JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

ID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein−Ligand Interactions

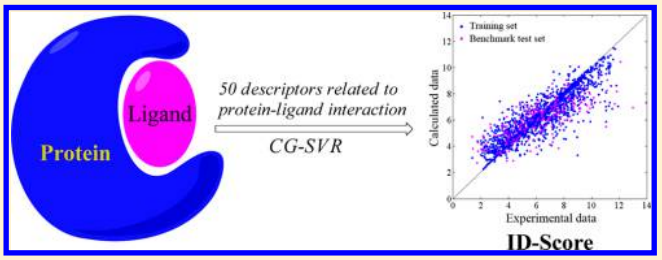
Guo-Bo Li,[†] Ling-Ling Yang,[†,‡] Wen-Jing Wang,[†] Lin-Li Li,[†] and Sheng-Yong Yang*[,†]

[†]State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, West China Medical School, Sichuan University, Sichuan 610041, China

[‡]College of Chemical Engineering, Sichuan University, Sichuan 610041, China

Supporting Information

ABSTRACT: Scoring functions have been widely used to assess protein−ligand binding affinity in structure-based drug discovery. However, currently commonly used scoring functions face some challenges including poor correlation between calculated scores and experimental binding affinities, target-dependent performance, and low sensitivity to analogues. In this account, we propose a new empirical scoring function termed ID-Score. ID-Score was established based on a comprehensive set of descriptors related to protein−ligand interactions; these descriptors cover nine categories: van der Waals interaction, hydrogen-bonding interaction, electrostatic interaction, π-system interaction, metal−ligand bonding interaction, desolvation effect, entropic loss effect, shape matching, and surface property matching. A total of 2278 complexes were used as the training set, and a modified support vector regression (SVR) algorithm was used to fit the experimental binding affinities. Evaluation results showed that ID-Score outperformed other selected commonly used scoring functions on a benchmark test set and showed considerable performance on a large independent test set. ID-Score also showed a consistent higher performance across different biological targets. Besides, it could correctly differentiate structurally similar ligands, indicating higher sensitivity to analogues. Collectively, the better performance of ID-Score enables it as a useful tool in assessing protein−ligand binding affinity in structure-based drug discovery as well as in lead optimization.



INTRODUCTION

Scoring functions have been routinely adopted to estimate the binding affinity between a small molecule and its biological target in structure-based drug discovery, such as molecular docking and de novo design.[1,2] Up to now, a variety of scoring functions have been established, which can be roughly classified into three categories: force field-based, knowledge-based, and empirical.[3,4] Among them, the empirical scoring functions are the most widely used. The empirical scoring functions estimate the binding affinity by explicitly accounting for the various physicochemical terms that can contribute to the binding free energy; these terms may include, for example, van der Waals (vdW), hydrogen bonding, desolvation effect, metal−ligand bonding, etc.[5−9] The coefficients of these terms are usually fit using multiple linear regression methods. This strategy is simple and, hence, computationally efficient. However, most of the empirical scoring functions currently used suffer some limitations, including poor correlation between the calculated scores and experimental binding affinities, target-dependent performance, and low sensitivity to analogues.[4,10] Many factors may contribute to these problems, which include at least the following three aspects. The first is that interaction energy terms involved in scoring functions are limited, generally less than 20. Although having a low number of scoring terms makes the scoring function chemically more interpretable and less prone to overfitting by training, the fewer scoring terms might not be able to sufficiently consider the complexity of interaction between ligands and their biological targets, hence influencing the predictive ability of scoring function. The second is that the training set used to fit the coefficients of the scoring functions is small. Generally speaking, it is difficult to derive a very good prediction model from a small training set. A large training set has recently been demonstrated to benefit the improvement of predictive ability of a scoring function.[5,11] The third is that the coefficients are generally obtained by linear regression. Although linear regression generally works fine, it might fail to consider the cooperative effects of noncovalent interactions because linear regression assumes that the interaction energy terms in a scoring function are independent; such cooperativity has only been considered recently.[12−15]

Recently some progresses in improving the performance of empirical scoring functions have been achieved.[15−20] For example, Kramer et al. adopted the traditional quantitative structure−activity relationship (QSAR) techniques in a proteochemometric approach to predict the binding affinity of protein−ligand complexes.[18] Li et al. employed the support vector machine (SVM) algorithm to derive two different scoring functions, SVR-KB and SVR-EP.

Received: October 15, 2012
Published: February 9, 2013

Table 1. 50 Descriptors Related to Protein-Ligand Interactions Together with Their Categories Used in the Development of ID-Score

category	number of descriptors	descriptors[a]
vdW interaction	20	$K^{vdW}_{C.3-C.3}$, $K^{vdW}_{C.3-C.ar/C.2}$, $K^{vdW}_{C.3-N.3/N.4}$, $K^{vdW}_{C.3-N.pl3/N.am}$, $K^{vdW}_{C.3-O.3}$, $K^{vdW}_{C.3-O.2}$, $K^{vdW}_{C.3-S}$, $K^{vdW}_{C.ar/C.2-C.ar/C.2}$, $K^{vdW}_{C.ar/C.2-N.pl3/N.am}$, $K^{vdW}_{C.ar/C.2-N.4/N.3}$, $K^{vdW}_{C.ar/C.2-O.2}$, $K^{vdW}_{C.ar/C.2-O.3}$, $K^{vdW}_{C.ar/C.2-S}$, K^{vdW}_{O-O}, K^{vdW}_{O-N}, K^{vdW}_{N-N}, $K^{vdW}_{X-C.ar/C.2}$, $K^{vdW}_{X-C.3}$, K^{vdW}_{X-N}, K^{vdW}_{X-O}
H bond interaction	10	$K^{hbond}_{O.2-NH}$, $K^{hbond}_{O.2-OH}$, $K^{hbond}_{O.co2-NH}$, $K^{hbond}_{O.co2-OH}$, $K^{hbond}_{N.ar-NH}$, $K^{hbond}_{N.ar-OH}$, $K^{hbond}_{O.3-NH}$, $K^{hbond}_{O.3-OH}$, $K^{hbond}_{S.3-NHOH}$, K^{hbond}_{F-NHOH}
electrostatic interaction	1	$K^{elect}_{pos-neg}$
π-system interaction	3	$K^{\pi}_{X-\pi}$, $K^{\pi}_{Y-\pi}$, $K^{\pi}_{C.ar-\pi}$
metal−ligand interaction	2	K^{metal}_{O-M}, K^{metal}_{N-M}
desolvation effect	6	$K^{desolv}_{lig-logP}$, $K^{desolv}_{lig-TPSA}$, $K^{desolv}_{lig-Vol}$, $K^{desolv}_{rec-SA/nonSA}$, $K^{desolv}_{lig-SASA}$, $K^{desolv}_{rec-SARA}$
entropy effect	1	K^{conf}_{lig}
shape matching	1	K^{shape}_{SM}
surface property matching	6	$K^{surf}_{pos-pos}$, $K^{surf}_{pos-neg}$, $K^{surf}_{neg-neg}$, $K^{surf}_{pos-nonpolar}$, $K^{surf}_{neg-nonpolar}$, $K^{surf}_{nonpolar-nonpolar}$

[a]A definition for each descriptor is given in Table S4 of the Supporting Information.

The SVR-KB was generated on the features consisting of the knowledge-based pairwise potentials, while the SVR-EP was based on physicochemical properties derived from the crystal complexes in the training set.[19] Very recently, Zheng et al. developed a new empirical scoring function called LISA, in which they designed four kinds of energy terms, including vdW contacts, H-bonding interactions, desolvation effects, and metal chelation, to fit the experimental equilibrium constants (K_d/K_i) using a linear model.[20]

In this investigation, we propose another new empirical scoring function termed ID-Score. Different from other commonly used empirical scoring functions, ID-Score was derived based on a comprehensive set of descriptors associated with protein−ligand interactions and by nonlinear regression. The new or improved strategies adopted in the development of ID-Score led to a better performance of ID-Score in assessing protein−ligand binding affinity than other commonly used empirical scoring functions. Here, we describe the establishment of ID-Score and its performance evaluations.

METHODS

Data Set. The size and quality of a training set used to develop a scoring function have significant influence on the effectiveness and performance of the final scoring function. Wang et al. have been maintaining a database called PDBbind, which provides a collection of crystal structures and experimentally measured binding affinities exclusively for the biomolecular complexes available in the Protein Data Bank (PDB).[21,22] On the basis of the PDBbind database, Wang et al. compiled a "refined set" that provided a high quality set of structures and binding data. The refined set is thought to be very suitable for docking/scoring studies (for more details, see ref 23). Further, they constructed a benchmark set called "core set"; this set was carefully selected through a systematic nonredundant sampling of the refined set with special focus on the diversity of structures and binding data. In this investigation, we shall adopt the latest version of "refined set" (v2011) excluding the "core set" (v2007) as the training set, which contains 2278 complexes. For a fair comparison with other scoring functions, we shall select the "core set" (v2007) consisting of 195 complexes as the benchmark test set. To further validate the scoring function, a large independent test set was constructed, which contains 2310 complexes. These test set complexes were selected from the "general set" collected by Wang et al. according to the following rules. (1) Complexes with the X-ray crystal structure resolution larger than 2.5 Å or solved by NMR techniques were not included. (2) Only complexes with pK_i/pK_d/pIC_{50} values distributed in the range between 2 and 12 were included. (3) Complexes already in "refined set" or "core set" were not included. Detailed information including PDB ID and experimental pK_i/pK_d/IC_{50} values of the training set and the test sets is given in Tables S1, S2, and S3 of the Supporting Information.

Descriptors Related to Protein−Ligand Interactions. A total of 50 descriptors (Table 1) related to protein−ligand interactions were carefully selected for the construction of ID-Score. These descriptors can be classified into nine categories, including van der Waals interaction, hydrogen-bonding interaction, electrostatic interaction, π-system interaction, metal−ligand bonding interaction, desolvation effect, entropic loss effect, shape matching, and surface property matching. A brief description of these descriptors is given as follows (a definition for each of these descriptors is given in Table S4 of the Supporting Information).

Descriptors Related to vdW Interaction. The vdW interaction is one of the most important protein−ligand interactions. Here, descriptors representing vdW interaction are calculated using the Lennard−Jones 6−12 potential formula (eq 1).

$$K^{vdW}_{I-J} = \sum_{i\in I}\sum_{j\in J}\left[\left(\frac{r_i + r_j}{d_{ij}}\right)^{12} - \left(\frac{r_i + r_j}{d_{ij}}\right)^{6}\right] \quad (1)$$

where K^{vdW}_{I-J} denotes the descriptor of vdW interaction between atom type I in the binding site of a protein and atom type J in the ligand (Table 2). d_{ij} is the distance between atom i in the binding site of a protein and atom j in the ligand. To avoid nonphysical attractive or repulsive vdW interaction, we initially set the d_{ij} cutoff of 3 to 5.5.[24] r_i and r_j are the vdW radius of atom i and atom j, respectively. A total of 20 vdW descriptors (Table 1) were involved in the ID-Score scheme.

Descriptors Related to Hydrogen-Bonding Interaction. Hydrogen-bonding interactions are another type of important interactions presented in protein−ligand complexes. In the ID-Score scheme, a similar formula as that given by Vedani et al. (eq 2) was used to calculate the hydrogen-bonding interactions.[25,26]

$$K^{hbond}_{A-D} = \sum_{A}\sum_{D}\left[\left(\frac{r_0}{d_{A-D}}\right)^{12} - 2\left(\frac{r_0}{d_{A-D}}\right)^{6}\right] \times \cos^2(\theta - \theta_0) \quad (2)$$

dx.doi.org/10.1021/ci300493w | J. Chem. Inf. Model. 2013, 53, 592−600

Table 2. AtomTypes Defined in the ID-Score Scheme

ID	atom type	description	vdW radius	atom polarity
1	C.3	carbon sp3 hybridization	1.94	nonpolar
2	C.2	carbon sp2 hybridization	1.90	nonpolar
3	C.1	carbon sp hybridization	1.80	nonpolar
4	C.ar	carbon in aromatic ring	1.85	nonpolar
5	C.cat	carbocation in guadinium	1.90	positively polar
6	N.4	nitrogen positively charged	1.83	positively polar
7	N.3	nitrogen sp3 hybridization	1.86	negatively polar
8	N.2	nitrogen sp2 hybridization	1.87	negatively polar
9	N.1	nitrogen sp hybridization	1.85	negatively polar
10	N.ar	nitrogen in aromatic ring	1.87	negatively polar
11	N.pl3	nitrogen in trigonal planar	1.86	negatively polar
12	N.am	nitrogen in amide	1.83	negatively polar
13	O.3	oxygen sp3 hybridization	1.74	negatively polar
14	O.2	oxygen sp2 hybridization	1.66	negatively polar
15	O.co2	oxygen in carboxylate/ phosphate	1.66	negatively polar
16	S.3	sulfur sp3 hybridization	2.09	negatively polar
17	S.2	sulfur sp2 hybridization	2.09	negatively polar
18	S.o	sulfur in sulfoxide	2.09	positively polar
19	S.o2	sulfur in sulfone	2.09	positively polar
20	P	phosphor	2.03	positively polar
21	F	fluorine	1.77	negatively polar
22	Cl	chlorine	2.00	negatively polar
23	Br	bromine	2.20	negatively polar
24	I	iodine	2.40	negatively polar
25	M	metal cation	0.50	positively polar

where d_{A-D} represents the distance between a hydrogen bond acceptor (HA) and a hydrogen bond donor (HD), and θ is the bond angle between HD and HA (Figure 1A). r_0 and θ_0 denote the optimal distance and angle for a specific pair of HA and HD, respectively, which are dependent on the types of donor atom and acceptor atom. The values of r_0 and θ_0 for different hydrogen-bonding pairs used in the ID-Score scheme are given in Table S5 of the Supporting Information.[20,27−29] In the ID-Score scheme, 10 kinds of descriptors associated with hydrogen-bonding interactions are included (Table 1).

Electrostatic Interaction Descriptor. The electrostatic interaction descriptor used here is computed according to the Coulomb's formula (eq 3).

$$K_{pos-neg}^{elect} = \frac{q_{pos} q_{neg}}{d_{pos,neg}} \tag{3}$$

where $d_{pos,neg}$ represents the distance between a positive charge center and a negative charge center. Definitions for charge centers in proteins and ligands are shown in Figure S1 of the Supporting Information. A distance cutoff of 3.5−6 Å was set to avoid unreasonable electrostatic interactions. In eq 3, q_{pos} and q_{neg} denote the numbers of positive charge and negative charge, respectively.

Descriptors Related to π-System Interaction. The conjugate system or π system widely exists in biomacromolecules and organic small molecules, which also contributes to the protein−ligand interaction, through for example, $\pi-\pi$ interaction, halogen$-\pi$ interaction, and electronegative atom (such as O, N)$-\pi$ interaction.[30−32] However, it is difficult to calculate this type of π-system related interaction accurately. We thus adopted a simple and effective approach to estimate the π-system related interactions, in which we just counted the number of π-system related interactions. Three types of π-system descriptors are involved, including $\pi-\pi$ interaction ($K^{\pi}_{C.ar-\pi}$), halogen$-\pi$ interaction ($K^{\pi}_{X-\pi}$), and electronegative atom$-\pi$ interaction ($K^{\pi}_{Y-\pi}$).

Descriptors Related to Metal−Ligand Interaction. Some metal cations, such as Zn^{2+}, Mg^{2+}, Cu^{2+}, Fe^{3+}, Ca^{2+}, and Mn^{2+}, may exist in binding sites of protein−ligand complexes. In such cases, the bonding interaction between the metal cation and ligand is often considered to be important for the stability of the complex and also contribute to the binding energy. By analyzing the refined data set of PDBbind v2011, it was found that the distances between O/N atoms and metal cations are mainly in the range of 2.0−3.0 Å (Figure S2, Supporting Information). We thus adopted a distance-dependent function (eqs 4 and 5) to calculate the O/N−metal interactions, which was originally proposed by Wang et al.[5]

$$K_{L-M}^{metal} = \sum_{i \in L} \sum_{j \in M} \delta(d_{ij}) \tag{4}$$

$$\delta(d_{ij}) = \begin{cases} 1.0 & d_{ij} < 2.0 \text{ Å} \\ 3.0 - d_{ij} & 2.0 \text{ Å} \le d_{ij} < 3.0 \text{ Å} \\ 0.0 & d_{ij} \ge 3.0 \text{ Å} \end{cases} \tag{5}$$

where L refers to atom O or N in ligand, and M refers to the metal cations in the binding site of protein. d_{ij} denotes the distance between ligand atom i and metal cation j.

Descriptors Related to Desolvation Effect. When a ligand associates with a protein, a bulk of water will be excluded out of the binding site. This process will result in a change in entropy as well as in enthalpy, which is called a desolvation

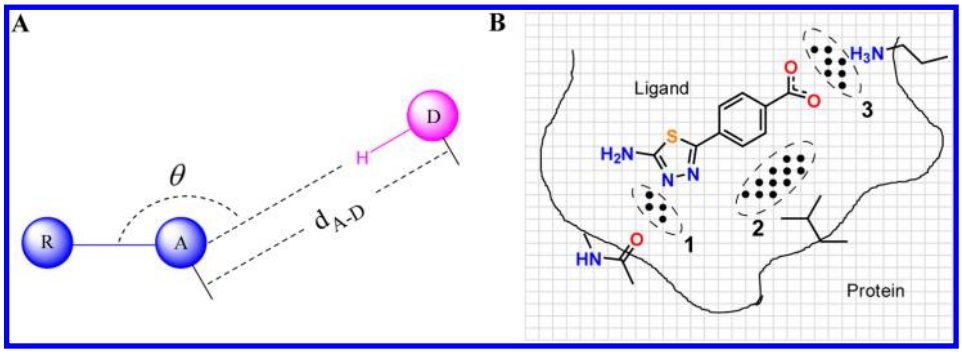


Figure 1. (A) Schematic hydrogen-bonding interaction model. (B) One example showing the calculation method for surface property matching descriptors.

dx.doi.org/10.1021/ci300493w | J. Chem. Inf. Model. 2013, 53, 592−600

effect.[33] In the ID-Score scheme, we introduced six descriptors to consider the desolvation effect. The first is $K^{desolv}_{lig-logP}$, which represents the hydrophilicity/lipophilicity of ligand and was designated as logP of the ligand.[34] The second is $K^{desolv}_{lig-TPSA}$, which denotes the polar surface area (PSA) of ligand; this descriptor was computed by an approach proposed by Ertl et al. based on the summation of tabulated surface contributions of polar fragments.[35] The third is $K^{desolv}_{lig-Vol}$, which is the volume of ligand. The forth is $K^{desolv}_{rec-SA/nonSA}$, representing the ratio of the polar and nonpolar surface areas of the protein active pocket, which is used to characterize the hydrophilic/hydrophobic effect of protein binding site.[36] The last two descriptors are $K^{desolv}_{lig-SASA}$ and $K^{desolv}_{rec-SARA}$, which are used to describe the solvent-accessible surface area (SASA) of ligand and binding site, respectively. These two descriptors are calculated using a similar strategy as that proposed by Zheng et al.[20]

Entropy Effect Descriptor. The entropy effect is another important component of free energy change when a ligand binds to its biomacromolecular target. In the ID-Score scheme, the descriptor related to the ligand conformational entropy was calculated by an empirical equation, eq 6, which is very similar with that used by Eldridge et al.[6]

$$K^{conf}_{lig} = N_{rot} + \sum r[P_{nl}(r) + P'_{nl}(r)]/2 \quad (6)$$

where N_{rot} is the number of frozen rotatable bonds. $P_{nl}(r)$ and $P'_{nl}(r)$ are the percentages of nonlipophilic heavy atoms on either side of the rotatable bond.

Shape Matching Descriptor. Shape complementarity between a ligand and its target binding site is necessary for protein−ligand binding. Thus, we introduced the shape matching descriptor to characterize the shape complementarity. This descriptor was calculated by eq 7.[36]

$$K^{shape}_{SM} = \left(\sum_{i\in L, j\in P} \delta_{ij}\right)/N_{Latom} \quad (7)$$

$$\delta_{ij} = \begin{cases} 1 & d_{ij} \le |r_i + r_j \pm \varepsilon| \\ 0 & d_{ij} > |r_i + r_j \pm \varepsilon| \end{cases} \quad (8)$$

where N_{Latom} is the number of heavy atoms in the ligand. r_i and r_j refer to the vdW radius of atom i and atom j. d_{ij} is the distance between atom i and atom j. ε represents the distance tolerance of the contact of atom i and atom j; here, ε was set to 0.4 Å.

Surface Property Matching Descriptors. In addition to the shape complementarity, the complementarity of physicochemical properties of the surfaces of a ligand and its target binding site is also important for protein−ligand binding. In the ID-Score scheme, the matching of positively polar (pos), negatively polar (neg), and nonpolar (nonpolar) atoms on the surfaces was considered. Thus, there are six matching cases, including pos−pos, pos−neg, pos−nonpolar, neg−neg, neg−nonpolar, and nonpolar−nonpolar, and each one corresponds to a descriptor. Calculations for these descriptors are briefly described as follows. First, each non-hydrogen atom on the surfaces of ligand and protein binding sites was assigned an atom polarity type (pos, neg, or nonpolar) according to its atom type (Table 2). Second, the binding pocket was partitioned into multiple equal-sized rectangular grids. Distances between each grid and each non-hydrogen atom on the surfaces of binding pocket and ligand were then calculated. A grid would be recorded if the distance from the grid to an atom on the binding site surface of protein, and that from the same grid to an atom on the ligand surface, are both less than the vdW radius of the corresponding non-hydrogen atom plus 0.5 Å. Finally, the recorded grids were ascribed to six categories, including pos−pos, pos−neg, neg−neg, pos−nonpolar, neg−-nonpolar, and nonpolar−nonpolar, according to the electrostatic properties of atoms on the surfaces of protein binding sites and ligands that are most close to the grid. Figure 1B shows one example. Grids in region **1** were ascribed to the neg−neg category because they are closest to a protein sp2 oxygen atom (negatively polar) and a ligand aromatic nitrogen atom (negatively polar). Similarly, grids in region **2** and region **3** belong to the nonpolar−nonpolar and pos−neg category, respectively. Finally, the number of grids in each category was added together and then assigned to the corresponding surface property descriptors, $K^{surf}_{pos-pos}$, $K^{surf}_{pos-neg}$, $K^{surf}_{neg-neg}$, $K^{surf}_{pos-nonpolar}$, $K^{surf}_{neg-nonpolar}$, and $K^{surf}_{nonpolar-nonpolar}$.

Support Vector Regression (SVR). Support vector regression (SVR) is a supervised machine learning method. The SVR theory has been well documented in literature.[37,38] Here, we make a short summary to the basic idea of SVR.

Suppose that a given training data $\{(x_1,y_1), ..., (x_l,y_l)\}$: where x_i are the input vectors and y_i are the associated output values of x_i. Basically, the support vector regression is an optimization problem (eq 9)

$$\underset{w,b,\xi,\xi^*}{\text{minimize}} \ \frac{1}{2}w^T w + C\sum_{i=1}^{l}(\xi_i + \xi_i^*)$$

$$\text{subject to} \begin{cases} y_i - w^T\phi(x_i) - b \le \varepsilon + \xi_i \\ w^T\phi(x_i) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, i = 1, ..., l \end{cases} \quad (9)$$

where l denotes the number of samples, b is the bias term, vector of i sample is data set mapped to a higher dimensional space by the kernel function φ vector, ξ_i and $\xi_i^*$are slack variables for measurements above and below the ε-insensitive zone $|y - (w^T\varphi(x) + b)| \le \varepsilon$, and C is the error cost. The basic idea in SVR is to project the input data set x_i into a high-dimensional feature space via a nonlinear manner using kernel functions. Up to now, a variety of kernel functions have been suggested. The Gaussian kernel (RBF, radial basis function) (eq 10) is one of the most popular one, which was also used in this study.

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\gamma^2}\right) \quad (10)$$

where γ^2 denotes the width of Gaussian kernel.

Previous studies have shown that the selection of error cost parameter C and kernel parameter γ has significant influence on the predictive ability of an SVR model.[39,40] We thus adopted the conjugate gradient (CG) method to optimize the two parameters and, hopefully, gained an SVR model with higher quality (for details of the CG algorithms, see ref 41).

Correlation and Performance Metrics. In order to assess the performance of ID-Score, several metrics were used here, including Pearson's correlation coefficient (R_p), Spearman correlation coefficient (R_s), root mean squared error (RMSE), and standard deviation (SD) in linear correlation.[19,23] Details for these metrics are presented in the Supporting Information.

dx.doi.org/10.1021/ci300493w | J. Chem. Inf. Model. 2013, 53, 592−600

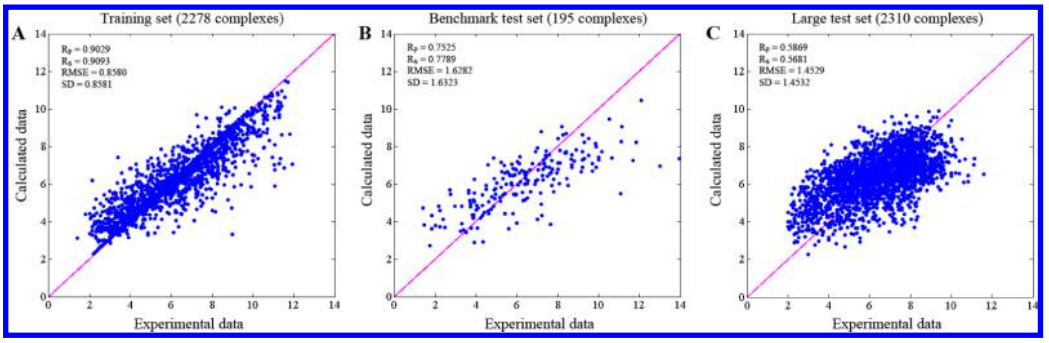


Figure 2. Performances (correlation between calculated scoring function values and experimenal bioactivities) of ID-Score evaluated on the training set (A), benchmark test set (B), and large independent test set (C).

RESULTS AND DISCUSSION

Establishment of ID-Score. Essentially, the establishment of ID-Score is a regression process; the purpose is to develop a relationship between descriptors related to protein−ligand interactions and binding affinities between ligands and their biological targets. The basic process of establishment of ID-Score is briefly described as follows.

The first step is the preparation of a training set. Here, we took the 2011 version of the PDBbind database as our training set, which contains experimentally measured binding affinity data and structural information of 2278 protein−ligand complexes (Table S1, Supporting Information); the PDBbind database is thought to be one of the best databases for various studies on molecular recognition in biological systems in terms of the size and quality. The second step is the calculation of descriptors related to protein−ligand interactions. In ID-Score, a total of 50 descriptors related to protein−ligand interactions were calculated (Table 1). These descriptors fall into nine categories: van der Waals interaction, hydrogen-bonding interaction, electrostatic interaction, π-system interaction, metal−ligand bonding interaction, desolvation effect, entropic loss effect, shape matching, and surface property matching. Because van der Waals and hydrogen-bonding interactions are the most important interactions between proteins and ligands, a set of elaborately selected descriptors were used here to differentiate the interactions between pairs of different atom types. Such a strategy might help to improve the sensitivity of the established scoring function in predicting binding affinity for ligands with little structural difference. In addition, other types of descriptors associated with protein−ligand interactions, such as π-system interaction, shape matching, and surface property matching, are also included. We expected that adoption of this sophisticated and comprehensive set of descriptors would improve the predictive ability of ID-Score. The third step corresponds to the regression process. Different from traditionally used regression processes in which multiple linear regression methods are generally used, we adopted a support vector regression (SVR) method that is a supervised machine learning method. Here, a modified SVR method, CG-SVR, was used, in which a conjugate gradient (CG) method was utilized to optimize the two parameters, C and γ, in SVR (Methods section). The optimized C and γ are 2.397617 and 0.458829, respectively. Finally, the calculated descriptors together with experimentally measured binding affinities for the training set were taken as the input of CG-SVR to generate the SVR prediction model.

The established SVR prediction model was used to calculate the binding affinities for the training set complexes. Figure 2A shows the correlation between experimentally measured binding affinities and those calculated. The calculated R_p, R_s, RMSE, and SD values are 0.9029, 0.9093, 0.8580, and 0.8581, respectively, indicating good consistency between the experimentally measured and calculated binding affinities for the training set complexes.

Evaluation of ID-Score. The performance of ID-Score was first evaluated using the 5-fold cross-validation method; the 5-fold cross-validation method is useful for overcoming the problem of overfitting and minimizing generalization errors.[41,42] In 5-fold cross-validation, first the training set is randomly divided into five subsets equally. Subsequently, one subset is tested using the SVR model trained on the other subsets. This procedure continues until each subset is tested once. Table 3 shows the correlation coefficients (R_p, R_s) and deviations (RMSE, SD) between experimentally measured and calculated affinities for the training set complexes calculated by the 5-fold cross validation. The R_p values for the 5-fold cross validation are in the range of 0.6885−0.7348, averaged at 0.7049. The R_s values are between 0.6822 and 0.7315, and its mean value is 0.7052. The RMSE and SD values are all stable and small (<1.5).

Table 3. Performance of ID-Score on Training Set Evaluated by 5-Fold Cross Validation

no.	R_p	R_s	RMSE	SD
1	0.7348	0.7315	1.2998	1.3012
2	0.6885	0.6922	1.4131	1.4147
3	0.7195	0.7213	1.4427	1.4443
4	0.6889	0.6822	1.4325	1.4341
5	0.6929	0.6987	1.4068	1.4084
average	0.7049	0.7052	1.3990	1.4005

To compare the performance of ID-Score with the commonly used scoring functions, the "core set" (v2007) consisting of 195 complexes (Table S2, Supporting Information) compiled by Wang et al. was adopted as a benchmark test set. Figure 2B depicts the relationship between calculated ID-Score values and experimentally determined binding affinities. The correlation coefficient is 0.7525 for Pearson and 0.7789 for Spearman. The RMSE and SD values are 1.6282, and 1.6323, respectively. Very recently, Cheng et al. used the same "core set" to assess 16 commonly used scoring functions, including LigScore, PLP, PMF, Jain, LUDI, D-Score, PMF-Score, G-Score, ChemScore, F-Score, GlideScore, GoldScore, ChemScore, ASP, DrugScore, and X-Score.[23] In terms of the correlation coefficients (R_p and R_s) and the standard deviation (SD), our ID-Score outperformed all of these scoring functions (Table 4).

Finally, another independent test set containing 2310 complexes (Table S3, Supporting Information) was further used to evaluate

dx.doi.org/10.1021/ci300493w | J. Chem. Inf. Model. 2013, 53, 592−600

Table 4. Performance Comparison of Different Scoring Functions against the Benchmark Set

scoring function	R_p	R_s	SD
SYBYL::F-Score	0.216	0.243	2.35
SYBYL::PMF-Score	0.268	0.273	2.29
GOLD::GoldScore	0.295	0.322	2.29
DS::Jain	0.316	0.346	2.24
SYBYL::D-Score	0.392	0.447	2.19
GOLD::ChemScore	0.441	0.452	2.15
DS::PMF	0.445	0.448	2.14
GlideScore-XP	0.457	0.435	2.14
DS::LigScore2	0.464	0.507	2.12
DS::LUDI3	0.487	0.478	2.09
SYBYL::G-Score	0.492	0.536	2.08
GOLD::ASP	0.534	0.577	2.02
DS::PLP1	0.545	0.588	2.00
SYBYL::ChemScore	0.555	0.585	1.98
DrugScoreCSD	0.569	0.627	1.96
X-Score::HMScore	0.644	0.705	1.83
ID-Score	0.753	0.779	1.63

the performance of ID-Score on a large set of complexes outside of the training set and the "core set". Figure 2C depicts the validation results by ID-Score. The calculated R_p, R_s, SD, and RMSE are 0.5869, 0.5681, 1.4532, and 1.4529, respectively. These results indicate that ID-Score has a considerable predictive ability to the large test set. Here, one may notice that the correlation coefficients (R_p and R_s) are not as larger as those in the case of "core set". One of the main reasons could be that the quality of the test set might not be adequately good. For example, in order to make the test set sufficiently large, the test set includes complexes whose bioactivities are expressed not only by K_i, and K_d, but also by IC_{50}. The bioactivities (pK_i/pK_d/pIC_{50}) are mainly distributed in a range between 4 and 9, which is not as broader as that of the training set or the benchmark test set (Figure 2C).

Assessment of Performance of ID-Score on Different Biological Targets. The performance of ID-Score on different biological targets will be evaluated in this section. For this purpose, we carefully constructed a test set consisting of 20 targets that cover different biological target types. For each target, we collected all the crystal structures of complexes that contain small molecular ligands with their bioactivity data available from RCSB Protein Data Bank (PDB). Complexes presented in the training set or benchmark test set were removed (for a detailed list of all the collected crystal structures of the 20 selected targets, see Table S6 of the Supporting Information). For each target, the performance of ID-Score on the collected complexes was evaluated, which was characterized by R_p, R_s, and SD. Table 5 shows the calculated R_p, R_s, and SD values for each target. The R_p values for the 20 targets are in the range of 0.627−0.851, and averaged at 0.764. The R_s values are between 0.400 and 0.841, and its average is 0.697. Most of the SD values are less than 1.5. These results clearly indicated that ID-Score had considerable and consistent performance on all the test targets, implying that it can be applied in a broad range of biological target types.

Test for Sensitivity of ID-Score. In this section, we will examine the sensitivity of ID-Score; here, sensitivity means the discrimination ability to molecules with little structural difference. For this purpose, we selected five CK2α inhibitors that are very similar in their structures to form a test set (Cpd1−Cpd5, Figure 3).[43] These compounds have almost the same chemical structure except for different heteroatoms in the five-member aromatic ring. However, their bioactivities (IC_{50}) are considerably different from each other. For the evaluation, the crystal structure of CK2α (PDB id: 3AXW) was used as the receptor structure. These compounds were docked into the binding site of CK2α by using the GOLD[44] program (CCDC, Cambridge, U.K.) or LigandFit[45] program in Discovery Studio 3.1 (Accelrys, San Diego, CA, USA). The docking poses of these compounds that are in accordance with the pose of 4-(5-amino-1,3,4-thiadiazol-2-yl) benzoic acid in the crystal structure of CK2α (Figure 3A) were kept for subsequent analysis; the selected poses are depicted in Figure 3B−F. Then Chemscore (GOLD), Goldscore (GOLD), ASP (GOLD), PMF (LigandFit), PLP1 (LigandFit), LigScore2 (LigandFit), and LUDI3 (LigandFit), as well as ID-Score, were calculated for complexes corresponding to these poses. Table 6 lists calculated scores by the seven selected scoring functions and ID-Scores for these compounds. From Table 6, we see that ID-Score gave the same ranking order as that sorted by experimentally measured bioactivities (IC_{50}), and none in the selected seven scoring functions gave a completely correct ranking order. These results show that ID-Score has a considerable ability to correctly sort structurally similar analogues.

Factors Leading to Better Performance of ID-Score. Obviously, the established ID-score displayed a better performance than currently commonly used empirical scoring functions. Several factors could contribute to the better performance of

Table 5. Performances of ID-Score Evaluated against Set Consisting of 20 Selected Diverse Biological Targets

no.	test target[a]	number of complexes	pIC_{50} range	R_p	R_s	SD
1	BACE-1	94	7	0.800	0.697	0.989
2	CDK2	105	7	0.744	0.631	0.845
3	CHK1	21	6	0.791	0.695	1.597
4	DPP4	27	7	0.755	0.400	1.637
5	ER	16	3	0.627	0.554	0.573
6	FTase	10	5	0.715	0.758	1.997
7	GluR2	11	5	0.773	0.698	1.117
8	HIV PR	16	5	0.735	0.659	0.794
9	HSP90	25	6	0.790	0.689	1.106
10	LTA-4H	25	6	0.743	0.817	2.576
11	MMP-13	14	6	0.717	0.742	1.337
12	P38a	22	6	0.751	0.741	1.077
13	PDE4B	28	7	0.775	0.825	1.444
14	PDK1	11	5	0.784	0.791	1.077
15	PPAR	13	4	0.743	0.599	0.656
16	PTP1B	24	5	0.744	0.540	1.087
17	RdRP	11	5	0.797	0.816	0.857
18	Renin	22	4	0.787	0.697	1.075
19	SRC	31	8	0.848	0.841	1.507
20	Thrombin	21	8	0.851	0.745	1.184

[a]BACE-1, beta-secretase 1; CDK2, cell division protein kinase 2; CHK1, serine/threonine-protein kinase chk1; DPP4, dipeptidyl peptidase 4; ER, estrogen receptor; FTase, farnesyltransferase alpha; GluR2, glutamate receptor 2; HIV PR, hiv-1 protease; HSP90, heat shock protein 90; LTA-4H, leukotriene A-4 hydrolase; MMP-13, collagenase 3; P38a, mitogen-activated protein kinase 14; PDE4B, camp-specific 3′,5′-cyclic phosphodiesterase 4b; PDK1, 3-phosphoinositide-dependent protein kinase 1; PPAR gamma, peroxisome proliferator activated receptor; PTP1B, protein tyrosine phosphatase 1B; RdRP, RNA-directed RNA polymerase; SRC, proto-oncogene tyrosine protein kinase src.

dx.doi.org/10.1021/ci300493w | J. Chem. Inf. Model. 2013, 53, 592−600

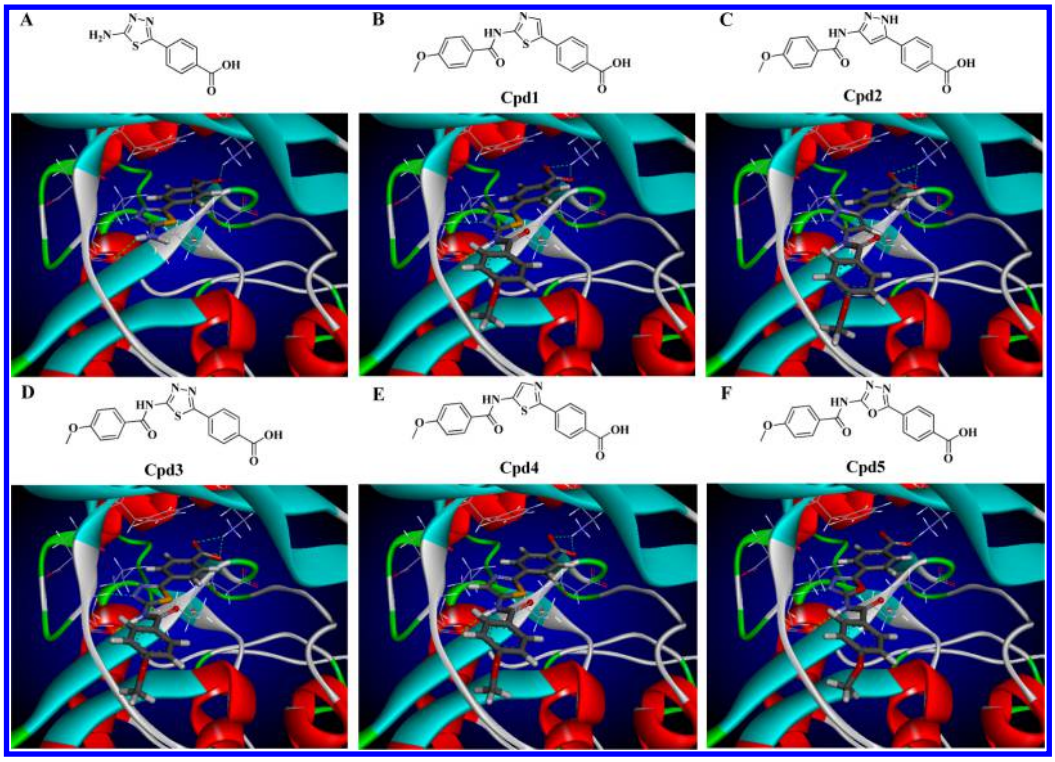


Figure 3. (A) Crystal structure (PDB id: 3AXW) of CK2α in complex with 4-(5-amino-1,3,4-thiadiazol-2-yl)benzoic acid, which was used as the receptor structure in molecular docking. (B−F) Five selected CK2α inhibitors (Cpd1−Cpd5) used for the evaluation of sensitivity of ID-Score together with their binding modes predicted by molecular docking.

Table 6. Evaluation of Sensitivity of Selected Scoring Functions Using a Set of CK2α Inhibitors

ID	exp. (IC_{50}, nM)	ID-Score	GOLD::Chemscore	GOLD::Goldscore	GOLD::ASP	DS::PMF	DS::PLP1	DS::LigScore2	DS::LUDI3
Cpd1	32	7.38	34.57	62.81	39.94	−117.27	−108.05	6.37	572
Cpd2	140	6.30	34.49	60.11	39.77	−118.89	−107.18	6.38	572
Cpd3	3400	5.83	31.76	62.89	38.80	−116.94	−106.38	6.17	571
Cpd4	50000	5.67	32.13	61.83	36.70	−115.14	−98.85	6.14	575
Cpd5	74000	5.34	26.79	60.70	38.08	−114.45	−106.66	6.25	573

ID-Score. The first factor could be the adoption of a set of comprehensive descriptors related to protein−ligand interactions in the development of ID-Score. Essentially, the establishment of an empirical scoring function is a regression process that is to develop a relationship between interaction energy terms and experimental binding affinities. In nature, the protein−ligand interactions are extremely complicated. In commonly used empirical scoring functions, limited descriptor terms were adopted. This might lead to some complicated protein−ligand interactions that were simplified or even ignored, for example, sweeping choices of atom types for protein and ligand atoms or neglect of three-body and higher order effects by approximating atomic interactions as a sum of pairwise terms. In ID-Score, a total of 50 descriptors were elaborately selected, which can reflect various protein−ligand interactions directly or indirectly. The second factor could be the use of SVR for the model regression in ID-Score. In commonly used scoring functions, the coefficients of energy terms are generally obtained by linear regression, which ignored the possible coupling of various energy terms.[12−15] SVR used in the development of ID-Score is a typical nonlinear modeling regression method, which can effectively avoid the inherit shortcomings of linear regression methods.[15] In addition, the adoption of a large training set containing 2278 complexes, which cover a variety of different biological target types, could be another important factor. Apparently, the use of the large training set can help to improve the quality of ID-Score, including prediction ability and applicability to different biological targets.[5,11]

CONCLUSIONS

In summary, we have presented a new empirical scoring function named ID-Score. The basic establishment process and evaluation of ID-Score are described here. ID-Score was developed based on a comprehensive set of molecular descriptors associated with protein−ligand interactions. Evaluation results indicated that ID-Score had a higher performance on a benchmark test set than other selected commonly used scoring functions and a considerable performance on a large independent test set. In a performance assessment on 20 different biological targets, ID-Score showed a consistent higher performance. Finally, ID-Score exhibited a better ability in discriminating analogues than the other selective seven scoring functions in a sensitive test. Collectively, due to its good performance, ID-Score should have a broad application in structure-based drug discovery, as well as in lead optimization. In particular, ID-Score can be reasonably expected to have considerable performance in cases if new compounds are excepted to bind in very similar ways to compounds in the training sets and to be particularly effective as a rescoring function in virtual screening.

dx.doi.org/10.1021/ci300493w | J. Chem. Inf. Model. 2013, 53, 592−600

■ ASSOCIATED CONTENT

Supporting Information

Tables listing detailed information for training set complexes (Table S1), benchmark test set complexes (Table S2), and independent test set complexes (Table S3); definitions for the 50 descriptors used in the ID-Score scheme (Table S4); parameters for different hydrogen-bonding descriptors (Table S5); and a table listing the collected crystal structures for the 20 test targets (Table S6). Figures showing definitions for charge centers in proteins and ligands (Figure S1), and distribution of the distances between ligand O/N atoms and metals of the complexes in PDBbind refined set (Figure S2). Supplementary methods giving calculation methods for the correlation and performance metrics. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: +86-28-85164063. Fax: +86-28-85164060. E-mail: yangsy@scu.edu.cn.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by The Major State Basic Research Development Program of China (2013CB967204), 863 Hi-Tech Program (2012AA020301, 2012AA020308), and The National Science and Technology Major Project (2012ZX09501001-003).

■ ABBREVIATIONS

SVR, support vector regression; SVM, support vector machine; CG, conjugate gradient; QSAR, quantitative structure−activity relationship; vdW, van der Waals; PDB, Protein Data Bank; HA, hydrogen bond acceptor; HD, hydrogen bond donor; PSA, polar surface area; SASA, solvent-accessible surface area; R_p, Pearson's correlation coefficient; R_s, Spearman correlation coefficient; RMSE, root mean squared error; SD, standard deviation

■ REFERENCES

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

(2) Blundell, T. L. Structure-based drug design. *Nature* **1996**, *384*, 23−26.

(3) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935−949.

(4) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein−ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **2006**, *49*, 5851−5855.

(5) Wang, R.; Liu, L.; Lai, L.; Tang, Y. SCORE: A new empirical method for estimating the binding affinity of a protein−ligand complex. *J. Mol. Model.* **1998**, *4*, 379−394.

(6) 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.

(7) Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M. LigScore: A novel scoring function for predicting binding affinities. *J. Mol. Graphics Modell.* **2005**, *23*, 395−407.

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

(9) Raub, S.; Steffen, A.; Kamper, A.; Marian, C. M. AIScore chemically diverse empirical scoring function employing quantum chemical binding energies of hydrogen-bonded complexes. *J. Chem. Inf. Model.* **2008**, *48*, 1492−1510.

(10) 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.

(11) Ballester, P. J.; Mitchell, J. B. A machine learning approach to predicting protein−ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26*, 1169−1175.

(12) Muley, L.; Baum, B.; Smolinski, M.; Freindorf, M.; Heine, A.; Klebe, G.; Hangauer, D. G. Enhancement of hydrophobic interactions and hydrogen bond strength by cooperativity: Synthesis, modeling, and molecular dynamics simulations of a congeneric series of thrombin inhibitors. *J. Med. Chem.* **2010**, *53*, 2126−2135.

(13) Baum, B.; Muley, L.; Smolinski, M.; Heine, A.; Hangauer, D.; Klebe, G. Non-additivity of functional group contributions in protein-ligand binding: A comprehensive study by crystallography and isothermal titration calorimetry. *J. Mol. Biol.* **2010**, *397*, 1042−1054.

(14) Williams, D. H.; Stephens, E.; O'Brien, D. P.; Zhou, M. Understanding noncovalent interactions: Ligand binding energy and catalytic efficiency from ligand-induced reductions in motion within receptors and enzymes. *Angew. Chem., Int. Ed.* **2004**, *43*, 6596−6616.

(15) Kinnings, S. L.; Liu, N.; Tonge, P. J.; Jackson, R. M.; Xie, L.; Bourne, P. E. A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *J. Chem. Inf. Model.* **2011**, *51*, 408−419.

(16) Tang, Y. T.; Marshall, G. R. PHOENIX: A scoring function for affinity prediction derived using high-resolution crystal structures and calorimetry measurements. *J. Chem. Inf. Model.* **2011**, *51*, 214−228.

(17) Rahaman, O.; Estrada, T. P.; Doren, D. J.; Taufer, M.; Brooks, C. L., 3rd; Armen, R. S. Evaluation of several two-step scoring functions based on linear interaction energy, effective ligand size, and empirical pair potentials for prediction of protein-ligand binding geometry and free energy. *J. Chem. Inf. Model.* **2011**, *51*, 2047−2065.

(18) Kramer, C.; Gedeck, P. Global free energy scoring functions based on distance-dependent atom-type pair descriptors. *J. Chem. Inf. Model.* **2011**, *51*, 707−720.

(19) Li, L.; Wang, B.; Meroueh, S. O. Support vector regression scoring of receptor−ligand complexes for rank-ordering and virtual screening of chemical libraries. *J. Chem. Inf. Model.* **2011**, *51*, 2132−2138.

(20) Zheng, Z.; Merz, K. M., Jr. Ligand identification scoring algorithm (LISA). *J. Chem. Inf. Model.* **2011**, *51*, 1296−1306.

(21) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.

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

(23) 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.

(24) 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.

(25) Vedani, A.; Dunitz, J. D. Lone-pair directionality in hydrogen-bond potential functions for molecular mechanics calculations: The inhibition of human carbonic anhydrase II by sulfonamides. *J. Am. Chem. Soc.* **1985**, *107*, 7653−7658.

(26) Vedani, A. YETI: An interactive molecular mechanics program for small-molecule protein complexes. *J. Comput. Chem.* **1988**, *9*, 269−280.

(27) Sarkhel, S.; Desiraju, G. R. N-H...O, O-H...O, and C-H...O hydrogen bonds in protein−ligand complexes: strong and weak interactions in molecular recognition. *Proteins* **2004**, *54*, 247−259.
(28) Panigrahi, S. K. Strong and weak hydrogen bonds in protein−ligand complexes of kinases: A comparative study. *Amino Acids* **2008**, *34*, 617−633.
(29) Muller, K.; Faeh, C.; Diederich, F. Fluorine in pharmaceuticals: Looking beyond intuition. *Science* **2007**, *317*, 1881−1886.
(30) Matter, H.; Nazare, M.; Gussregen, S.; Will, D. W.; Schreuder, H.; Bauer, A.; Urmann, M.; Ritter, K.; Wagner, M.; Wehner, V. Evidence for C-Cl/C-Br...pi interactions as an important contribution to protein−ligand binding affinity. *Angew. Chem., Int. Ed.* **2009**, *48*, 2911−2916.
(31) Grimme, S. Do special noncovalent pi−pi stacking interactions really exist? *Angew. Chem., Int. Ed.* **2008**, *47*, 3430−3434.
(32) Quiñonero, D.; Frontera, A.; Escudero, D.; Ballester, P.; Costa, A.; Deyà, P. M. MP2 study of synergistic effects between X−H/π (X = C,N,O) and π−π interactions. *Theor. Chem. Acc.* **2008**, *120*, 385−393.
(33) 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.
(34) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A.* **1998**, *102*, 3762−3772.
(35) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.
(36) Head, R. D.; Smythe, M. L.; Oprea, T. I.; Waller, C. L.; Green, S. M.; Marshall, G. R. VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.* **1996**, *118*, 3959−3969.
(37) Vapnik, V. N. *Statistical Learning theory*; Wiley Press: New York, 1998; pp 1−55.
(38) Burges, C. J. C. A Tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **1998**, *2*, 121−167.
(39) Wu, C.-H.; Tzeng, G.-H.; Lin, R.-H. A novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression. *Expert Syst. Appl.* **2009**, *36*, 4725−4735.
(40) Chapelle, O.; Vapnik, V.; Bousquet, O.; Mukherjee, S. Choosing multiple parameters for support vector machines. *Mach. Learn.* **2002**, *46*, 131−159.
(41) Yang, S. Y.; Huang, Q.; Li, L. L.; Ma, C. Y.; Zhang, H.; Bai, R.; Teng, Q. Z.; Xiang, M. L.; Wei, Y. Q. An integrated scheme for feature selection and parameter setting in the support vector machine modeling and its application to the prediction of pharmacokinetic properties of drugs. *Artif. Intell. Med.* **2009**, *46*, 155−163.
(42) Schumacher, M.; Binder, H.; Gerds, T. Assessment of survival prediction models based on microarray data. *Bioinformatics* **2007**, *23*, 1768−1774.
(43) Hou, Z.; Nakanishi, I.; Kinoshita, T.; Takei, Y.; Yasue, M.; Misu, R.; Suzuki, Y.; Nakamura, S.; Kure, T.; Ohno, H.; Murata, K.; Kitaura, K.; Hirasawa, A.; Tsujimoto, G.; Oishi, S.; Fujii, N. Structure-based design of novel potent protein kinase CK2 (CK2) inhibitors with phenyl-azole scaffolds. *J. Med. Chem.* **2012**, *55*, 2899−2903.
(44) *GOLD*, version 5.0; Cambridge Crystallographic Data Centre: Cambridge, U.K.
(45) *LigandFit*, Discovery Studio version 3.1; Accelrys: San Diego, CA.