JOURNAL OF CHEMICAL INFORMATION AND MODELING

Perspective

pubs.acs.org/jcim

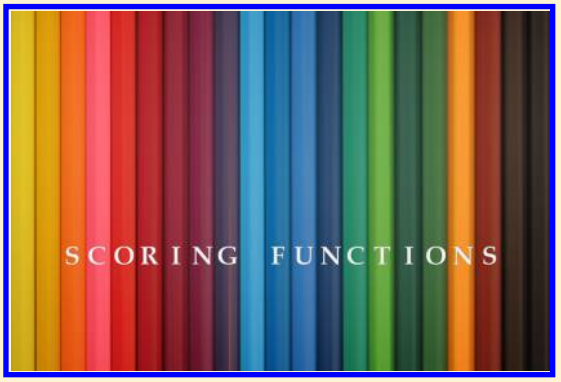
Classification of Current Scoring Functions

Jie Liu[†] 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, Shanghai, 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

ABSTRACT: Scoring functions are a class of computational methods widely applied in structure-based drug design for evaluating protein−ligand interactions. Dozens of scoring functions have been published since the early 1990s. In literature, scoring functions are typically classified as force-field-based, empirical, and knowledge-based. This classification scheme has been quoted for more than a decade and is still repeatedly quoted by some recent publications. Unfortunately, it does not reflect the recent progress in this field. Besides, the naming convention used for describing different types of scoring functions has been somewhat jumbled in literature, which could be confusing for newcomers to this field. Here, we express our viewpoint on an up-to-date classification scheme and appropriate naming convention for current scoring functions. We propose that they can be classified into physics-based methods, empirical scoring functions, knowledge-based potentials, and descriptor-based scoring functions. We also outline the major difference and connections between different categories of scoring functions.



INTRODUCTION

Scoring functions are widely applied in structure-based drug design for evaluating protein−ligand interactions. Scoring functions do not attempt to account for the physics in a protein−ligand binding process with a high-level theory. Instead, they make various approximations in order to provide a compromise between speed and accuracy. Scoring functions are thus particularly suitable for high-throughput tasks, such as molecular docking, virtual screening, library design, and so on.[1−5]

The first batch of scoring functions appeared in the early 1990s. After 20 years, scoring functions have grown into a large diverse family. Our estimation based on literature survey is that over a hundred scoring functions have already been published in the literature. A classification scheme of scoring functions has been given in some classical works,[1,2] which is typically narrated as "According to how they are derived, current scoring functions can be grouped into three categories: force-field-based, empirical, and knowledge-based." In fact, this scheme is still repeatedly quoted by some recent publications.[6−8] We want to point out that this scheme does not reflect the recent progress in this field. An update is thus needed. Besides, the naming convention for describing different types of scoring functions has been somewhat jumbled in the literature, which could be confusing or even misleading for newcomers to this field. To facilitate the discussion on scoring functions, a generally acceptable naming convention is desired. To our knowledge, this issue has not been addressed seriously in the public literature in recent years.

Therefore, we would like to express our viewpoint on this issue by writing this article to the *Journal of Chemical Information and Modeling*, a journal that has published many high-quality studies on scoring functions over all these years. The main purpose of our article is not to analyze the current status of scoring functions or to provide perspectives for future development. Those topics have been addressed by many excellent reviews.[1−5] Instead, the main purpose of our article is to propose an updated classification scheme and an appropriate naming convention of current scoring functions. We give our discussion on each type of scoring function in a separate section below. The strengths and weaknesses of each approach are also outlined to orient the readers.

CATEGORY I: "FORCE-FIELD-BASED" OR "PHYSICS-BASED"?

We start our discussion with the so-called "force-field-based scoring functions". At the beginning, there was no scoring function especially developed for evaluating protein−ligand interactions. Nevertheless, force field had been gradually recognized as a powerful tool for modeling biological macromolecules since the pioneering works by Martin Karplus and co-workers in the 1970s.[9,10] Researchers thus relied on available force fields to compute the direct interactions between protein and ligand. Often the noncovalent energy terms in a force field, including the van der Waals and the electrostatic energy terms, were used for this purpose due to the nature of

Received: December 10, 2014
Published: February 3, 2015

DOI: 10.1021/ci500731a
J. Chem. Inf. Model. 2015, 55, 475−482

protein−ligand interactions. Hydrogen bonding was taken into account with an additional term or being included implicitly in the electrostatic energy term. For example, early versions of both DOCK[11,12] and AutoDock[13] employed energy functions based on the AMBER force field[14−16] as their internal scoring engine. Force field energy functions are designed to compute potential energy in the gas phase, which is only one component of the free energy change in a protein−ligand binding process. Later, scoring functions based on force field were augmented by solvation energy terms,[17,18] which were computed with either Poisson−Boltzmann (PB) or Generalized Born (GB) continuum solvation models. The general functional form below thus became more widely adopted:

$$\Delta G_{\text{binding}} = \Delta E_{\text{vdw}} + \Delta E_{\text{electrostatic}} + [\Delta E_{\text{H-bond}}] + \Delta G_{\text{desolvation}} \quad (1)$$

Besides the ones already mentioned above, a few more examples of such scoring functions include COMBINE,[19] GoldScore,[20] and MedusaScore.[21]

In fact, another group of computational methods share a similar theoretical framework, including the Linear Interaction Energy (LIE) method[22,23] and the Linear Response Approximation (LRA) method[24,25] which appeared in the early 1990s as well as the popular MM-PBSA/GBSA method which appeared in the late 1990s.[26] These methods also employ standard force fields to compute potential energies, and they usually rely on molecular dynamics simulations in explicit solvent for configurational sampling. Unlike full-scale free energy perturbation (FEP) or thermodynamics integration (TI) computations, these methods only consider protein and ligand in their unbound and bound states to compute the free energy change in the binding process. Therefore, they are also referred to as "end-point approximations" in the literature. Such methods are normally not considered as scoring functions because they compute energies as ensemble averages in an attempt to incorporate more physics at the expense of significantly increased computational cost. But in reality, these methods are often applied to compute one or a few selected snapshots of a protein−ligand complex structure for scoring/ranking purposes just like standard scoring functions. For example, the MM-PBSA/GBSA method can be applied in this manner as a reranking tool in virtual screening to improve hit rates.[27] Besides, these methods may also incorporate empirical parameters to improve their performance on certain molecular systems, such as the α, β, and γ parameters in LIE or LRA models. This practice further blurs the boundary between these methods and standard scoring functions.

Warmed up by the above descriptions, one can see that "force-field-based scoring function" is not an accurate term to summarize these computational methods. First, the PB or GB solvation model, usually as an indispensable component in the energy function, is not based on force field. Second, end-point approximation methods may be applied in practice as advanced scoring functions, but they are not covered by this term in concept. Third, given the power of today's computers, quantum mechanics (QM) computation may replace the role of force field in evaluating the direct interactions between protein and ligand. Apparently there are still many technical difficulties along this approach, but a number of studies have started to show promising results.[28−31] This represents a notable new trend in this field for the coming years. Therefore, our opinion is that all these methods should be generally referred to as "physics-based methods". This term actually has long been used in the literature. We simply want to emphasize that it is the appropriate term to describe the broad scope of this type of methods.

An obvious advantage of physics-based methods is that they can ride on the progress of modern force fields, quantum mechanics methods, solvation models, and others. Theoretical chemists have been working ceaselessly on these subjects, and the whole field of computational chemistry has advanced greatly as compared to 20 years ago. If scoring functions can reliably compute protein−ligand binding free energies one day, physics-based methods, if properly designed, should have the best chance. However, due to the extremely complicated nature of binding free energy prediction, so far rather limited successes have been achieved even by high-level theories.[32−34] It still needs to be testified if new methods, such as polarizable force fields,[35] can lead to a real breakthrough. In the technical aspect, current physics-based methods often produce unrealistic energies regardless of whether they are based on force fields or QM models. Experimentally measured binding free energy changes are in small quantities, typically between 3−18 kcal/mol. Considering the intrinsic error of each individual energy term in a physics-based method, one probably should not expect that those terms automatically cancel out to give the correct result. That is why current physics-based methods often need empirical scaling parameters to fit their results to experimental binding data.

CATEGORY II: "EMPIRICAL" OR "REGRESSION-BASED"?

The second type of method has been called "empirical scoring function". Most literature points to the method published by Böhm in 1994[36] as the first general-purpose empirical scoring function. This scoring function is still available today in the Discovery Studio software. A few more examples of popular empirical scoring functions, ranked by their first publication date, include PLP,[37] ChemScore,[38,39] X-Score,[40] and GlideScore.[41,42]

An empirical scoring function computes the fitness of protein−ligand binding by summing up the contributions of a number of individual terms, each representing an important energetic factor in protein−ligand binding. For example, ChemScore implemented in the GOLD software uses the following formula:[43]

$$\text{ChemScore} = S_{\text{H-bond}} + S_{\text{metal}} + S_{\text{lipophilic}} + P_{\text{rotor}} + P_{\text{strain}} + P_{\text{clash}} + [P_{\text{covalent}} + P_{\text{constraint}}] \quad (2)$$

It consists of rewarding scores ("S") for hydrogen bonding, coordinate bonds with metal ions, and lipophilic contacts and penalties ("P") for frozen rotatable bonds and internal strain energy of the ligand as well as steric clashes between protein and ligand. Additional penalties may be invoked if covalent or restrained docking is required. Since multiple terms with different implications are combined to give the final binding score, an empirical scoring function normally relies on multivariate linear regression (MLR) or partial least-squares (PLS) analysis to derive the weight factor before each term. A training set of protein−ligand complexes with known three-dimensional structures and binding affinity data is required to perform the regression analysis. Therefore, empirical scoring functions are calibrated at the first place to reproduce protein−ligand binding affinities. Indeed, some comparative studies of

DOI: 10.1021/ci500731a
J. Chem. Inf. Model. 2015, 55, 475−482

scoring functions[44−46] indicated that empirical scoring functions on average are more capable than other types of scoring functions in this aspect.

"Empirical scoring function" is an appropriate term for describing this type of methods for two reasons. First, although some terms in an empirical scoring function share the same physical meaning as their counterparts in a physics-based method, the functional form of an empirical scoring function is often more intuitive than the energy function in a physics-based method. Moreover, the energy terms may be implemented in rather intuitive ways. For example, almost every empirical scoring function has a term accounting for hydrophobic effect, but this term is implemented in very different algorithms in different scoring functions. Second, an empirical scoring function relies on regression analysis to derive the weight factors for its individual terms. That is why sometimes these methods are also referred to as "regression-based" methods in the literature. But this term reflects the technical aspect of empirical scoring functions rather than their theoretical basis. Therefore, our opinion is that "empirical scoring function" is the more appropriate term.

The boundary between an empirical scoring function and a physics-based method is often not as distinct as one may think. In fact, both of them decompose protein−ligand binding free energy into individual energy terms. In addition, a physics-based method may introduce empirical parameters to reconcile the contributions of its energy terms just like an empirical scoring function. Yet, it is still helpful for discussion to separate empirical scoring functions from physics-based methods. The major difference between them is that a physics-based method borrows the complete theoretical framework, including the energy function and the associated parameters, from other well-established models; whereas an empirical scoring function usually adopts a flexible, intuitive functional form that is composed from scratch.

The consequence of adopting intuitive functional forms is actually double-edged for empirical scoring functions. On one hand, it is a technical advantage because it is convenient to implement any reasonable idea. Such an example is GlideScore-XP,[42] which is arguably the most sophisticated empirical scoring function at present. GlideScore-XP is designed with an emphasis on recognizing the diversity in protein binding sites by rewarding or penalizing certain interaction patterns. Of particular interest is the classification of hydrogen bonds into neutral−neutral, neutral−charged, and charged−charged types and use of separate terms accounting for "hydrophobic enclosure" in addition to consideration of hydrophobic contacts between protein and ligand. The convenience of adding or removing individual terms also makes it possible to develop customized scoring functions for certain molecular systems to achieve better performance.[47−49] On the other hand, adopting intuitive functional forms adds to the empirical nature of these methods. Empirical scoring functions include only common protein−ligand interaction patterns. Less common interaction patterns, despite being strong and specific such as cation-π interaction, are usually ignored because they are not significant in the regression analysis. Or, if a certain factor is not interpretable by human in a straightforward manner, such as entropic factors, it is not likely to be included either. Thus, it is rather difficult, if not impossible, to establish a comprehensive and consistent description of all possible factors in protein−ligand binding within the framework of an empirical scoring function.

As mentioned a couple of times above, data sets of protein−ligand complexes with known three-dimensional structures and binding affinity data are needed to derive the critical weight factors in an empirical scoring function. Such data sets themselves used to be a bottleneck for scoring function development in early years. For example, Böhm's pioneering study (SCORE1 or LudiScore) employed a set of 34 protein−ligand complexes as the training set.[36] The empirical scoring functions published in late 1990s were normally calibrated on a training set of fewer than a hundred protein−ligand complexes. Robust statistical models are not likely to be obtained on such limited data sets. In fact, we demonstrated with X-Score, an empirical scoring function at roughly the same level of complexity as LudiScore, that converged regression models were obtained only when the training set consisted of approximately two hundred samples.[40] Luckily, structural information and binding data of protein−ligand complexes have accumulated rapidly over these years. By the end of 2014, over 105 000 experimentally resolved structures have already been deposited in the Protein Data Bank (PDB),[50,51] nearly half of which are valid protein−ligand complexes.[52] Comprehensive collections of binding data are also available from several public-domain databases, such as BindingDB,[53,54] ChEMBL,[55,56] and PubChem Bioassay.[57,58] Some focused databases, such as PDBbind[52,59] and Binding MOAD,[60,61] collect experimental binding data particularly for the protein−ligand complexes in PDB, which provide direct aids to scoring function development.

Nowadays, researchers are able to use several thousands of protein−ligand complexes at a better quality to calibrate or test their scoring functions. Empirical scoring functions used to be questioned if they were applicable to chemotypes or genotypes outside their limited training sets. This doubt may be largely removed now because current empirical scoring functions are based on much more comprehensive data sets. A major remaining concern is that the experimental binding data collected from public literature are not measured under a consistent condition. Statistical surveys revealed that binding data for the same protein−ligand complex could scatter in a wide range if they are obtained from different sources.[62] The intrinsic error in experimental binding data of course sets the ceiling for any computational prediction. Therefore, one should always choose to use high-quality, consistent binding data whenever possible.

CATEGORY III: "POTENTIAL OF MEAN FORCE" OR "KNOWLEDGE-BASED POTENTIAL"?

The third type of method is often referred to as "knowledge-based scoring function" in literature. To the best of our knowledge, the first general-purpose scoring function of this type was the one implemented in SMoG, a de novo design program published in 1996.[63,64] A rapid spread of such methods has been witnessed in the following ten years or so. A few more examples, ranked by their first publication date, include Muegge's PMF,[65−67] DrugScore,[68−70] IT-Score,[71−73] and KECSA.[74]

Although differing in technical aspects, these scoring functions follow the same principle. They sum pairwise statistical potentials between protein and ligand:

$$A = \sum_{i}^{\text{lig}} \sum_{j}^{\text{prot}} \omega_{ij}(r) \tag{3}$$

DOI: 10.1021/ci500731a
J. Chem. Inf. Model. 2015, 55, 475−482

The distance-dependent potential between atom pair i–j, i.e., $\omega_{ij}(r)$, is derived from an inverse Boltzmann analysis as

$$\omega_{ij}(r) = -k_B T \ln[g_{ij}(r)] = -k_B T \ln\left[\frac{\rho_{ij}(r)}{\rho_{ij}^*}\right] \quad (4)$$

Here, $\rho_{ij}(r)$ is the numeric density of atom pair i–j at distance r and ρ_{ij}^* is the numeric density of the same atom pair in a reference state where interatomic interactions are assumed to be zero. With this approach, the occurrence frequency of a pairwise contact is assumed to be a measure of its energetic contribution to protein–ligand binding. If a specific pairwise contact occurs more frequently than that in the reference state, i.e. a random distribution, it indicates an energetically favorable interaction between the given atom pair; if it occurs less frequently, then it indicates an unfavorable interaction. In order to derive the desired pairwise potentials, the standard approach is to use a large set of protein–ligand complex structures from PDB as the training set, i.e. the "knowledge base". Atoms on the protein side and the ligand side are classified into a number of degenerated atoms types according to their molecular environment. Then, distance-dependent potentials for each possible atom pair are derived from the occurrence frequency of this atom pair observed in the training set with formula 4.

As a basic idea rooted in statistical mechanics analysis of liquids,[75,76] inverse Boltzmann analysis leads to efficient conversion of a histogram of interatomic distances into potentials of mean force. Formula 4 certainly resembles this approach, and thus some researchers name this type of scoring functions as "potential of mean force". Nevertheless, neither the protein nor the ligand is a randomized assembly of atoms as in liquid. For instead, atoms in a molecule are constrained by covalent bonds in certain orders. The reference state considered in formula 4 thus does not comply with the definition of a true reference state. Due to the same reason, the occurrence of different atom pairs are not totally independent. Some analyses pointed out that the occurrence frequency of an certain atom pair in real protein–ligand complex structures should not be assumed to be in a Boltzmann distribution.[77,78] Although the statistical potentials derived by formula 4 are often considered as approximation of potentials of mean force, this interpretation is actually not firmly grounded.

Thus, our opinion is that the term potential of mean forces used for describing this type of scoring function causes unnecessary confusion. To make it even more complicated, a class of studies in which protein–ligand binding free energy changes are computed as a function of certain reaction coordinates through molecular dynamics sampling are also labeled as "potential of mean force" in the literature.[79] For these reasons, we recommend to use the term "knowledge-based potential" to describe this type of scoring function. In fact, this term has been used for describing the statistical potentials derived from protein structures. Before they were applied to modeling protein–ligand interactions, knowledge-based potentials had been applied to evaluation of protein structures and study of protein folding.[80−82] Note that the term "knowledge-based scoring function" should be avoided too because it is ambiguous in meaning: all scoring functions are in fact more or less knowledge-based.

A major attraction of knowledge-based potentials is their conceptual and computational simplicity. Compared to the physics-based methods that often need computationally expensive treatment of solvent, these methods are much more efficient due to their pairwise characteristics. Unlike empirical scoring functions, knowledge-based potentials attempt to capture all of the energetic factors in protein–ligand interaction implicitly with the pairwise potentials. They are thus not troubled by choosing an optimal functional form. Moreover, they are derived through statistical analysis of pure structural information without the need of experimental binding data. Considering their theoretical basis, knowledge-based potentials should be applied to reproduce protein–ligand binding poses rather than binding energies. But researchers correlate such potentials to binding affinity data anyway, and the observed correlation is often comparable or even better than other types of scoring functions.[71−74] Some methods actually combine knowledge-based potentials with solvation and entropy terms.[68,73,83,84] If so, they have crossed the border and in fact become hybridized methods between pure knowledge-based potential and empirical scoring function.

CATEGORY IV: "DESCRIPTOR-BASED" OR "MACHINE-LEARNING BASED"?

The fourth type of scoring function represents a new trend in this field. These methods introduce modern quantitative structure–activity relationship (QSAR) analysis into protein–ligand interaction evaluation. QSAR analysis has been applied widely to the modeling of various physicochemical, biological, and pharmaceutical properties of small-molecule compounds since the dawn of computer-aided drug design.[85] If the properties of the ligand and the protein as well as their interaction patterns can be coded with certain descriptors, then those sophisticated machine-learning techniques employed in modern QSAR analysis can be applied to derive statistical models that compute protein–ligand binding scores.

It seems that this type of study started to emerge around 2004.[86,87] A few recently published examples include NNScore,[88,89] RF-Score,[90,91] SFCscore^{RF},[92] and ID-Score.[93] The starting point of such a method is usually a large pool of descriptors. One popular choice is to use atom pairs or structural interaction fingerprints[94] between a protein and a ligand. Descriptors accounting for specific interactions (electrostatic interactions, hydrogen bonds, or aromatic stacking), geometrical descriptors (surface or shape properties), and conventional ligand-based descriptors (molecular weight, number of rotatable single bonds, etc.) are considered as well. Then, a variety of machine-learning algorithms, such as random forest, Bayesian classifiers, neural network, and support vector machine, are employed for variable selection. Similar to empirical scoring functions, these methods also need a training set of protein–ligand complexes with known structures and binding data to derive their final models.

Here, we suggest to use the term "descriptor-based scoring functions" to describe these methods. In the literature, some researchers have used the term "machine learning-based scoring function". This term is less instructive since it again emphasizes on the technical aspect rather than the theoretical basis. Besides, its implication is not obvious either because knowledge-based potentials, as discussed in the previous section, are also "machine-learned". One may argue that descriptor-based scoring functions are also empirical scoring functions. This argument is true in a sense. Nevertheless, our opinion is that these methods should be separated from conventional empirical scoring functions for good reasons. First, virtually all empirical scoring functions, as well as physics-based methods and knowledge-based potentials, are in linear functional forms;

DOI: 10.1021/ci500731a
J. Chem. Inf. Model. 2015, 55, 475−482

whereas a descriptor-based scoring function, depending on the machine-learning technique employed, is not necessarily in a linear functional form. A descriptor-based scoring function usually consists of a considerably larger number of descriptors than an empirical scoring function. It is unclear if one can always obtain a converged model if different sets of descriptors are supplied as the input for machine learning. Second, an empirical scoring function adopts a theory-inspired functional form that is predetermined by human; whereas a descriptor-based scoring function relies on machine-learning technique to select the final model. The individual terms in an empirical scoring function normally have interpretable physical meanings; while in the case of a descriptor-based method, the rationale for selecting a certain combination of descriptors is often vague. In this sense, a descriptor-based scoring function is essentially a "black box" as many QSAR models.

Very interestingly, it is repeatedly reported that these methods achieved better correlations to protein−ligand binding data than other types of scoring functions. For example, on the 195 diverse protein−ligand complexes included in the CASF-2007 benchmark,[44] a number of conventional scoring functions produced Pearson correlation coefficients (R) in a range of 0.216−0.644 between their binding scores and experimental binding data. On the same benchmark, the R values produced by SFCscore^{RF}, RF-Score, and ID-Score were 0.779, 0.803, and 0.753, respectively, according to the authors of these methods. The superior performance demonstrated by descriptor-based scoring functions still needs to be understood properly. Some researchers have already expressed their concerns on this type of method.[8] Extensive validations on other benchmarks are perhaps necessary to cast insights into this issue.

SUMMARY

As outlined in this viewpoint, the whole field of scoring function has been evolving continually over the past 20 years. We propose that current scoring functions can be classified into four main categories, i.e. *physics-based methods*, *empirical scoring functions*, *knowledge-based potentials*, and *descriptor-based scoring functions*. The major difference among different types of scoring functions has been discussed here. But one should also keep in mind that different types of scoring functions may be connected on their theoretical basis. Indeed, there are hybridized methods that cannot be easily classified into any category listed above. This article does not intend to be a comprehensive review. We apologize that many important studies in this field are not cited here. We hope that an appropriate classification scheme and naming convention can help the discussion on scoring functions and provide general guidance for newcomers to this field.

Compared to the scenario 20 years ago, scoring functions are standing on a much more solid ground now. Large data sets of protein−ligand complexes in improved quality are available to the public. Some special benchmarks have been created, such as the CASF benchmark[44−46] and the CSAR exercise,[95−97] so that different scoring functions can be compared in a more objective manner. Despite the numerous successful applications of scoring functions, generally speaking, their performance is still not satisfactory. The outcomes of standard benchmarks suggest that current scoring functions can provide promising results in terms of "docking/posing", i.e. predicting the correct ligand binding pose in molecular docking, but their performance is generally poor in terms of "scoring/ranking", i.e. predicting absolute or relative protein−ligand binding affinities. In virtual screening, it is still challenging for scoring functions to differentiate low- to medium-affinity binders from non-binders. Currently, there is simply no consensus on which category of scoring function represents the right direction. Nevertheless, inspired by the progress made over the past years, we are optimistic that new, exciting approaches will be brought up to this field in the coming years.

AUTHOR INFORMATION

Corresponding Author

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

Present Address

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.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Muegge, I.; Rarey, M. Small molecule docking and scoring. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH Inc.: New Jersey, 2001; Vol. *17*, pp 1−60.

(2) 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 Inc.: New Jersey, 2002; Vol. *18*, pp 41−88.

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

(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) Rajamani, R.; Good, A. C. Ranking poses in structure-based lead discovery and optimization: Current trends in scoring function development. *Curr. Opin. Drug Discovery Develop.* **2007**, *10* (3), 308−315.

(6) Huang, S. Y.; Zou, X. Chapter 14. Mean-Force Scoring Functions for ProteinLigand Binding. *Annu. Rep. Comput. Chem.* **2010**, *6*, 281−296.

(7) Liu, Y.; Xu, Z.; Yang, Z.; Chen, K.; Zhu, W. A knowledge-based halogen bonding scoring function for predicting protein-ligand interactions. *J. Mol. Model.* **2013**, *19*, 5015−5030.

(8) Gabel, J.; Desaphy, J.; Rognan, D. Beware of Machine Learning-Based Scoring Functions - On the Danger of Developing Black Boxes. *J. Chem. Inf. Model.* **2014**, *54*, 2807−2815.

(9) Karplus, M.; Weaver, D. L. Protein-Folding Dynamics. *Nature* **1976**, *260*, 404−406.

(10) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4* (2), 187−217.

(11) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **1992**, *13*, 505−524.

(12) Makino, S.; Kuntz, I. D. Automated flexible ligand docking: method and its application for database search. *J. Comput. Chem.* **1997**, *18*, 1812−1825.

(13) Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated docking of flexible ligands: applications of AutoDock. *J. Mol. Recog.* **1996**, *9*, 1−5.

(14) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765−784.

DOI: 10.1021/ci500731a
J. Chem. Inf. Model. 2015, 55, 475−482

(15) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.* **1986**, *7*, 230−252.
(16) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(17) Gilson, M. K.; Given, J. A.; Head, M. S. A new class of models for computing receptor-ligand binding affinities. *Chem. Biol.* **1997**, *4*, 87−92.
(18) Zou, X.; Sun, Y.; Kuntz, I. D. Inclusion of solvation in ligand binding free energy calculations using the generalized Born model. *J. Am. Chem. Soc.* **1999**, *121*, 8033−8043.
(19) Ortiz, A. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. Prediction of Drug Binding Affinities by Comparative Binding Energy Analysis. *J. Med. Chem.* **1995**, *38*, 2681−2691.
(20) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(21) Yin, S.; Biedermannova, L.; Vondrasek, J.; Dokholyan, N. V. MedusaScore: An Accurate Force Field-Based Scoring Function for Virtual Drug Screening. *J. Chem. Inf. Model.* **2008**, *48*, 1656−1662.
(22) Aqvist, J.; Medina, C.; Samuelsson, J. E. New method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **1994**, *7*, 385−391.
(23) Almlof, M.; Brandsdal, B. O.; Aqvist, J. Binding Affinity Prediction with Different Force Fields: Examination of the Linear Interaction Energy Method. *J. Comput. Chem.* **2004**, *25*, 1242−1254.
(24) Carlson, H. A.; Jorgensen, W. L. Extended linear response method for determining free energies of hydration. *J. Phys. Chem.* **1995**, *99*, 10667−10673.
(25) Jones-Hertzog, D. K.; Jorgensen, W. L. Binding affinities for sulfonamide inhibitors with human thrombin using monte carlo simulations with a linear response method. *J. Med. Chem.* **1997**, *40*, 1539−1549.
(26) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc. Chem. Res.* **2000**, *33*, 889−897.
(27) Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M. Validation and Use of the MM-PBSA Approach for Drug Discovery. *J. Med. Chem.* **2005**, *48*, 4040−4048.
(28) Hensen, C.; Hermann, J. C.; Nam, K.; Ma, S.; Gao, J.; Hotje, H. D. A Combined QM/MM Approach to Protein-Ligand Interactions: Polarization Effects of the HIV-1 Protease on Selected High Affinity Inhibitors. *J. Med. Chem.* **2004**, *47*, 6673−6680.
(29) Raha, K.; Merz, K. M. Large-Scale Validation of a Quantum Mechanics Based Scoring Function: Predicting the Binding Affinity and the Binding Mode of a Diverse Set of Protein-Ligand Complexes. *J. Med. Chem.* **2005**, *48*, 4558−4575.
(30) Zhou, T.; Huang, D.; Cafisch, A. Is Quantum Mechanics Necessary for Predicting Binding Free Energy? *J. Med. Chem.* **2008**, *51*, 4280−4288.
(31) Chaskar, P.; Zoete, V.; Röhrig, U. F. Toward On-The-Fly Quantum Mechanical/Molecular Mechanical (QM/MM) Docking: Development and Benchmark of a Scoring Function. *J. Chem. Inf. Model.* **2014**, *54*, 3137−3152.
(32) Gohlke, H.; Klebe, G. Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644−2676.
(33) Gilson, M. K.; Zhou, H.-X. Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21−42.
(34) Zhou, H.-X.; Gilson, M. K. Theory of Free Energy and Entropy in Noncovalent Binding. *Chem. Rev.* **2009**, *109* (9), 4092−4107.
(35) Khoruzhii, O.; Donchev, A. G.; Galkin, N.; Illarionov, A.; Olevanov, M.; Ozrin, V.; Queen, C.; Tarasov, T. Application of a polarizable force field to calculations of relative protein−ligand binding affinities. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 10378−10383.
(36) Bohm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput. Aided. Mol. Des.* **1994**, *8*, 243−256.
(37) Verkhivker, G.; Appelt, K.; Freer, S. T.; Villafranca, J. E. Empirical free energy calculations of ligand-protein crystallographic complexes. I. Knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Eng.* **1995**, *8*, 677−691.
(38) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided. Mol. Des.* **1997**, *11*, 425−445.
(39) Murray, C. W.; Auton, T. R.; Eldridge, M. D. Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of Bayesian regression to improve the quality of the model. *J. Comput. Aided. Mol. Des.* **1998**, *12*, 503−519.
(40) 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.
(41) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739−1749.
(42) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J. Med. Chem.* **2006**, *49*, 6177−6196.
(43) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins: Struct. Funct. Bioinf.* **2003**, *52*, 609−623.
(44) 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.
(45) Li, Y.; Liu, Z. H.; Han, L.; Li, J.; Liu, J.; Zhao, Z. X.; Li, C. K.; Wang, R. X. Comparative Assessment of Scoring Functions on an Updated Benchmark: I. Compilation of the Test Set. *J. Chem. Inf. Model.* **2014**, *54* (6), 1700−1716.
(46) Li, Y.; Han, L.; Liu, Z. H.; Wang, R. X. Comparative Assessment of Scoring Functions on an Updated Benchmark: II. Evaluation Methods and General Results. *J. Chem. Inf. Model.* **2014**, *54* (6), 1717−1736.
(47) Fornabaio, M.; Spyrakis, F.; Mozzarelli, A.; Cozzini, P.; Abraham, D. J.; Kellogg, G. E. Simple, Intuitive Calculations of Free Energy of Binding for Protein-Ligand Complexes. 3. The Free Energy Contribution of Structural Water Molecules in HIV-1 Protease Complexes. *J. Med. Chem.* **2004**, *47*, 4507−4516.
(48) Kerzmann, A.; Neumann, D.; Kohlbacher, O. SLICK - Scoring and Energy Functions for Protein-Carbohydrate Interactions. *J. Chem. Inf. Model.* **2006**, *46*, 1635−1642.
(49) Catana, C.; Stouten, P. F. W. Novel, Customizable Scoring Functions, Parameterized Using N-PLS, for Structure-Based Drug Discovery. *J. Chem. Inf. Model.* **2007**, *47* (1), 85−91.
(50) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, I. E. The Protein Data Bank. *Nucleic Acids Res.* 2000, *28*, 235−242; http://www.rcsb.org/pdb/.
(51) Berman, H. M.; Henrick, K.; Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **2003**, *10*, 98.
(52) Liu, Z. H.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z. X.; Nie, W.; Liu, Y. C.; Wang, R. X. PDB-wide Collection of Binding Data: Current

DOI: 10.1021/ci500731a
J. Chem. Inf. Model. 2015, 55, 475−482

Status of the PDBbind Database. *Bioinformatics* **2014**, DOI: 10.1093/bioinformatics/btu626.
(53) Chen, X.; Liu, M.; Gilson, M. K. Binding DB: A web-accessible molecular recognition database. *J. Combi. Chem. High-Throughput Screen* **2001**, *4*, 719−725.
(54) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.
(55) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−1107.
(56) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krü ger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **2014**, *42*, D1083−D1090.
(57) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623−633.
(58) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's BioAssay Database. *Nucleic Acids Res.* **2012**, *40* (1), D400−D412.
(59) 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.
(60) Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother of All Databases). *Proteins: Struct. Funct. Bioinfor.* **2005**, *60*, 333−340.
(61) Benson, M. L.; Smith, R. D.; Khazanov, N. A.; Dimcheff, B.; Beaver, J.; Dresslar, P.; Nerothin, J.; Carlson, H. A. Binding MOAD, A high-quality protein-ligand database. *Nucleic Acids Res.* **2008**, *36*, D674−D678.
(62) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public Ki Data. *J. Med. Chem.* **2012**, *55*, 5165−5173.
(63) DeWitte, R. S.; Shakhnovich, E. I. SMoG: de Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence. *J. Am. Chem. Soc.* **1996**, *118*, 11733−11744.
(64) Grzybowski, B. A.; Ishchenko, A. V.; Shimada, J.; Shakhnovich, E. I. From Knowledge-Based Potentials to Combinatorial Lead Design in Silico. *Acc. Chem. Res.* **2002**, *35*, 261−269.
(65) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.
(66) Muegge, I. A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. *Perspectives in Drug Discovery & Design.* **2000**, *20*, 99−114.
(67) Muegge, I. Effect of ligand volume correction on PMF scoring. *J. Comput. Chem.* **2001**, *22*, 418−425.
(68) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(69) Velec, H. F. G.; Gohlke, H.; Klebe, G. DrugScore(CSD): Knowledge-Based Scoring Function Derived from Small Molecule Crystal Data with Superior Recognition Rate of Near-Native Ligand Poses and Better Affinity Prediction. *J. Med. Chem.* **2005**, *48*, 6296−6303.
(70) Neudert, G.; Klebe, G. DSX: A Knowledge-Based Scoring Function for the Assessment of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2011**, *51*, 2731−2745.
(71) Huang, S.-Y.; Zou, X. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. *J. Comput. Chem.* **2006**, *27*, 1865−1875.
(72) Huang, S. Y.; Zou, X. An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. *J. Comput. Chem.* **2006**, *27*, 1876−1882.
(73) Huang, S. Y.; Zou, X. Inclusion of solvation and entropy in the knowledge-based scoring function for protein-ligand interactions. *J. Chem. Inf. Model.* **2010**, *50*, 262−273.
(74) Zheng, Z.; Merz, K. M. Development of the Knowledge-Based and Empirical Combined Scoring Algorithm (KECSA) To Score Protein−Ligand Interactions. *J. Chem. Inf. Model.* **2013**, *53*, 1073−1083.
(75) McQuarrie, D. A. *Statistical Mechanics*; Harper Collins Publishers, New York, 1976.
(76) Chandler, D. *Introduction to Modern Statistical Mechanics*; Oxford University Press, New York, 1987.
(77) Ben-Naim, A. Statistical Potentials Extracted from Protein Structures: Are these Meaningful Potentials? *J. Chem. Phys.* **1997**, *107*, 3698−3706.
(78) Thomas, P. D.; Dill, K. Statistical Potentials Extracted from Protein Structures: how accurate are they? *J. Mol. Biol.* **1996**, *257*, 457−469.
(79) Jensen, M. O.; Park, S.; Tajkhorshid, E.; Schulten, K. Energetics of glycerol conduction through aquaglyceroporin GlpF. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 6731−6736.
(80) Tanaka, S.; Scheraga, H. A. Medium-and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* **1976**, *9*, 945−950.
(81) Miyazawa, S.; Jernigan, R. L. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules* **1985**, *18*, 534−552.
(82) Sippl, M. J. Knowledge-based potentials for proteins. *Curr. Opin. Strucut. Biol.* **1995**, *5*, 229−235.
(83) Zheng, Z.; Ucisik, M. N.; Merz, K. M. The Movable Type Method Applied to Protein-Ligand Binding. *J. Chem. Theory Comput.* **2013**, *9*, 5526−5538.
(84) Zheng, M.; Xiong, B.; Luo, C.; Li, S.; Liu, X.; Shen, Q.; Li, J.; Zhu, W.; Luo, X.; Jiang, H. Knowledge-Based Scoring Functions in Drug Design: 3. A Two-Dimensional Knowledge-Based Hydrogen-Bonding Potential for the Prediction of Protein-Ligand Interactions. *J. Chem. Inf. Model.* **2011**, *51*, 2994−3004.
(85) Hansch, C.; Fujita, T. Rho-Sigma-Pi Analysis. Method For Correlation Of Biological Activity + Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86* (8), 1616−1626.
(86) Deng, W.; Breneman, C.; Embrechts, M. J. Predicting Protein-Ligand Binding Affinities Using Novel Geometrical Descriptors and Machine-Learning Methods. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 699−703.
(87) Zhang, S.; Golbraikh, A.; Tropsha, A. Development of Quantitative Structure-Binding Affinity Relationship Models Based on Novel Geometrical Chemical Descriptors of the Protein-Ligand Interfaces. *J. Med. Chem.* **2006**, *49*, 2713−2724.
(88) Durrant, J. D.; McCammon, J. A. NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2010**, *50*, 1865−1871.
(89) Durrant, J. D.; McCammon, J. A. NNScore 2.0: A Neural-Network Receptor-Ligand Scoring Function. *J. Chem. Inf. Model.* **2011**, *51*, 2897−2903.
(90) Ballester, P. J.; Mitchell, J. B. O. A machine learning approach to predicting protein−ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26*, 1169−1175.
(91) Ballester, P. J.; Schreyer, A.; Blundell, T. L. Does a More Precise Chemical Description of Protein−Ligand Complexes Lead to More Accurate Prediction of Binding Affinity? *J. Chem. Inf. Model.* **2014**, *54*, 944−955.
(92) Zilian, D.; Sotriffer, C. A. SFCscore^{RF}: A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2013**, *53*, 1923−1933.
(93) Li, G. B.; Yang, L. L.; Wang, W. J.; Li, L. L.; Yang, S. Y. ID-Score: A New Empirical Scoring Function Based on a Comprehensive

DOI: 10.1021/ci500731a
J. Chem. Inf. Model. 2015, 55, 475−482

Set of Descriptors Related to Protein−Ligand Interactions. *J. Chem. Inf. Model.* **2013**, *53*, 592−600.
(94) Deng, Z.; Chuaqui, C.; Singh, J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions. *J. Med. Chem.* **2004**, *47*, 337−344.
(95) 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.
(96) 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.
(97) 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.

DOI: 10.1021/ci500731a
J. Chem. Inf. Model. 2015, 55, 475−482