\Delta G_{\text{binding}} = -2.303RT \log K_a = -1.364 \log K_a$. [b] The PDB code of the corresponding protein−ligand complex structure.

On the PDBbind refined set, the ligand efficiency limit at the 99% level is −1.09 kcal/mol. The limit at the 95% level decreases further to −0.79 kcal/mol, which is merely 40% of the top value. It indicates that extremely high ligand efficiency can be found among only a small number of samples. In comparison, the ligand efficiency limits at the 100% and 99% levels on the PDBbind core set are considerably lower than the corresponding values derived on the refined set. It is because some complexes with particularly high ligand efficiency are not included in the core set. Nevertheless, the ligand efficiency limit at the 95% level, i.e., −0.76 kcal/mol per atom, is very close to its counterpart derived on the larger refined set. Interestingly, both values are close to the value reported by Carlson et al., i.e., −0.83 kcal/mol per non-hydrogen atom,[53] which was derived at the 95% level on their data set.

Scatter plots of binding constants versus ΔSAS or ΔVOL of ligand molecules are given in Figure 6 and Figure 7, respectively. Compared to NHA, which is an indicator of the size of the ligand molecule, these two descriptors are related to protein−ligand interactions more straightforwardly. Indeed, the correlation between binding constants and these two descriptors is more obvious on the core set (R^2 = 0.371 and 0.325). A rough correlation can be observed on the refined set as well (R^2 = 0.224 and 0.223). If ligand efficiency is computed as log K_a/ΔSAS, the top value observed on the refined set is −82 (cal/mol)/Å^2. It is a complex formed between carbonic anhydrase II and histamine (PDB entry 1AVN, Figure 8B). Its dissociation constant is reported to be 125 μM.[55] If ligand efficiency is computed by ΔVOL, the top value observed on the refined set is −330 (cal/mol)/Å^3. It is a complex formed between HIV-1 RNase H and a pyrimidinol carboxylic acid inhibitor (PDB entry 3QIN). Its dissociation constant is reported to be 136 nM.[56]

Here, we want to point out that neither ΔSAS nor ΔVOL is a preferred descriptor for computing ligand efficiency. If ligand efficiency is computed so, a ligand molecule binding to the target protein through a small contacting interface will be favored. Take the complex with the highest log K_a/ΔSAS for example (Figure 8B); histamine forms a single hydrogen bond with Glu92 on carbonic anhydrase II, but the rest parts of this molecule are largely exposed to the solvent. Such a picture is certainly not what one would expect by the concept of ligand efficiency. Given the fact that there is an obvious correlation between binding affinity and ΔSAS or ΔVOL, a ligand molecule binding to the target protein through a small contacting interface should be abandoned rather than be pursued.

Based on our results, we conclude that it is still too early to define what the upper limit of ligand efficiency is. If this limit is derived from a statistical survey, it is dependent on the data set under survey. New records probably will be set in the future once even larger data sets become available. But the limit at an appropriate level, such as 95%, is probably a converged value. The results obtained by us as well as other researchers suggest that this value is around −0.8 kcal/mol per non-hydrogen atom. Such a limit of ligand efficiency has more practical meanings because the top 5% ligand molecules are not very druglike anyway.

3.3. Advantages and Limitations of Our Data Set. As described in the Introduction, the primary test set used in our benchmark is desired to have two essential qualities. One quality is that it should consist of diverse protein−ligand complexes

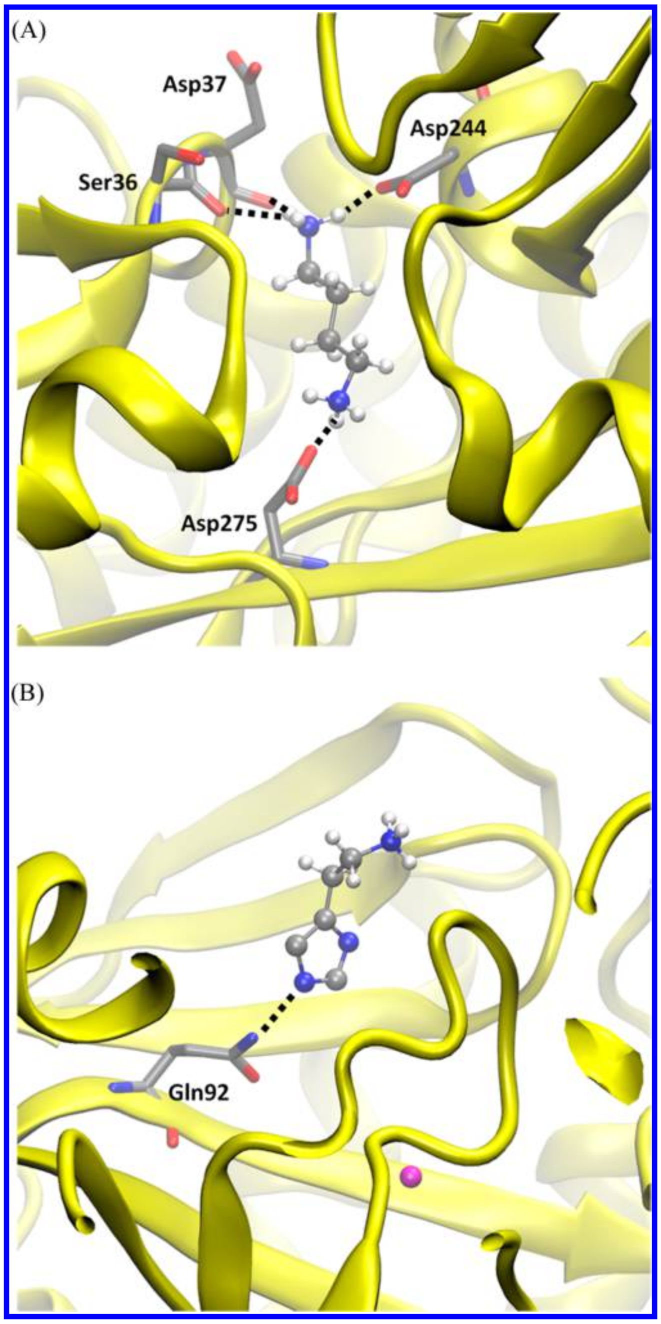


Figure 8. (A) Interactions between putrescine and polyamine transport protein SpuD (PDB entry 3TTM). Ligand efficiency of putrescine is −1.94 kcal/mol per non-hydrogen atom or −65 cal/mol per Å^2 buried surface area. (B) Interactions between histamine and carbonic anhydrase II (PDB entry 1AVN). Ligand efficiency of histamine is −0.66 kcal/mol per non-hydrogen atom or −82 cal/mol per Å^2 buried surface area.

rather than a handful of families of protein−ligand complexes. It is because the goal of our benchmark is to test the general performance of scoring functions, not the performance on certain types of complexes. Here, our data set consists of a wide range of important protein molecules in PDB. Structural diversity at the protein side is imposed by considering sequence similarity; whereas structural diversity at the ligand side is imposed by selection of samples at different levels of binding affinities. Moreover, tight control on sample redundancy is applied, where each cluster of complexes is represented by three selected members so that the contribution from each cluster is equal.

Another essential quality of our test set is that it should include only protein−ligand complexes with both high-quality crystal structures and reliable binding data. As explained in Methods, a set of rules with concerns on crystal structures, binding data, and the nature of complexes were applied to the selection of the refined set. These rules are more stringent than those employed by many other studies. As a matter of fact, only 36% of the protein−ligand complexes in the PDBbind database (2,959 in 8,302) were accepted into the refined set, which reflects the level of quality required in sample selection. Since the core set is technically a subset of the refined set, all of the complexes in the core set automatically meet all of these requirements. Besides, all of the complexes in the final core set need to pass the visual examination of electron density maps to ensure that there are no obvious flaws in the ligand binding pose or nearby residues. This serves as another critical control on structural quality because a significant fraction of the protein−ligand complex structures in PDB actually have problems.[28−30] In fact, over one-third of the candidates selected from the refined set failed to pass this examination.

The community structure−activity resource exercise[31−34] organized by Prof. Carlson's group at the University of Michigan is also a high-impact benchmark in this field. The first CSAR exercise (CSAR-2010) assessed the performance of a panel of scoring functions on a set of 343 protein−ligand complexes, i.e., the CSAR-NRC HiQ set.[31] This data set was mainly selected from the binding MOAD database through a multistep process. The readers may want to know how our data set compares to the CSAR-NRC HiQ set. Among the 343 protein−ligand complexes in the CSAR-NRC HiQ set, 52 are not included in the PDBbind database (version 2013). Some of those 52 complexes contain ligand molecules that are not "valid" by our definition, such as inorganic ions, while others have binding data that are not recorded in PDBbind. Among the other 291 complexes in the CSAR-NRC HiQ set, 42 complexes are not included in the PDBbind refined set due to miscellaneous problems in structures or binding data. Therefore, up to 86% complexes (249 in 291) in the CSAR NRC-HiQ set meet the standard of the PDBbind refined set.

Moreover, a protein−ligand complex needs to pass the electron density map examination to qualify for the PDBbind core set. We did not check the complex structures in the CSAR-NRC HiQ set one by one; thus, we do not know how many of them will pass this examination. But we did come across a few complex structures in the CSAR NRC-HiQ set in our examination and found that they were problematic. Two examples (PDB entry 2PGZ and 2AZR) are given in Figure 9 to illustrate our concerns in this examination. In both cases, one can see notable "positive" or "negative" density regions around the ligand molecule. It means that fitting of the ligand molecule into the electron density map is somewhat problematic in these regions. Consequently, the ligand binding pose may not be very accurate. Since accurate ligand binding pose is important for our subsequent evaluation of scoring functions, we consider such a complex structure as problematic even if its overall quality is good.

Besides the advantages previously discussed, it is also important to be aware of the limitations embedded in our data set. The protein−ligand complexes included in this data set may still have minor problems in crystal structures, binding data, or other aspects. As for structure quality, it is possible to apply even

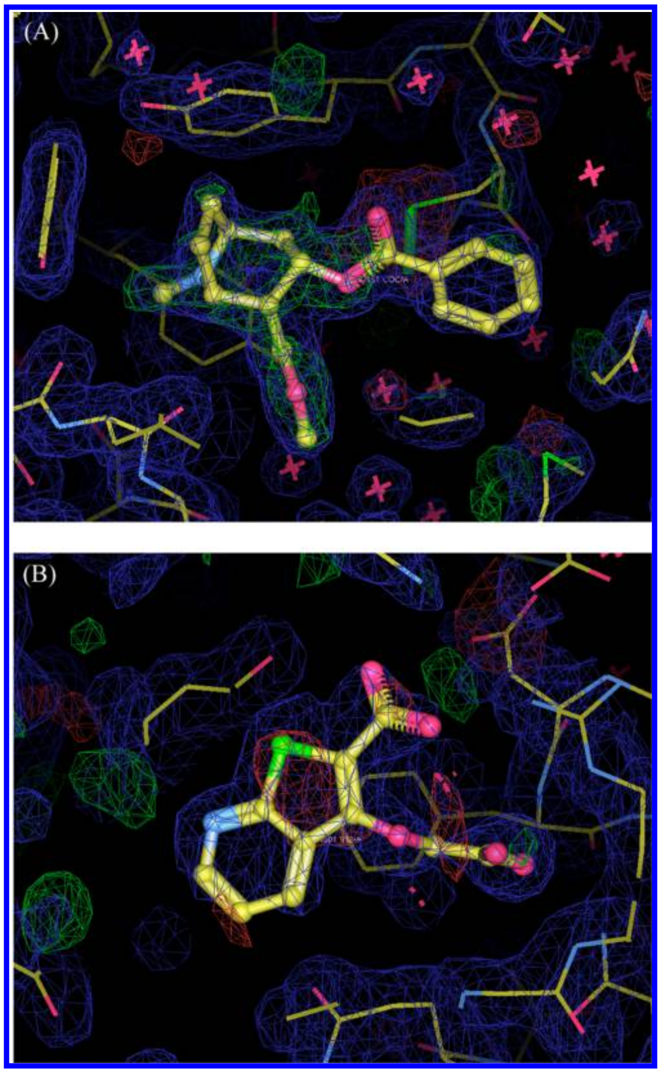


Figure 9. Examples of unqualified protein−ligand complex structures identified in our electron density map examination although the overall quality of these crystal structures is good: (A) PDB entry 2PGZ, resolution = 1.76 Å, R_{free} = 0.210; (B) PDB entry 2AZR, resolution = 2.00 Å, R_{free} = 0.250. Structure factors were downloaded from PDB. Electron density maps were displayed by the COOT software (version 0.7.1). Positive and negative regions in electron density fitting are indicated in red and green, respectively.

more stringent criteria in sample selection. For example, there are other metrics for measuring overall or local quality of a crystal structure, such as the diffraction-component precision index (DPI),[57−59] the real-space *R*-factor (RSR), and the real-space correlation coefficient (RSCC).[60] If those metrics were also applied in our study, we certainly would have obtained a smaller pool of qualified samples. Thus, a compromise has to be made to ensure that the final data set still has an acceptable size. Besides, we conducted electron density map examination in sample selection. We hope that this examination plays, if not more stringent, an equivalent role as other metrics previously mentioned.

The second problem lies in the experimental binding data. All of the binding data in the PDBbind database are collected from public literature. Those data are generated in various binding assays under different experimental settings. One should not expect that they are all in the same level of quality. Thus, there is a certain level of intrinsic error in our binding data. In a recent analysis of multiple protein−ligand binding data in the ChEMBL database,[61] Kramer et al. found 2540 complexes with at least two binding data. Among them, 67% measurements are within one log unit of one another; whereas the remaining are not. As a whole, the binding data included in the PDBbind core set is supposed to have a better quality than a large-scale database such as ChEMBL because some controls were made during our sample selection to remove less reliable binding data. Yet, our estimation is that the average error among the binding constants in this data set is at least 0.5 log units.

Third, all of the protein and ligand structures used in our computation were processed through a standard procedure for technical convenience. It is not good enough if some special factors in the given complex structure turn out to be important for protein−ligand interactions. For example, some residues at the protein side and some chemical groups at the ligand side need to be protonated properly to set the electrostatic interaction or hydrogen bonding network correct between the protein and the ligand. In our study, a simple protonation scheme was applied to each complex. We did not attempt to adjust the protonation states of each complex according to the experimental conditions used in crystal growth or binding assay. Another example is the water molecules mediating protein−ligand interactions. It is well-known that some of them are actually critical for protein−ligand binding.[62−64] Although water molecules are actually kept in our processed structural files, they were totally ignored in our evaluation of scoring functions. It is more practical to fix the problems mentioned above if one deals with only a handful classes of protein−ligand complexes. For a diverse data set such as ours, apparently a much higher level of effort is needed to accomplish this mission. Anyway, this aspect of our data set can be improved in the future by utilizing available information in the literature.

The last issue regards the size of our data set. Recently, Carlson et al. provided interesting discussion on the ideal size of a test set for comparing scoring functions.[65] They stated that if one wants to confirm the statistical difference of two scoring functions differing by $\Delta R = 0.1$ (for example $R = 0.75$ versus $R = 0.65$) at the 90% confidence level, a test set of at least 362 samples is needed. If the performance difference is narrowed to $\Delta R = 0.05$, then a test set of at least 1439 samples is needed. These estimations were made through hypothetical analyses by assuming a standard error in experimental binding data of 0.5 log units. If their statements are meaningful, the size of our current data set, i.e., 195 protein−ligand complexes, is about half of the smallest desired size for evaluating real scoring functions. We certainly agree that a larger test set is better for deriving statistically meaningful results. But other important qualities should not be sacrificed just to obtain a larger data set. This size problem will be solved once even more high-quality structures and binding data are available. Besides, it should be pointed out that Carlson's estimations were made by assuming that binding data in the test set is in a normal distribution. In reality, this requirement is not always met and it may not be necessary at all, such as in the case of our data set (Figure 3A).

The current PDBbind core set (version 2013) happens to have the same size as the one employed in our previous CASF-2007 benchmark,[39] i.e., the core set version 2007. However, only 25 complexes (~13%) in the previous one are still included in the current one. The current one was selected out of ~8300 protein−ligand complexes, whereas the previous one was selected out of ~3100 protein−ligand complexes. It reflects the higher standards adopted by us during compilation of the current version. We expect that future versions of the PDBbind core set will grow with the growth of PDBbind itself. We also plan to consider other possibilities, for example, allowing more members in each complex cluster, which will increase the size of this data set considerably. Our expectation is that the PDBbind core set will double in size in the next 3 years or so while we will strive to adopt even higher standards for sample selection.

4. CONCLUSION

Our CASF benchmark aims at providing an objective, third-party evaluation of scoring functions. The latest study, CASF-2013, is described in this and the companion work in this issue. As the basis of our benchmark, a set of protein−ligand complexes was selected out of 8302 protein−ligand complexes recorded in the PDBbind database (version 2013) through a fairly complicated process. Our emphasis in sample selection was on the quality of the complex structure, the binding data, and the nature of the complex. The standards employed in our sample selection are more stringent than those in similar works. Finally, qualified protein−ligand complexes were clustered by 90% similarity of protein sequences. Only three representative complexes were chosen from each cluster to control sample redundancy. It should be mentioned that besides the final data set, setting up such a workflow is also a major contribution of our work. It reflects our understanding of how a high-quality data set for validating docking/scoring methods should be compiled.

The final data set, namely, the PDBbind core set (version 2013), consists of a total of 195 protein−ligand complexes in 65 clusters. Binding constants of these complexes span nearly 10 orders of magnitude. In this data set, 82% of ligand molecules are druglike and 78% of protein molecules are validated or potential drug targets. Survey conducted on this data set as well as the PDBbind refined set reveals that the correlation between binding constants and ligand sizes is rather weak. But the correlation between binding constants and buried solvent-accessible surface areas or molecular volumes of ligands is more obvious. Moreover, our results suggest that the upper limit of ligand efficiency at the 95% level is probably a converged value around −0.8 kcal/mol per non-hydrogen atom.

This data set served as the primary test set in our CASF-2013 benchmark for evaluating a panel of 20 scoring functions, which will be described in the companion work in this issue. This data set has been released to the public on our PDBbind-CN Web site (http://www.pdbbind-cn.org). We hope that this data set becomes a community resource for evaluation or development of docking/scoring methods.

■ ASSOCIATED CONTENT

Supporting Information

Figure showing distributions of five key properties of 1489 small-molecule drugs recorded in the DrugBank database. This material is available free of charge via the Internet at http://pubs.acs.org. The PDBbind database, including the core set described here, is freely available at http://www.pdbbind-cn.org/ .

■ AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interests.

■ ACKNOWLEDGMENTS

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

■ REFERENCES

(1) Kuntz, I. D. Structure-based strategies for drug design and discovery. *Science* **1992**, *257*, 1078−1082.

(2) Babine, R. E.; Bender, S. L. Molecular recognition of protein-ligand complexes: Applications to drug design. *Chem. Rev.* **1997**, *97*, 1359−1472.

(3) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303*, 1813−1818.

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

(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, USA, 2001; Vol. *17*, pp 1−60.

(6) Böhm, H. J.; Stahl, M. The use of scoring functions in drug discovery applications. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: Hoboken, NJ, USA, 2002; Vol. *18*, pp 41−88.

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

(8) Schulz-Gasch, T.; Stahl, M. Scoring functions for protein-ligand interactions: a critical perspective. *Drug Discovery Today* **2004**, *1*, 231−239.

(9) 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.

(10) 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.

(11) Zhong, S.; Zhong, S.; Zhang, Y.; Xiu, Z. Rescoring ligand docking poses. *Curr. Opin. Drug Discovery Dev.* **2010**, *13* (3), 326−34.

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

(13) 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.

(14) Bursulaya, B.; Totrov, M.; Abagyan, R.; Brooks, C. Comparative Study of Several Algorithms for Flexible Ligand Docking. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 755−763.

(15) Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On evaluating molecular-docking methods for pose prediction and enrichment factors. *J. Chem. Inf. Model.* **2006**, *46*, 401−415.

(16) 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.

(17) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455−1474. Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.

(18) 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.

(19) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy. *Proteins* **2004**, *57*, 225−242.

(20) Perola, E.; Walters, W. P.; Charifson, P. S. A Detailed Comparison of Current Docking and Scoring Methods on Systems of Pharmaceutical Relevance. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 235−249.

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

(22) Kontoyianni, M.; Sokol, G. S.; MCclellan, L. M. Evaluation of Library Ranking Efficacy in Virtual Screening. *J. Comput. Chem.* **2005**, *26*, 11−22.

(23) 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.

(24) 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.

(25) Kim, R.; Skolnick, J. Assessment of programs for ligand binding affinity prediction. *J. Comput. Chem.* **2008**, *29*, 1316−1331.

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

(27) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(28) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein−ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726−741.

(29) Warren, G. L.; Do, T. D.; Kelley, B. P.; Nicholls, A.; Warren, S. D. Essential considerations for using protein−ligand structures in drug discovery. *Drug Discovery Today* **2012**, *17*, 1270−1281.

(30) Søndergaard, C. R.; Garrett, A. E.; Carstensen, T.; Pollastri, G.; Nielsen, J. E. Structural Artifacts in Protein−Ligand X-ray Structures: Implications for the Development of Docking Scoring Functions. *J. Med. Chem.* **2009**, *52*, 5673−5684.

(31) Dunbar, J. B., Jr.; Smith, R. D.; Yang, C. Y.; Ung, P. M.; Lexa, K. W.; Khazanov, N. A.; Stuckey, J. A.; Wang, S.; Carlson, H. A. CSAR benchmark exercise of 2010: Selection of the protein−ligand complexes. *J. Chem. Inf. Model.* **2011**, *51*, 2036−2046.

(32) 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.

(33) 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.

(34) 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.

(35) Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother of All Databases). *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 333−340.

(36) Smith, R. D.; Hu, L.; Falkner, J. A.; Benson, M. L.; Nerothin, J. P.; Carlson, H. A. Exploring protein−ligand recognition with Binding MOAD. *J. Mol. Graphics Modell.* **2006**, *24*, 414−425.

(37) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47* (12), 2977−2980.

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

(39) 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.

(40) Li, Y.; Han, L.; Liu, Z.; Li, J.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results. *J. Chem. Inf. Model.* **2014**, DOI: 10.1021/ci500081m, (companion work in this issue).

(41) 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.

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

(43) Lee, B.; Richards, F. M. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379−400.

(44) Bondi, A. van der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441−451.

(45) Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L. Computation of octanol−water partition coefficients by guiding an additive model with knowledge. *J. Chem. Inf. Model.* **2007**, *47*, 2140−2148.

(46) 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.

(47) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: A comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035−D1041.

(48) Abad-Zapatero, C.; Metz, J. T. Ligand efficiency indices as guideposts for drug discovery. *Drug Discovery Today* **2005**, *10*, 464−469.

(49) Reynolds, C. H.; Tounge, B. A.; Bembenek, S. D. Ligand binding efficiency: Trends, physical basis, and implications. *J. Med. Chem.* **2008**, *51*, 2432−2438.

(50) Bembenek, S. D.; Tounge, B. A.; Reynolds, C. H. Ligand efficiency and fragment-based drug discovery. *Drug Discovery Today* **2009**, *14*, 278−283.

(51) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 9997−10002.

(52) Carlson, H. A.; Smith, R. D.; Khazanov, N. A.; Kirchhoff, P. D.; Dunbar, J. B., Jr.; Benson, M. L. Differences between High- and Low-Affinity Complexes of Enzymes and Nonenzymes. *J. Med. Chem.* **2008**, *51*, 6432−6441.

(53) Smith, R. D.; Engdahl, A. L.; Dunbar, J. B., Jr.; Carlson, H. A. Biophysical Limits of Protein-Ligand Binding. *J. Chem. Inf. Model.* **2012**, *52*, 2098−2106.

(54) Wu, D.; Lim, S. C.; Dong, Y.; Wu, J.; Tao, F.; Zhou, L.; Zhang, L.-H.; Song, H. Structural Basis of Substrate Binding Specificity Revealed by the Crystal Structures of Polyamine Receptors SpuD and SpuE from Pseudomonas aeruginosa. *J. Mol. Biol.* **2012**, *416*, 697−712.

(55) Briganti, F.; Mangani, S.; Orioli, P.; Scozzafava, A.; Vernaglione, G.; Supuran, C. T. Carbonic Anhydrase Activators: X-ray Crystallographic and Spectroscopic, Investigations for the Interaction of Isozymes I and II with Histamine. *Biochemistry* **1997**, *36*, 10384−10392.

(56) Lansdon, E. B.; Liu, Q.; Leavitt, S. A.; Balakrishnan, M.; Perry, J. K.; Lancaster-Moyer, C.; Kutty, N.; Liu, X.; Squires, N. H.; Watkins, W. J.; Kirschberg, T. A. Structural and Binding Analysis of Pyrimidinol Carboxylic Acid and N-Hydroxy Quinazolinedione HIV-1 RNase H Inhibitors. *Antimicrob. Agents Chemother.* **2011**, *55*, 2905−2915.

(57) Cruickshank, D. W. J. Remarks about protein structure precision. *Acta Crystallogr., Sect. D: Biol Crystallogr.* **1999**, *D55*, 583−601.

(58) Cruickshank, D. W. J.Coordinate Uncertainty in International Tables for Crystallography . In *Crystallography of Biological Macromolecules*, Vol. F; Rossmann, M. G., Arnold, E., Eds.; Kluwer Academic: Dordrecht, The Netherlands, 2001; Chapter 18.5, pp 403−414,

(59) Blow, D. M. Rearrangement of Cruickshank's formulae for the diffractioncomponent precision index. *Acta Crystallogr., Sect. D: Biol Crystallogr.* **2002**, *D58*, 792−797.

(60) Jones, T. A. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1991**, *A47*, 110−119.

(61) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The experimental uncertainty of heterogeneous public Ki data. *J. Med. Chem.* **2012**, *55*, 5165−5173.

(62) Barillari, C.; Taylor, J.; Viner, R.; Essex, J. W. Classification of Water Molecules in Protein Binding Sites. *J. Am. Chem. Soc.* **2007**, *129*, 2577−2587.

(63) Amadasi, A.; Surface, J. A.; Spyrakis, F.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E. Robust Classification of "Relevant" Water Molecules in Putative Protein Binding Sites. *J. Med. Chem.* **2008**, *51*, 1063−1067.

(64) Lu, Y.; Wang, R.; Yang, C.-Y.; Wang, S. Analysis of Ligand-Bound Water Molecules in High-Resolution Crystal Structures of Protein−Ligand Complexes. *J. Chem. Inf. Model.* **2007**, *47*, 668−675.

(65) Carlson, H. A. Check Your Confidence: Size Really Does Matter. *J. Chem. Inf. Model* **2013**, *53*, 1837−1841.