J Mol Model (2013) 19:5015–5030
DOI 10.1007/s00894-013-2005-7

ORIGINAL PAPER

A knowledge-based halogen bonding scoring function for predicting protein-ligand interactions

Yingtao Liu · Zhijian Xu · Zhuo Yang ·
Kaixian Chen · Weiliang Zhu

Received: 22 July 2013 / Accepted: 8 September 2013 / Published online: 27 September 2013

Abstract Halogen bonding, a non-covalent interaction between the halogen σ-hole and Lewis bases, could not be properly characterized by majority of current scoring functions. In this study, a knowledge-based halogen bonding scoring function, termed XBPMF, was developed by an iterative method for predicting protein-ligand interactions. Three sets of pairwise potentials were derived from two training sets of protein-ligand complexes from the Protein Data Bank. It was found that two-dimensional pairwise potentials could characterize appropriately the distance and angle profiles of halogen bonding, which is superior to one-dimensional pairwise potentials. With comparison to six widely used scoring functions, XBPMF was evaluated to have moderate power for predicting protein-ligand interactions in terms of "docking power", "ranking power" and "scoring power". Especially, it has a rather satisfactory performance for the systems with typical halogen bonds. To the best of our knowledge, XBPMF is the first halogen bonding scoring function that is not dependent on any dummy atom, and is practical for high-throughput virtual screening. Therefore, this scoring function should be useful for the study and application of halogen bonding interactions like molecular docking and lead optimization.

Keywords Halogen bonding · Knowledge-based scoring function · Potential of mean forces · Protein-ligand interaction · σ-hole

Electronic supplementary material The online version of this article (doi:10.1007/s00894-013-2005-7) contains supplementary material, which is available to authorized users.

Y. Liu · Z. Xu · Z. Yang · K. Chen · W. Zhu (✉)
Drug Discovery and Design Center, CAS Key Laboratory of Receptor Structure and Function, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China
e-mail: wlzhu@mail.shcnc.ac.cn

Y. Liu
e-mail: ytliu@mail.shcnc.ac.cn

Introduction

In recent years, halogen bonding (XB), a non-covalent interaction, has attracted the interest of many chemists and structural biologists [1–17]. As reported, organic halogens display an anisotropic charge distribution, with an equatorial ring of negative charge and a region of positive charge, namely the σ-hole, along the extension of the R-X bonds [11, 18–22]. Thus, XB is a directional attraction occurring between σ-hole of a halogen atom and a nucleophilic region of another atom, which makes it an important and specific non-covalent interaction in many fields including drug discovery and development [14, 23].

Lots of launched drugs are halogenated compounds, and the halogen atoms are now intentionally introduced in new bioactive entities, owing to the ubiquitous nature of halogens, such as increasing the binding affinity and membrane permeability, facilitating the blood–brain barrier crossing and prolonging the lifetime of the drug, and so on [12, 24–26].

Dobes *et al.* reported that semiempirical quantum mechanical method PM6-DH2X described the geometry and energetics of CK2-inhibitor complexes involving halogen bonds well, while the Amber empirical potentials failed [27]. Indeed, molecular mechanics methods characterize halogens as negatively-charged atoms without anisotropic profile, therefore, they are unable to correctly describe halogen bonding interaction. Recently, some approaches were developed to describe the σ-holes [28–33] by introducing a positively-charged and optionally massless dummy atom. Nevertheless, all these approaches were implemented in molecular mechanical force fields, aiming at studying the dynamic behavior of macromolecules, which are not practical for high-throughput virtual screening due to limited transferability of molecular mechanics force fields, resulting from parameterizations on specific systems. More recently, Kolář and Hobza *et al.* successfully incorporated a molecular-mechanical approach into a docking program suite for the first time [34], which represents the σ-hole with a dummy positive charge as well. Intuitively, to introduce extra dummy atoms equals to an increase of extra degrees of freedom. Furthermore, proper parameterization of the location and the charge of the extra atoms would be challenging owing to the complexity of halogenated moieties and the diversity of possible extra atom construction schemes [31]. To introduce an extra atom is a good way to research σ-holes (*e.g.,* halogen bond), but not the only one. Indeed, Carter *et al.* have developed a set of force-field based potential functions, which are independent of extra dummy atoms, to properly model the anisotropic structure-energy relationships for halogen bonding [35]. However, to the best of our knowledge, there are no scoring functions that describe halogen bonding well, which does not depend on an extra dummy atom and is practical for high-throughput virtual screening at the same time. Thus, it is of interest to develop a scoring function to deal with halogen bonding without introducing any dummy atom.

Current scoring functions can be generally classified into three groups: (*i*) force-field-based [34, 36–39]; (*ii*) empirical [40–46]; and (*iii*) knowledge-based [47–53]. The first two groups of scoring functions are highly dependent on systematic parameterizations on diverse systems. Although many studies have been completed during the last few decades [1, 6–13, 17, 22, 26, 29, 30, 33, 34, 54–56], we still have limited understandings about distance and angle preferences of halogen bonding. Different from the first two groups of scoring functions, knowledge-based scoring functions can model the behavior observed in experimentally determined structures through inference of interaction energy landscapes directly from protein-ligand complexes without any knowledge of other information, and thus are expected to be general.

The fundamental strategy of knowledge-based scoring function is to convert structural information from protein-ligand complexes into distance-dependent pairwise potentials. However, halogen bonding is both distance- and angle-dependent that is similar to hydrogen bonding (HB), therefore, only distance-dependent potentials (1D potential) can not fully characterize the directionality of a XB or a HB. Hence, a multidimensional statistical approach should be employed to study the geometric and energetic preferences of XB [57]. However, there exists an inherent limitation for knowledge-based scoring function because it involves calculation of an ideal reference state, which is theoretically not achievable [58]. In order to circumvent this problem, an iterative method is usually applied to extract the interaction potentials [50, 51, 59].

In this study, we combined the iterative method with the multidimensional statistical model to develop a knowledge-based scoring function based on two high-quality training datasets of protein-ligand complexes. XBs and HBs are characterized by two-dimensional (2D) potentials for various atom types, and the characteristics of their 2D potentials were discussed in order to compare the geometric and energetic preferences of halogen and hydrogen bonding. Subsequently, the scoring function based on the derived potentials, termed XBPMF, was evaluated using a test set with binding affinities and compared to six popular scoring functions. The evaluations demonstrated that the new scoring function is of moderate power for predicting halogen bonding interaction. Thus, the scoring function should be able to facilitate the study and application of halogen bonding.

Materials and methods

Preparation of protein-ligand complexes

A large set of crystal structures of protein-ligand complexes are available in the Protein Data Bank (PDB) [60], hereby serving as the source of our training datasets. Since not all the crystal structures are qualified enough for the purpose of pairwise potential extraction, a number of filters were applied to select the qualified structures:

(1) It must be experimentally determined by X-ray diffraction.
(2) Resolution is better than 3.0 Å.
(3) No DNAs, RNAs or multiple models are included in the crystal structure.
(4) Complexes with severe steric clashes (here, steric clash distance threshold is set as 1.75 Å) are also discarded.
(5) The crystal structure should be composed of at least one protein and one valid ligand. For an entry including one protein and multiple ligands, each ligand and the protein were reassembled as a discrete complex; for an entry including identical ligands in duplicate chains, only the first one was kept.
(6) A valid ligand must fulfill the following criteria: (*i*) it is not covalently bound to any protein or other ligands; (*ii*) it should not contain any metal atoms or other uncommon elements, such as B, Si, Se, *etc.*; (*iii*) it should not be a part of solvent, cofactor, coenzyme or buffer; and (*iv*) it can be an oligopeptide with less than eight residues.

Proteins in the extracted complexes were prepared by Schrödinger software package (version: 2010). Specifically, missing atoms or residues were amended if necessary; all hydrogens in the original PDB file were removed and readded; all waters were retained; all hydrogens were optimized by Protassign module for appropriate protonation and tautomerization states of His residues and appropriate "chi-flips" conformations in Asn, Gln and His residues. Ligands were prepared by OpenBabel (version: 2.3) [61] for adding hydrogens.

Preparation of ligand decoy sets of the training sets

The key idea of an iterative method is to tune the pairwise potentials by iteration until they can discriminate native binding poses from decoy ligand poses, therefore, a diverse set of decoys for the ligands in the training sets were needed. AutoDock (version: 4.2) [62] was used in this study to generate 50 decoy poses (including the native binding pose) for each ligand in the training sets.

Derivation of pairwise potentials

Iterative method to tune potentials

Since the ideal reference state is not achievable for knowledge-based scoring function [58], an iterative method was introduced to circumvent this problem [50, 51, 59]. Based on the assumption of pairwise additivity of atomic interactions, the pairwise potentials were iteratively adjusted using a training set of protein-ligand complexes with a set of decoys for each ligand.

At the beginning, a set of initial values, $u_{ij}^{(0)}(r,\theta)$, for all the pairwise potentials was calculated (see the section: Derivation of the initial potentials), where i and j represent a protein atom type and a ligand atom type, respectively, r is the distance between the two atoms, and θ is the angle of a halogen bond (R-XBD···XBA) or hydrogen bond (HBD-H···HBA) (XBD: halogen bonding donor; XBA: halogen bonding acceptor; HBD: hydrogen bonding donor; HBA: hydrogen bonding acceptor). At the n-th iteration, the binding score, $U_{\mathrm{XBPMF}}^{(n)}$, of every ligand pose for every complex in the training set was calculated by

$$U_{XBPMF}^{(n)} = U_{XB}^{(n)} + U_{HB}^{(n)} + U_{1D}^{(n)} = \sum_{i=1,\cdots,N_P;j=1,\cdots,N_L} u_{ij}^{(n)}(r,\theta), \tag{1}$$

where $U_{\mathrm{XB}}^{(n)}$, $U_{HB}^{(n)}$ and $U_{1D}^{(n)}$ are the sum of the pairwise potentials for halogen bonding, hydrogen bonding and other 1D potentials, respectively. N_P and N_L denote number of protein atoms and ligand atoms, respectively, and $u_{ij}^{(n)}(r,\theta)$ is the pairwise potential for protein atom i and ligand atom j at the n-th iteration.

Then, the best-scored ligand pose was identified for each complex. If the following criterion converges, the iteration stops.

$$\eta = \frac{1}{M}\sum_{m}^{RMSD_m<2} 1 > \eta_0, \tag{2}$$

where M is the number of complexes in the training set, and RMSD_m is the root-mean-square-derivation between the best-scored ligand pose and the native ligand pose for the m-th complex. If RMSD is less than 2 Å, it was recorded as a success. η represents the success rate and η_0 is the predefined convergence threshold.

The pairwise potentials were updated through the iterative process as

$$u_{ij}^{(n+1)}(r,\theta) = u_{ij}^{(n)}(r,\theta) + \lambda k_B T\left(g_{ij}^{(n)}(r,\theta) - g_{ij}^{obs}(r,\theta)\right), \tag{3}$$

where λ is a parameter to control the convergence rate, k_B is the Boltzmann constant, and T is the temperature. $g_{ij}^{\mathrm{obs}}(r,\theta)$ is the experimentally observed pairwise distribution function for the native ligand pose in the training set and $g_{ij}^{(n)}(r,\theta)$ is the predicted pairwise distribution function at the n-th iteration (see the section: Derivation of 2D halogen-bonding and hydrogen-bonding potentials).

At each iteration, the pairwise potentials were updated and the convergence criterion (Eq. (2)) was checked. A final set of pairwise potentials for the prediction of protein-ligand interactions were obtained till the convergence criterion was satisfied.

Derivation of 2D halogen-bonding and hydrogen-bonding potentials

The most widely used statistical model [49] is to extract distance-dependent potentials from crystal structures, which is a 1D model with only one degree of freedom. While both halogen bonding and hydrogen bonding interactions are distance- and angle-dependent, a 1D model is not suitable, thus a 2D model is inferred in a similar way as reported by Zheng *et al.* [57]. Another degree of freedom is designed for halogen bonding and hydrogen bonding, which is R-XBD···XBA or HBD-H···HBA angle (θ). Therefore, for those atom type pairs that can not form a XB or a HB, the extra degree of freedom is automatically lost. In other words, those atom type pairwise potentials were calculated as 1D Muegge's model [49].

In the 2D model, the experimentally observed pair distribution function $g_{ij}^{\mathrm{obs}}(r,\theta)$ was calculated as:

$$g_{ij}^{obs}(r,\theta) = {}^{\rho_{ij}^{obs}(r,\theta)}\Big/_{\rho_{ij,bulk}^{obs}(r,\theta)}, \tag{4}$$

where $\rho_{ij}^{\mathrm{obs}}(r,\theta)$ is the number density of the pair of atoms i from protein and j from ligand at specific distance r and angle θ, and $\rho_{ij,\mathrm{bulk}}^{\mathrm{obs}}(r,\theta)$ is the number density in a reference sphere of radius $R_{\max}$ ($R_{\max}$=10 Å). In order to calculate the number densities for halogen bonding and hydrogen bonding interactions, the surrounding three-dimensional (3D) space of a halogen bonding donor (XBD) or a hydrogen donated by a hydrogen bonding donor is radially divided into multiple spherical bins (geometrical parameters for a bin are Δr= 0.1 Å and $\Delta\theta$=5°). The volume $V(r,\theta)$ of a bin at specific distance r and angle θ is calculated as

$$V(r,\theta) = \frac{4}{3}\pi\left((r+\Delta r)^3 - r^3\right)\sin\left(\theta + \frac{\Delta\theta}{2}\right)\sin\frac{\Delta\theta}{2}. \tag{5}$$

So the number densities $\rho_{ij}^{\mathrm{obs}}(r,\theta)$ and $\rho_{ij,\mathrm{bulk}}^{\mathrm{obs}}(r,\theta)$ were calculated as

$$\rho_{ij}^{obs}(r,\theta) = \frac{1}{M}\sum_{m}^{M}\frac{n_{ij}^{m}(r,\theta)}{V(r,\theta)} \tag{6}$$

$$\rho_{ij,bulk}^{obs}(r,\theta) = \frac{1}{M}\sum_{m}^{M}\frac{N_{ij}^{m}}{V(R_{\max})} \tag{7}$$

$$V(R_{\max}) = \frac{4}{3}\pi R_{\max}^{3}, \tag{8}$$

where $n_{ij}^{m}(r,\theta)$ and N_{ij}^{m} are the numbers of atom pair ij in the spherical bin and the reference sphere for the m-th experimentally observed complex, respectively, and $V(R_{\max})$ is the volume of the reference sphere.

Based on the decoys for each ligand in each complex, the predicted pairwise distribution function $g_{ij}^{(n)}(r,\theta)$ was calculated as

$$g_{ij}^{(n)}(r,\theta) = {}^{\rho_{ij}^{(n)}(r,\theta)}\Big/\rho_{ij,bulk}^{(n)}(r,\theta), \tag{9}$$

where number densities $\rho_{ij}^{(n)}(r,\theta)$ and $\rho_{ij,\mathrm{bulk}}^{(n)}(r,\theta)$ at the n-th iteration were calculated over different decoys as follows:

$$\rho_{ij}^{(n)}(r,\theta) = \frac{1}{ML}\sum_{m}^{M}\sum_{l}^{L}\frac{n_{ij}^{ml}(r,\theta)e^{-\beta U_{ml}^{(n)}}}{V(r,\theta)} \tag{10}$$

$$\rho_{ij,bulk}^{(n)}(r,\theta) = \frac{1}{ML}\sum_{m}^{M}\sum_{l}^{L}\frac{N_{ij}^{ml}e^{-\beta U_{ml}^{(n)}}}{V(R_{\max})} \tag{11}$$

$$\beta = {}^{1}\big/_{k_B T}, \tag{12}$$

where L is the number of decoys for each ligand prepared by AutoDock, and $n_{ij}^{ml}(r,\theta)$ and N_{ij}^{ml} are the numbers of the atom pair ij in the spherical bin and the reference sphere for the l-th decoy of the m-th complex, respectively. $U_{ml}^{(n)}$ was the binding score of the l-th decoy of the m-th complex calculated by Eq. (1).

Derivation of the initial potentials

For the iterative process in Eq. (3), the initial pairwise potentials $u_{ij}^{(0)}(r,\theta)$ needed to be assigned, which were defined as extracted potentials $w_{ij}(r,\theta)$ from experimentally observed complexes in the training set.

$$w_{ij}(r,\theta) = -k_B T \ln g_{ij}^{obs}(r,\theta). \tag{13}$$

We ignored the potentials of the atom type pairs whose occurrences were statistically insufficient (<500) [50], in other words, the pairwise potentials $w_{ij}(r,\theta)$ or $u_{ij}^{(n+1)}(r,\theta)$ for low occurrences were set to zero. And if no atom type pair ij was found in a certain spherical bin, the corresponding potential in this bin was set to 3 kcal mol^{-1} [49, 50, 57] as an unfavorable interaction.

As indicated by Huang *et al.* [50, 51], an effective short-distant repulsive component was necessary for avoiding steric clashes, therefore, we adopted the same repulsive component as Huang *et al.*. The Lennard-Jones 6–12 potentials, $v_{ij}(r,\theta)$, were introduced in Eq. (14),

$$v_{ij}(r,\theta) = \frac{\varepsilon r_{eqm}^{12}}{r^{12}} - \frac{2\varepsilon r_{eqm}^{6}}{r^{6}}, \tag{14}$$

where the equilibrium radii r_{eqm} were taken from the AMBER force filed [63] and the well depths ε were set to three times of the corresponding value in the AMBER force field [50]. In order to remove possible fluctuations in large distances, the initial potentials were set to zero when the distance was larger than r_c=6 Å. Therefore, the initial potentials $u_{ij}^{(0)}(r,\theta)$ were calculated as

$$u_{ij}^{(0)}(r,\theta) = \begin{cases} w_{ij}(r,\theta) & r{\le}r_c\ \textit{for XB\&HB pairs} \\ \dfrac{v_{ij}(r,\theta)e^{-v_{ij}(r,\theta)} + w_{ij}(r,\theta)e^{-w_{ij}(r,\theta)}}{e^{-v_{ij}(r,\theta)} + e^{-w_{ij}(r,\theta)}} & r{\le}r_c\ \textit{for}\ \mathrm{1D}\ \textit{pairs} \\ 0 & r > r_c\ \textit{for all} \end{cases} . \tag{15}$$

Preparation of test set

In order to evaluate the scoring function in this study, we selected a diverse set of 162 protein-ligand complexes with binding affinities from the PDBbind database (version: 2012) [64, 65] as the primary test set, in which all the ligands are halogenated (including Cl, Br or I). There are over 9000 protein-ligand complexes with binding affinities in PDBbind database. Before preparing the primary test set, systematic mining and filtering were implemented, which were summarized as follows:

(1) Ligand in the complex must contain at least one halogen, such as Cl, Br or I, since the knowledge-based scoring function we developed was designed for halogen bonding.
(2) Resolution is better than 2.5 Å.
(3) Only the protein-ligand complexes with known dissociation constants (K_d) or inhibition constants (K_i) were considered.
(4) Ligand is not covalently bound to the protein.
(5) There should be only one protein and one ligand in the complex.
(6) Ligand contains no uncommon elements, such as B, Si, Se, *etc.*
(7) Molecular weight of the ligand should not exceed 1000.
(8) Oligopeptide with less than eight residues and oligonucleotides with no more than three residues are also considered as valid ligands.
(9) Protein in the complex must be complete.

In order to evaluate the powers of the scoring function more pertinently, we selected a subset of complexes in which typical halogen bonds formed. We obtained two secondary test subsets: TestSet-S1 (halogen bond distance <=3.5 Å, halogen bond angle >=140°, size: 24) and TestSet-S2 (halogen bond distance <=3.2 Å, halogen bond angle >=140°, size: 7). In addition, in order to evaluate the "ranking power" (refer to "Evaluation methods" section for details) of the scoring function, 162 complexes in the primary test set were clustered with a sequence similarity cutoff of 90 %. Eight clusters with no less than five members were extracted for Spearman correlation analysis, which were carbonic anhydrase II (CA, size: 13), casein kinase-1 (CK, size: 9), coagulation factor X (CFX, size: 16), heat shock protein 90-alpha (HSP, size: 12), human immunodeficiency virus protease (HIVP, size: 6), tyrosine-protein phosphatase non-receptor type 1 (TPPNRT, size: 12), beta-trypsin (BT, size: 16), and urokinase-type plasminogen activator (UTPA, size: 5).

Besides, a set of decoys for each ligand in the primary test set were necessary to be prepared for evaluating the "docking power" of a scoring function (see "Evaluation methods" for details). In order to sample the binding poses of a ligand as completely as possible, three molecular docking tools, including Glide in Schrödinger (version: 2010), AutoDock (version: 4.2) [62] and Dock (version: 6.5) [66, 67], were used to generate the initial set of decoy poses. For each ligand, two initial conformations were prepared before implementing a docking job, including the native conformation in the crystal structure and a random conformation outside the binding pocket. In this regard, about 200 binding poses for each ligand were generated for each docking tool, resulting in an initial set of ~600 binding poses for each ligand. Then, the ~600 binding poses were clustered into 100 clusters according to the RMSD of the native binding pose. The non-covalent interaction energies between each pose and its receptor were calculated by SYBYL, and then the pose with the lowest energy in each cluster was extracted to compose the decoy set for each complex. Therefore, 100 nonredundant and low-energy binding poses of the ligand for each complex were generated.

In addition, some scoring functions are sensitive to steric clashes because there are some repulsive terms in the functions, so it is quite possible that unexpected binding scores might be computed for some complexes. To address this problem, all the complexes in the three test sets were optimized by SYBYL with the protein fixed. Therefore, two separate results based on the original and optimized complexes were evaluated and discussed.

Evaluation methods

The evaluation for our scoring function was performed with the methods developed by Wang *et al.* [68], which were "docking power", "ranking power" and "scoring power".

Docking power

The ability to identify the native binding pose from a couple of decoys is defined as "docking power". The scoring function was utilized to score all the decoys for each complex in the test set, and the RMSD between each decoy and the native binding pose was calculated. If the RMSD for one of the best-scored decoys was less than a threshold, for example, 2.0 Å, a success was counted. Finally, the overall success rate on the test set was measured as the docking power of the scoring function.

Ranking power

The ability to reproduce the rank of a couple of ligands bound to a common protein according to their binding affinities is defined as "ranking power". As described earlier, we extracted eight clusters of protein-ligand complexes. Each cluster consisted of a number of complexes formed between a protein and its various ligands with different binding affinities. The scoring functions were applied to score each complex in the eight clusters, and then a Spearman correlation analysis was implemented to examine whether the rank of the binding scores for all the complexes in each cluster was consistent with their binding affinities. The larger the Spearman correlation coefficient, the stronger the "ranking power" of the scoring function. Spearman correlation coefficient of "1" stands for identical order between the rank of the binding scores and the binding affinities of all the complexes, while "-1" represents a totally reversed order.

Scoring power

The ability to correlate the predicted binding scores and the experimentally determined binding affinities of a diverse set of complexes in a linear way is defined as "scoring power". The scoring power of each scoring function on three test sets, which were the primary test set and the two typical halogen bonding sets (TestSet-S1, TestSet-S2), was measured by the Pearson correlation coefficient between the binding scores and the binding affinities of all the complexes. There are some cases that scoring functions fail to compute a favorable score for a specific protein-ligand complex due to various reasons, in these cases the complex would not be counted in the Pearson correlation analysis.

Results and discussion

Training sets of protein-ligand complexes

Through a couple of filtering steps, a total of 31,145 complexes were obtained from the Protein Data Bank, among

Table 1 Logarithm of selected HB and XB donor-acceptor pair occurrences in two training sets

$\log_{10}(N_{ij})$		Protein HB donor atom type (i)							
		TraningSet-1				TrainingSet-2			
		NC	ND	OD	SD	NC	ND	OD	SD
Ligand HB acceptor atom type (j)	OA	3.60	4.51	3.73	3.05	4.81	5.66	4.89	3.93
	OD	3.46	4.70	3.92	2.69	5.24	6.37	5.58	4.21
	NA	2.10	3.06	2.29	1.34	3.01	3.92	3.18	2.27
	ND	3.44	4.75	3.92	2.95	4.46	5.71	4.91	3.76
	N0	1.98	3.20	2.23	1.04	2.53	3.67	2.72	1.70
	SA	2.76	3.79	2.96	2.49	3.64	4.65	3.86	3.00
$\log_{10}(N_{ij})$	Ligand HB donor atom type (j)								
		NC	ND	OD	SD	NC	ND	OD	SD
Protein HB acceptor atom type (i)	OC	3.18	3.68	3.60	0.95	4.37	4.68	5.36	3.27
	OA	3.76	4.41	4.35	1.83	4.95	5.36	6.02	4.02
	OD	3.13	3.92	3.92	0.95	4.28	4.91	5.58	3.23
	NA	2.37	3.09	3.10	0.60	3.75	4.11	4.87	2.85
	ND	3.81	4.75	4.70	1.91	5.00	5.71	6.37	4.06
	SA	2.26	2.89	2.90	NA	3.39	3.82	4.38	2.47
$\log_{10}(N_{ij})$	Ligand XB donor atom type (j)								
		Cl			Br		I		
Protein XB acceptor atom type (i)	OC	3.65			3.10		2.77		
	OA	4.54			3.97		3.66		
	OD	3.73			3.10		2.84		
	NA	3.33			2.75		2.33		
	ND	4.58			4.00		3.69		
	SA	3.13			2.45		2.33		

Refer to Tables S1 and S2 for explanations of the names of the protein atom types and the ligand atom types

which 1591 complexes have halogenated ligands. In order to analyze the possible effects of different training sets on the extracted pairwise potentials, two separate training sets were prepared: one, namely TrainingSet-1, consisted of all 1591 halogenated complexes, and the other, namely TrainingSet-2, consisted of all 31,145 complexes. In addition, we adopted a set of atom types (see Tables S1 and S2 in Supporting information), which is an updated version of Muegge's atom types [49], including 17 protein atom types and 31 ligand atom types.

Extracted pairwise potentials

Before extracting the pairwise potentials, occurrences of all the possible atom type pairs were recorded. For good statistics, the pairwise potentials of only those atom type pairs whose occurrences were no less than 500 ($\log_{10}500\approx2.70$) were retained. Thus, based on TrainingSet-1, there were 321 pairs of effective 1D potentials, 28 pairs of effective 2D HB potentials and 15 pairs of effective 2D XB potentials. Likewise, based on TrainingSet-2, there were 378 pairs of effective 1D

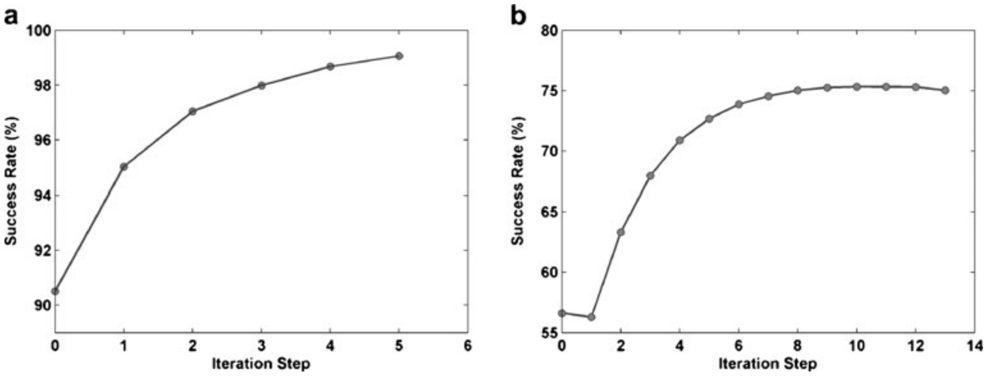


Fig. 1 Convergence parameter η as a function of the iteration step based on (**a**) TrainingSet-1 and (**b**) TrainingSet-2

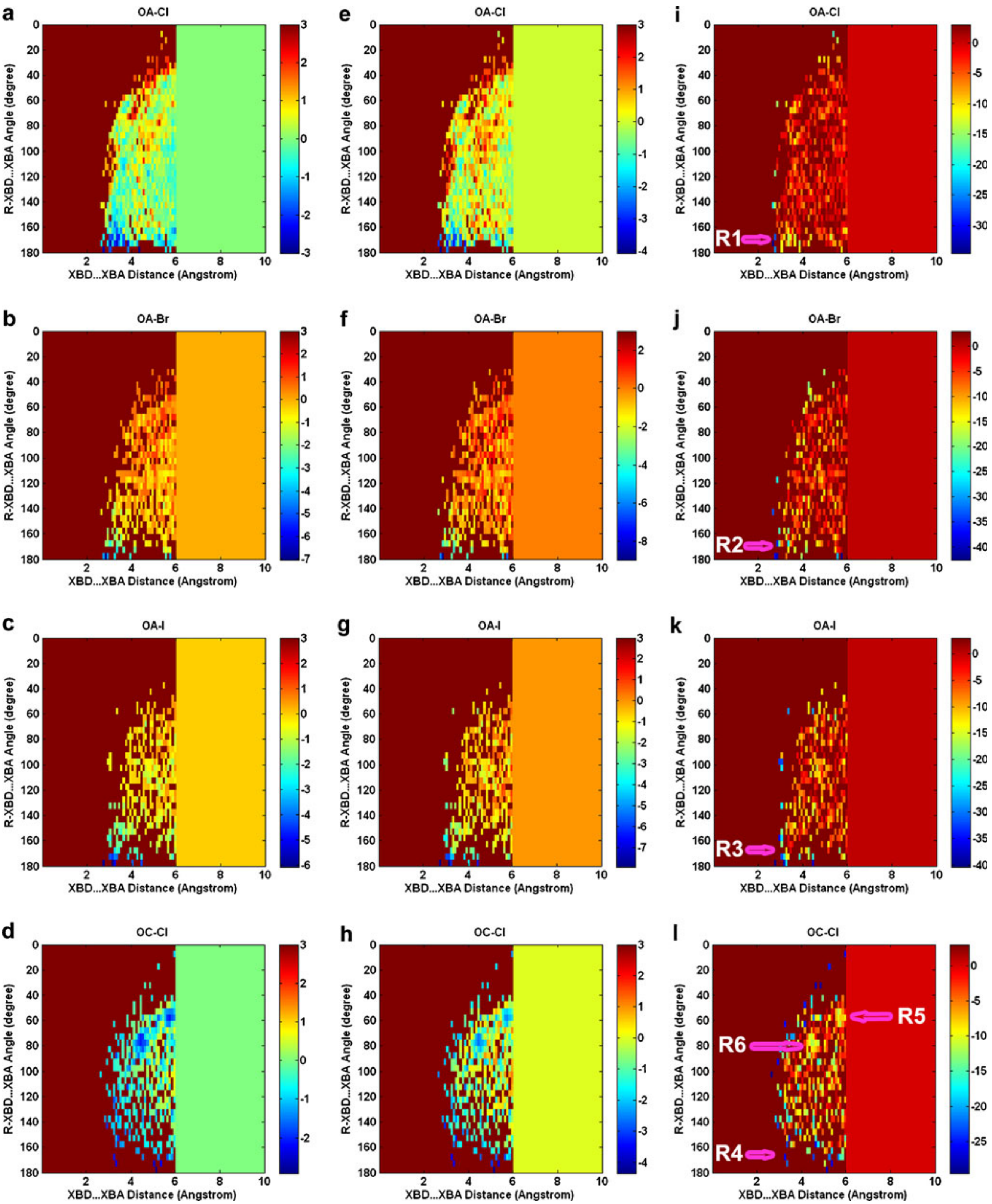
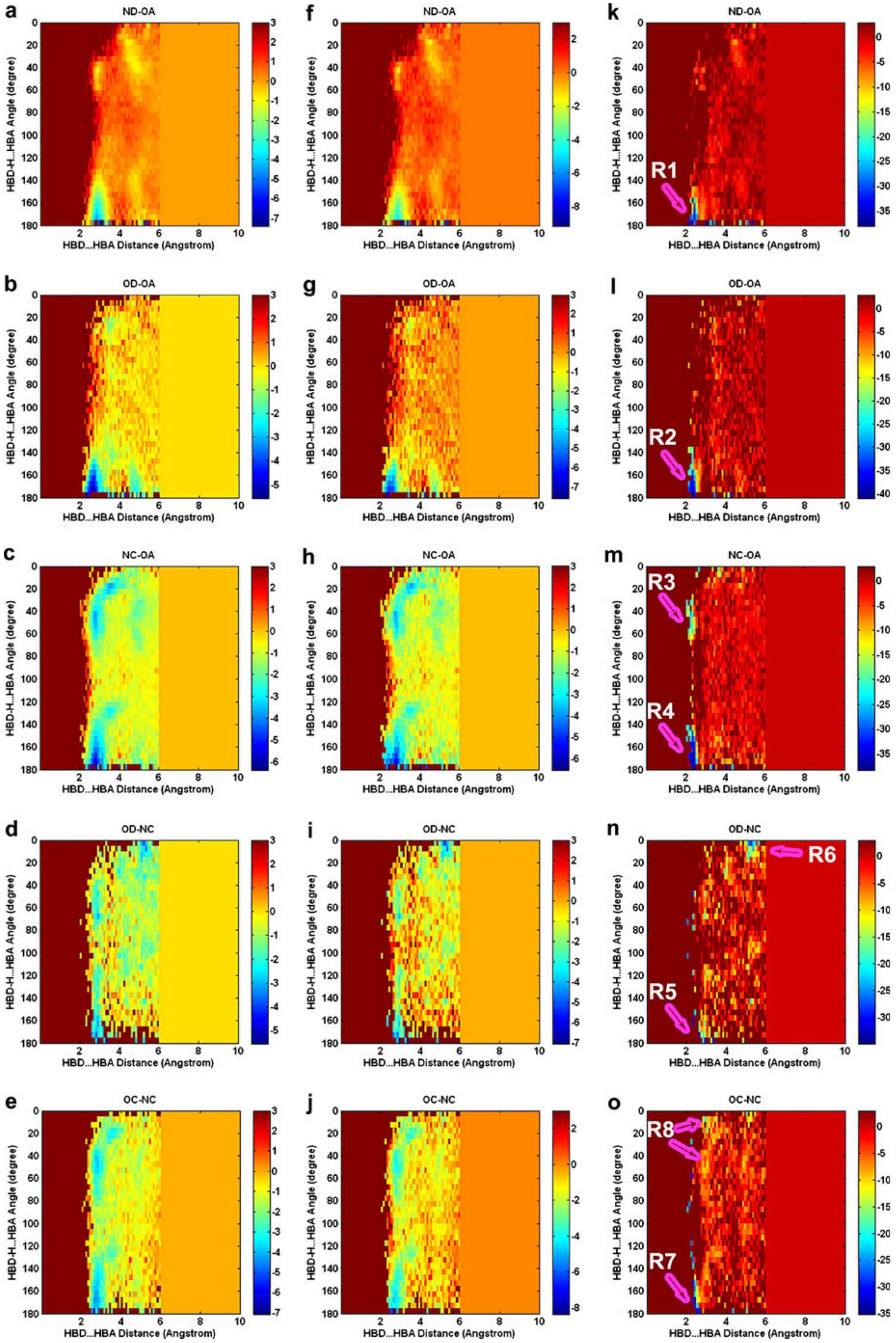


Fig. 2 Heat maps of 2D XB potentials for four selected donor-acceptor pairs based on TrainingSet-2: (a), (e), (i) OA-Cl; (b), (f), (j) OA-Br; (c), (g), (k) OA-I; (d), (h), (l) OC-Cl. R1-R6 stand for favorable interaction regions. The four- or five-letter code, *e.g.*, OA-Cl, refers to the atom type pair, where the letters before the dash, OA, refer to protein XB acceptor atom type, and the letters after the dash, Cl, refer to ligand XB donor atom type. (**a**)-(**d**) are heat maps of 2D XB potentials extracted from observed complexes before iteration starts; (**e**)-(**h**) are heat maps of 2D XB potentials at the first iteration; and (**i**)-(**l**) are heat maps of 2D XB potentials at the converged iteration


a
ND-OA
f
ND-OA
k
ND-OA
R1
b
OD-OA
g
OD-OA
l
OD-OA
R2
c
NC-OA
h
NC-OA
m
NC-OA
R3
R4
d
OD-NC
i
OD-NC
n
OD-NC
R6
R5
e
OC-NC
j
OC-NC
o
OC-NC
R8
R7
HBD-H...HBA Angle (degree)
HBD...HBA Distance (Angstrom)
0
20
40
60
80
100
120
140
160
180
2
4
6
8
10

◀ **Fig. 3** Heat maps of 2D HB potentials for five selected donor-acceptor pairs based on TrainingSet-2: (a), (f), (k) ND-OA; (b), (g), (l) OD-OA; (c), (h), (m) NC-OA; (d), (i), (n) OD-NC; (e), (j), (o) OC-NC. R1-R8 stand for favorable interaction regions. The five-letter code refers to the atom type pair, where the letters before the dash refer to protein HB donor atom type, and the letters after the dash refer to ligand HB acceptor atom type. (**a**)-(**e**) are heat maps of 2D HB potentials extracted from observed complexes before iteration starts; (**f**)-(**j**) are heat maps of 2D HB potentials at the first iteration; and (**k**)-(**o**) are heat maps of 2D HB potentials at the converged iteration

potentials, 40 pairs of effective 2D HB potentials and 15 pairs of effective 2D XB potentials. Based on the two training sets, the occurrences of selected HB and XB donor-acceptor pairs in logarithm scale were listed in Table 1.

The convergence parameters η, defined in Eq. (2), based on the two training sets are shown in Fig. 1. As for the smaller TrainingSet-1, the convergence parameter rapidly reached 99 % in five iterations (Fig. 1a), while for the bigger TrainingSet-2, the convergence parameter converged at ~75 % quickly as well (Fig. 1b), indicating that the majority of native binding poses of the protein-ligand complexes in the training sets were successfully identified based on the extracted pairwise potentials. Besides, fast convergence also demonstrated the effectiveness of our method.

Eventually, three representative sets of extracted pairwise potentials were selected for evaluation, termed XBPMF1, XBPMF2 and XBPMF3, respectively. XBPMF1 is the extracted potentials based on TrainingSet-1 at the converged iteration, while XBPMF2 and XBPMF3 are the extracted potentials at the first and the converged iterations, respectively, based on TrainingSet-2.

Characteristics of 2D halogen bonding potentials

Heat maps of 2D halogen bonding pairwise potentials for four atom type pairs based on TrainingSet-2 are shown in Fig. 2 (refer to Fig. S1 in Supporting information for the heat maps based on TrainingSet-1). In order to analyze what changed and what resulted in the convergence during the iteration process, all the pairwise potentials at the native state (Fig. 2a-d), the first iteration (Fig. 2e-h) and the converged iteration (Fig. 2i-l) were displayed. In general, during the iteration process, the minimal potentials of favorable XB pairs (deep blue region) was enlarged from $-3\sim-7$ kcal mol^{-1} to $-30\sim-40$ kcal mol^{-1} for all four atom type pairs. Similar trends were observed in the heat maps of 1D (Fig. S2) and HB (Fig. 3) pairwise potentials, indicating that the iteration process could enhance the discrimination capability between the favorable (negative potential) and unfavorable interactions (positive potential). For example, Fig. 2a and i show that the potential threshold for a strong halogen bonding interaction between OA (oxygen as XB acceptor) and Cl (chlorine as XB donor) was about -2 kcal mol^{-1} and -30 kcal mol^{-1}, respectively, and likewise for the other three atom type pairs (OA-Br, OA-I, OC-Cl). In addition, the favorable region (blue region) shrunk for the four atom type pairs during the iteration process, demonstrating that the iterative pairwise potentials became more sensitive to geometric location (interaction distance and angle) between the two interaction atom types.

Figure 2i depicts the halogen bonding interactions between OA (oxygen as XB acceptor) and Cl (chlorine as XB donor) at the converged iteration. If the threshold for a halogen bond is set to -25 kcal mol^{-1}, the favorable halogen bonding region is restricted to a small region R1: distance [2.6 2.8] Å, angle [165° 180°], and similar results were obtained for atom type pairs OA-Br (Fig. 2j) and OA-I (Fig. 2k) (Table 2), demonstrating that halogen bonding is angle-dependent and consistent with the anisotropic charge distribution of halogens. In addition, with the decrease of the R-XBD XBA angle, the repulsive region (potential: 3 kcal mol^{-1}) grows bigger and bigger, which also agrees well with negative charge distribution around the equatorial region of halogens. Thus, 2D pairwise potential is superior to 1D potential for properly characterizing halogen bonding.

As shown in Table 2, the geometric and energetic parameters of optimal interaction for OC-Cl is different from OA-Cl, OA-Br and OA-I, which could be attributed to the complicated interaction that the negatively charged OC is involved in which is not only halogen bonding interaction but also electrostatic interaction with Cl. Apparently, there is only one favorable interaction region between neutral XB acceptor and donor, for example, R1 region in Fig. 2i, R2 region in Fig. 2j, and R3 region in Fig. 2k, which should be caused by the angular preference of halogen bonding. However, Fig. 2l revealed three favorable regions (R4, R5 and R6), which is possibly caused by long-range electrostatic interaction between OC and Cl. Table 3 summarizes the mean potentials of selected XB atom type pairs at statistically effective regions (occurrences >=500, potential <3.0 kcal mol^{-1}, distance <=6.0 Å). Interestingly, the pairwise potentials have an order of OA-Cl>OA-Br>OA-I, which is consistent with the reported results: the strength of the interaction decreases in the following order I>Br>Cl [11]. Meantime, strengthened by attractive electrostatic interaction, the mean pairwise potential of OC-Cl is comparable to OA-I.

Characteristics of 2D hydrogen bonding potentials

Heat maps of 2D hydrogen bonding pairwise potentials for five atom type pairs based on TrainingSet-2 are shown in Fig. 3 (refer to Fig. S3 in Supporting information for the heat maps based on TrainingSet-1). For neutral atom pairs: ND-OA (Fig.ure 3k) and OD-OA (Fig. 3l), the favorable hydrogen bonding region is restricted to R1 (threshold: -25 kcal mol^{-1}, distance ∊ [2.1 2.6] Å, angle ∊ [160° 180°])

Table 2 Geometric and energetic parameters of optimal interactions of selected XB and HB atom type pairs at the converged iteration

Atom type pair	TrainingSet-1			TrainingSet-2		
	Angle (°)	Distance (Å)	Min_potential (kcal mol^{-1})	Angle (°)	Distance (Å)	Min_potential (kcal mol^{-1})
	Geometric and energetic parameters of optimal XB interaction					
OA-Cl	175	2.65	−8.408	175	2.65	−34.578
OA-Br	180	2.85	−14.612	180	2.85	−40.533
OA-I	180	2.75	−13.701	180	2.75	−40.347
OC-Cl	180	5.15	−8.617	130	2.35	−35.830
	Geometric and energetic parameters of optimal HB interaction					
ND-OA	180	2.65	−11.732	180	3.05	−38.060
ND-OA$^{\alpha}$				180	2.35	−34.238
OD-OA	180	2.55	−14.244	180	2.35	−41.048
NC-OA	180	2.65	−13.991	180	2.45	−38.371
OD-NC	175	2.95	−11.236	165	2.35	−34.878
OC-NC	10	2.95	−11.160	170	2.45	−35.166
OC-NC$^{\beta}$	170	2.55	−10.677			

Refer to Tables S1 and S2 for explanations of the names of the atom types

$^{\alpha}$ Interaction parameters of another optimal spot for ND-OA pair based on TrainingSet-2, which is similar to the parameters of other atom type pairs (OD-OA, NC-OA, OD-NC, OC-NC) based on TrainingSet-2

$^{\beta}$ Interaction parameters of another optimal spot for OC-NC pair based on TrainingSet-1, which is similar to the parameters of other atom type pairs (ND-OA, OD-OA, NC-OA, OD-NC) based on TrainingSet-1

and R2 (threshold: -30 kcal mol^{-1}, distance ∈ [2.1 2.5] Å, angle ∈ [155° 180°]), respectively. The geometric and energetic parameters of the most preferred hydrogen bonding interactions for five atom type pairs were summarized in Table 2. Linear interaction is preferred, and the repulsive region (potential: 3 kcal mol^{-1}) gradually expands with the decrease of HBD-H HBA angle, clearly reflecting the angular preference of hydrogen bonding.

Comparatively, heat maps of NC-OA (Fig. 3m), OD-NC (Fig. 3n) and OC-NC (Fig. 3o) are different from the aforementioned two neutral atom pairs. Multiple favorable regions can be identified, for example, R3 and R4 in Fig. 3m, R5 and R6 in Fig. 3n, R7 and R8 in Fig. 3o. But the repulsive regions almost did not change with the decrease of HBD-H HBA angle, which might be caused by both hydrogen bonding and electrostatic interaction

Table 3 Mean pairwise potentials of statistical spherical bins of selected XB and HB atom type pairs at the converged iteration

Statistical region: occurrences >=500, potential <3 kcal mol^{-1}, distance <=6.0 Å

Atom type pair	TrainingSet-1			TrainingSet-2		
	Sum (kcal mol^{-1})	#Spherical bin	Mean (kcal mol^{-1})	Sum (kcal mol^{-1})	#Spherical bin	Mean (kcal mol^{-1})
	Halogen bonding					
OA-Cl	−30.36	710	−0.043	−1238.94	560	−2.212
OA-Br	−394.24	469	−0.841	−2089.82	382	−5.471
OA-I	−638.74	343	−1.862	−2294.18	319	−7.192
OC-Cl	−684.49	399	−1.716	−2604.92	351	−7.421
	Hydrogen bonding					
ND-OA	−132.61	903	−0.147	−1015.70	1006	−1.010
OD-OA	−932.41	591	−1.578	−1823.75	917	−1.989
NC-OA	−1431.37	871	−1.643	−3096.35	1171	−2.644
OD-NC	−1223.85	344	−3.558	−2991.82	847	−3.532
OC-NC	−1889.88	550	−3.436	−2995.66	995	−3.011

Refer to Tables S1 and S2 for explanations of the names of the atom types

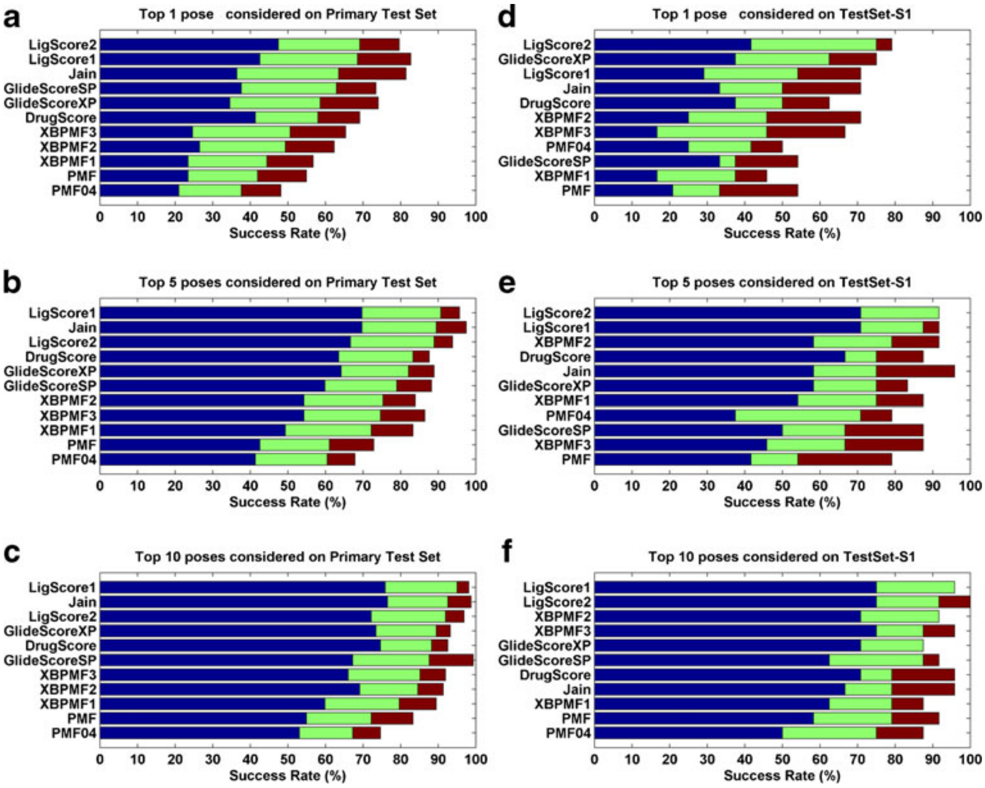


Fig. 4 Comparison of the success rates of selected scoring functions on the primary test set and TestSet-S1 considering top 1, 5 and 10 poses, when the rmsd cutoff is 1.0 Å (*blue bar*), 2.0 Å (*green bar*) or 3.0 Å (*red bar*), respectively. Scoring functions are ranked by the success rates when the rmsd cutoff is 2.0 Å. (**a**) top 1 pose considered on primary test set; (**b**) top 5 poses considered on primary test set; (**c**) top 10 poses considered on primary test set; (**d**) top 1 pose considered on TestSet-S1; (**e**) top 5 poses considered on TestSet-S1; (**f**) top 10 poses considered on TestSet-S1

since at least one atom in all three pairs is charged. R4 in Fig. 3m, R5 in Fig. 3n and R7 in Fig. 3o are hydrogen bonding regions, while R3 in Fig. 3m, R6 in Fig. 3n and R8 in Fig. 3o are caused by electrostatic interaction rather than hydrogen bonding interaction. Among Fig. 3c, d and e, the minimal potential of OC-NC (−7.08 kcal mol^{-1}) is deeper than those of NC-OA (−6.40 kcal mol^{-1}) and OD-NC (−5.54 kcal mol^{-1}), suggesting that there is a synergic effect between attractive electrostatic and hydrogen bonding interaction. The mean potentials of selected HB atom type pairs at statistical regions at the converged iteration are summarized in Table 3. General trends in the native state were basically maintained. Apparently, halogen and hydrogen bonding share similar geometric and energetic preferences, and both are specifically directional interactions.

Evaluation of scoring functions

Six widely used scoring functions were comparatively assessed with XBPMF, including the four scoring functions (LigScore [69], PMF [49], PMF04 [70], and Jain [40]) implemented in Discovery Studio software (version: 3.0), GlideScore [41, 42, 44] in Schrödinger software (version: 2010) and DrugScore [48, 52, 53]. Three scoring functions have multiple variants: LigScore (LigScore1 and LigScore2), GlideScore (GlideScore-SP and GlideScore-XP) and XBPMF (XBPMF1, XBPMF2 and XBPMF3). All these variants were evaluated in this study. In addition, different scoring functions generate scores in different units and different signs, hereby, binding scores generated by LigScore, PMF, PMF04 and Jain were reversed as negative scores for the sake of convenience.

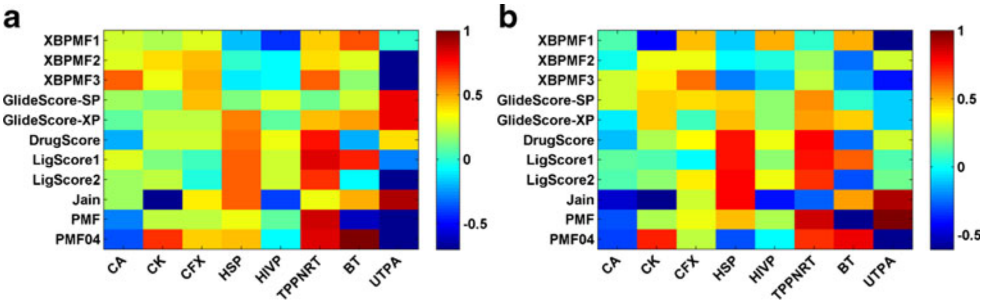


Fig. 5 Comparison of Spearman correlation coefficients of selected scoring functions based on (**a**) original protein-ligand complexes and (**b**) optimized protein-ligand complexes

Table 4 Comparison of ranking power of seven scoring functions with Spearman correlation coefficient cutoff >=0.6

Scoring function	CA	CK	CFX	HSP	HIVP	TPPNRT	BT	UTPA
XBPMF	√		√			√	√	
PMF04		√				√	√	
LigScore				√		√	√	
Jain				√				√
DrugScore				√		√		
GlideScore								√
PMF						√		

Docking power evaluation

The straightforward criterion for evaluating the docking power of a scoring function is whether the scoring function can discriminate a native binding pose from decoys by assigning the native pose a best binding score. All the decoys for each complex in the test set were scored by the selected scoring functions, and the RMSD between each decoy and the native binding pose was calculated, and then the overall success rates for all the scoring functions were calculated. Success rates of all the selected scoring functions based on two test sets are shown in Fig. 4. Based on the primary test set, LigScore, Jain, GlideScore and DrugScore outperform XBPMF, no matter that the top-1, top-5 or top-10 poses were considered (Fig. 4a-c). However, XBPMF outperforms PMF and PMF04, indicating that iterative 2D pairwise potentials were superior to 1D pairwise potentials on the primary test. Apparently, when more top-ranked binding poses are considered, success rates of all scoring functions increase considerably. In particular, when top-10 binding poses are considered, discrepancy between all selected scoring functions diminishes (Fig. 4c). When the RMSD cutoff is 2.0 Å, LigScore, Jain, GlideScore, DrugScore and XBPMF(3) achieve high success rates of over 85 %, and even to over 90 % when the RMSD cutoff is 3.0 Å, suggesting that more top-ranked binding poses should be considered in molecular docking.

When TestSet-S1 was applied (Fig. 4d-f), the ranking of all the scoring functions changed, especially for XBPMF. When top-5 or top-10 poses were considered with the RMSD cutoff of 2.0 Å, XBPMF achieves high success rates of about 80 % (Fig. 4e) and over 90 % (Fig. 4f), respectively, outranked only by LigScore. As TestSet-S1 is composed of the complexes with typical halogen bonds, the above result indicates that XBPMF is quite suitable for the systems with halogen bonding. In addition, there are some other interesting discoveries with respect to docking power: (*i*) GlideScore-XP is more appropriate than GlideScore-SP for the halogenated ligand, no matter if there exist halogen bonds or not in the protein-ligand complex of interest; (*ii*) Jain might be more appropriate in tackling the halogenated ligands with no halogen bonds rather than that with halogen bonds; (*iii*) in general, empirical scoring functions might be superior to knowledge-based scoring functions for halogenated ligands. The possible reasons might be as follows: (*i*) the limited occurrences of XB donor-acceptor pairs might result in that the pairwise potential as some specific distances or angles are not so statistically accurate, so that the scores calculated for some structures might not be so accurate; (*ii*) knowledge-based scoring functions rely on the geometries observed in the crystal structures, so that the inherent strength of some individual interactions in the crystal structures might be masked by perturbations associated with distortions to the protein or the ligand.

Ranking power evaluation

Different from discriminating native binding poses from decoy poses, the ranking power refers to the ability of correctly ranking active ligands bound to a common target according to the order of their binding affinities. As mentioned above, eight clusters of protein-ligand complexes were prepared. Each complex in the eight clusters was scored, and then a Spearman correlation analysis was implemented. Heat maps of Spearman correlation coefficients between the experimentally determined binding affinities and the binding scores computed by selected scoring functions are shown in Fig. 5. The ranking power evaluation were carried out with the Spearman correlation coefficient cutoff >=0.6 (Table 4). XBPMF outperforms other scoring functions followed by PMF04 and LigScore. For carbonic anhydrase II (CA), only XBPMF3 can generate acceptable results with Spearman correlation coefficient over 0.6 on original complexes, while no scoring functions perform well for HIV protease (HIVP). For beta-trypsin (BT), PMF04 achieves a Spearman correlation coefficient of 1.0, indicating that PMF04 correctly ranked all 16 ligands. In addition, PMF, LigScore and PMF04 perform well on tyrosine-protein phosphatase non-receptor type 1 (TPPNRT) with high Spearman correlation coefficient of over 0.8, so do Jain and GlideScore on urokinase-type plasminogen activator (UTPA).

Furthermore, moderate correlation between some scoring functions is observed in Fig. 6 (refer to Fig. S4 and Table S3 in Supporting information as well). Although different scoring functions are in different forms, they are essentially used to characterize some common interactions, such as hydrophobic contacts, hydrogen bonding, and electrostatic interactions. Hence, certain intercorrelations are in our expectations to some extent (Table S3).

Scoring power evaluation

The binding scores of the complexes in the three test sets were calculated, and the Pearson correlation coefficient between the binding scores and the binding affinities was measured as the scoring power of each scoring function, and the results were summarized in Table 5. Based on the primary test set, the results of only three scoring functions, viz. GlideScore-SP, LigScore2 and DrugScore, are acceptable with standard deviations of ~2.00 kcal mol^{-1} if we set Pearson correlation coefficient cutoff >=0.4. XBPMF performs not well on this test set though it is superior to Jain and scoring functions extracted from 1D pairwise potentials (PMF and PMF04).

However, when applying two test sets with typical halogen bonds in the complexes, XBPMF (at least one of XBPMF1, XBPMF2 and XBPMF3) produced Pearson correlation coefficient of about 0.5 on either original or optimized complexes. Especially on TestSet-S2 (halogen bond distance <=3.2 Å, halogen bond angle >=140°), XBPMF is only outranked by

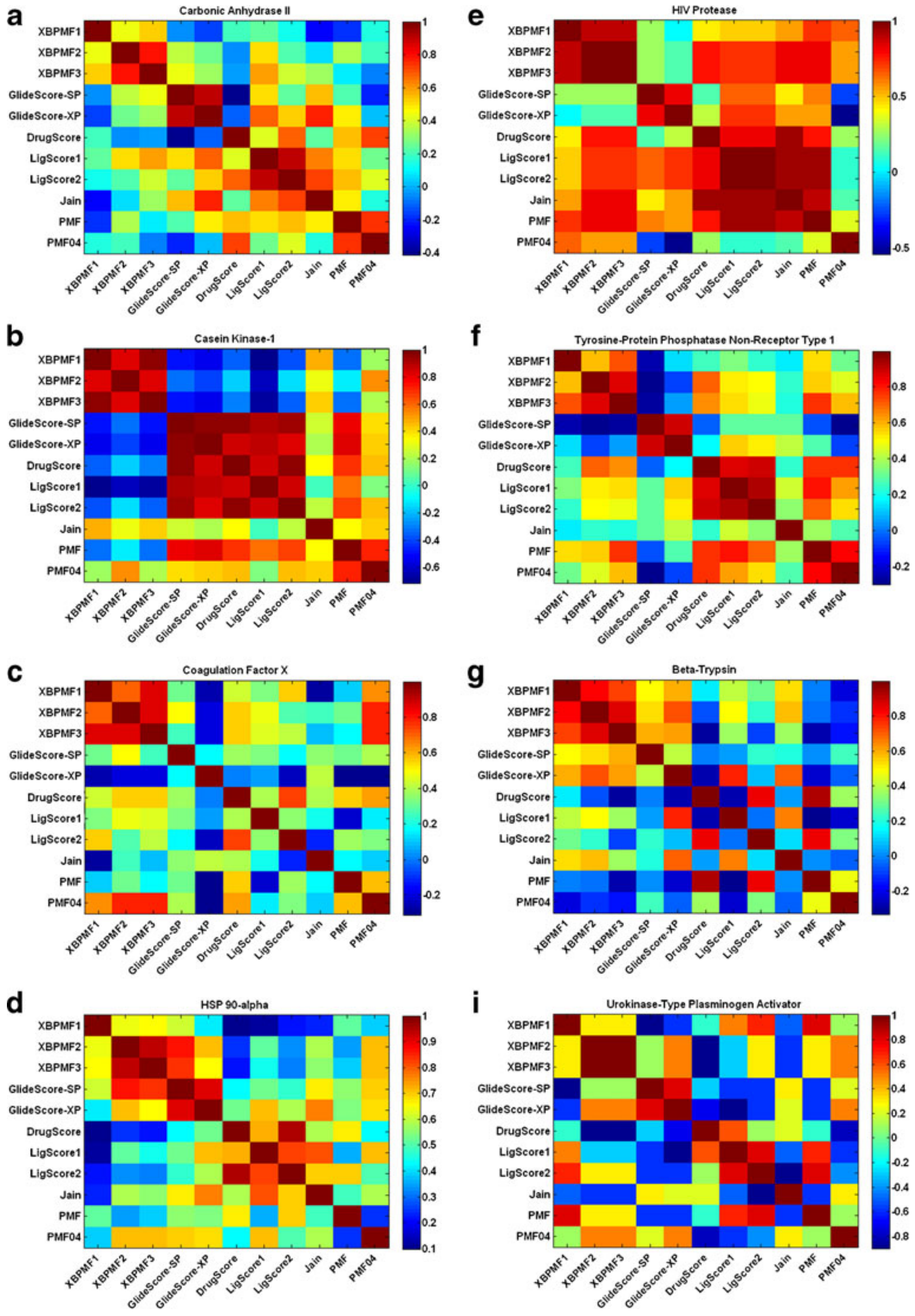


Fig. 6 Intercorrelation coefficients of selected scoring functions based on original protein-ligand complexes of eight clusters: (**a**) Carbonic anhydrase II; (**b**) Casein kinase-1; (**c**) Coagulation factor X; (**d**) HSP 90-alpha; (**e**) HIV protease; (**f**) Tyrosine-protein phosphatase non-receptor type 1; (**g**) Beta-trypsin; (**h**) Urokinase-type plasminogen activator

Table 5 Correlations between the experimentally determined binding constants and the binding scores computed by selected scoring functions on three test sets

Scoring function[a]	On original complexes				On optimized complexes			
	#entry[b]	R_p[c]	SD[d]	R_s[e]	#entry	R_p	SD	R_s
On TestSet (size: 162)								
GlideScore-SP	139	0.457	1.980	0.426	159	0.422	2.034	0.394
LigScore2	161	0.445	1.999	0.470	162	0.484	1.947	0.529
DrugScore	162	0.435	2.004	0.460	162	0.440	1.998	0.461
GlideScore-XP	152	0.259	2.176	0.231	161	0.250	2.162	0.230
LigScore1	162	0.223	2.170	0.202	162	0.236	2.163	0.211
XBPMF3	161	0.204	2.185	0.221	161	0.229	2.167	0.196
XBPMF2	161	0.203	2.163	0.201	159	0.161	2.156	0.151
XBPMF1	144	0.201	2.234	0.233	104	0.217	2.281	0.198
PMF	154	0.120	2.248	0.037	154	0.135	2.246	0.053
PMF04	138	0.013	2.282	−0.007	139	0.039	2.351	0.010
Jain	126	−0.057	2.223	0.006	127	−0.006	2.209	0.077
On TestSet-S1 (size: 24)								
GlideScore-SP	17	0.733	1.804	0.691	22	0.592	2.167	0.548
LigScore2	24	0.614	2.040	0.566	24	0.661	1.939	0.594
DrugScore	24	0.569	2.125	0.473	24	0.602	2.065	0.535
XBPMF2	24	0.431	2.333	0.424	24	0.541	2.174	0.517
LigScore1	24	0.380	2.391	0.237	24	0.362	2.410	0.217
XBPMF1	21	0.371	2.522	0.368	19	0.529	2.350	0.465
XBPMF3	24	0.342	2.429	0.288	23	0.528	2.205	0.487
GlideScore-XP	21	0.328	2.602	0.374	24	0.239	2.510	0.299
PMF	23	0.234	2.523	0.172	23	0.225	2.529	0.115
PMF04	22	0.116	2.610	0.121	21	0.054	2.667	0.039
Jain	15	0.104	2.716	0.186	15	0.259	2.663	0.232
On TestSet-S2 (size: 7)								
GlideScore-SP	6	0.525	0.968	0.657	7	0.679	0.778	0.750
XBPMF2	7	0.495	0.921	0.536	7	0.390	0.976	0.464
XBPMF3	7	0.483	0.928	0.250	7	0.469	0.936	0.357
DrugScore	7	0.427	0.958	0.679	7	0.362	0.988	0.214
LigScore2	7	0.380	0.980	0.714	7	0.482	0.928	0.714
XBPMF1	7	0.364	0.987	0.214	6	0.512	0.977	0.257
GlideScore-XP	6	0.153	1.124	0.257	7	0.137	1.050	0.393
LigScore1	7	0.092	1.055	0.464	7	0.016	1.059	0.286
PMF04	7	0.083	1.056	0.321	7	0.115	1.052	0.321
PMF	7	−0.242	1.028	−0.071	7	−0.207	1.037	−0.036
Jain	1	-	-	-	1	-	-	-

[a] Scoring functions are ranked by the Pearson correlation coefficients on the original complexes

[b] Number of complexes with negative (favorable) binding scores by the scoring function

[c] Pearson correlation coefficients

[d] Standard deviations in linear correlation (in kcal mol^{-1} units)

[e] Spearman correlation coefficients

GlideScore-SP, and the standard deviation decreased below 1.00 kcal mol^{-1}, suggesting that XBPMF is appropriate for protein-ligand complexes with halogen bonding.

Conclusions

Based on two training sets of protein-ligand complexes (size of TrainingSet-1: 1,591, size of TrainingSet-2: 31,145), an iterative multidimensional knowledge-based halogen bonding scoring function was developed, termed XBPMF, with three variants (XBPMF1, XBPMF2 and XBPMF3). The extracted 2D pairwise potentials can characterize appropriately the distance and angle preferences of halogen bonding and hydrogen bonding, which agree well with empirical observations or theoretical computational results. In addition, XBPMF was evaluated in three aspects: “docking power”, “ranking power” and “scoring power”, suggesting that the iterative 2D pairwise potentials are superior to 1D pairwise potential, and XBPMF are quite appropriate for the protein-ligand complexes, in which halogen bonds can be identified. In addition, XBPMF performs well in ranking active ligands bound to a common target according to the order of their binding affinities for four cluster families, viz. carbonic anhydrase II, coagulation factor X, tyrosine-protein phosphatase non-receptor type 1 and beta-trypsin. Although the novel halogen bonding scoring function is by no means perfect, it should help to improve our understanding of halogen bonding and provide a more accurate access to study on some systems with halogen bonds.

Acknowledgments This work was supported by the National Natural Science Foundation of China (81273435, 21021063 and 91013008) and the State Key Laboratory of Drug Research (SIMM1203KF-01). Computational resources were supported by Computer Network Information Center (CNIC), Chinese Academy of Science (CAS) and Shanghai Supercomputing Center (SCC).

References

1. Auffinger P, Hays FA, Westhof E, Ho PS (2004) Halogen bonds in biological molecules. Proc Natl Acad Sci U S A 101:16789–16794
2. Voth AR, Khuu P, Oishi K, Ho PS (2009) Halogen bonds as orthogonal molecular interactions to hydrogen bonds. Nat Chem 1:74–79
3. Voth AR, Hays FA, Ho PS (2007) Directing macromolecular conformation through halogen bonds. Proc Natl Acad Sci U S A 104:6188–6193
4. Bertani R, Sgarbossa P, Venzo A, Lelj F, Amati M et al (2010) Halogen bonding in metal-organic-supramolecular networks. Coord Chem Rev 254:677–695
5. Cavallo G, Metrangolo P, Pilati T, Resnati G, Sansotera M et al (2010) Halogen bonding: a general route in anion recognition and coordination. Chem Soc Rev 39:3772–3783
6. Li HY, Lu YX, Liu YT, Zhu X, Liu HL et al (2012) Interplay between halogen bonds and pi-pi stacking interactions: CSD search and theoretical study. Phys Chem Chem Phys 14:9948–9955
7. Li HY, Lu YX, Wu WH, Liu YT, Peng CJ et al (2013) Noncovalent interactions in halogenated ionic liquids: theoretical study and crystallographic implications. Phys Chem Chem Phys 15:4405–4414
8. Lu Y, Liu Y, Xu Z, Li H, Liu H et al (2012) Halogen bonding for rational drug design and new drug discovery. Expert Opin Drug Discov 7:375–383
9. Lu YX, Liu YT, Li HY, Zhu X, Liu HL et al (2012) Energetic effects between halogen bonds and anion-pi or lone pair-pi interactions: a theoretical study. J Phys Chem A 116:2591–2597
10. Lu YX, Liu YT, Li HY, Zhu X, Liu HL et al (2012) Mutual influence between halogen bonds and cation-p interactions: a theoretical study. Chemphyschem 13:2154–2161
11. Lu YX, Shi T, Wang Y, Yang HY, Yan XH et al (2009) Halogen bonding-a novel interaction for rational drug design? J Med Chem 52:2854–2862
12. Lu YX, Wang Y, Zhu WL (2010) Nonbonding interactions of organic halogens in biological systems: implications for drug discovery and biomolecular design. Phys Chem Chem Phys 12:4543–4551
13. Metrangolo P, Neukirch H, Pilati T, Resnati G (2005) Halogen bonding based recognition processes: a world parallel to hydrogen bonding. Acc Chem Res 38:386–395
14. Metrangolo P, Resnati G (2008) Chemistry. Halogen versus hydrogen. Science 321:918–919
15. Parisini E, Metrangolo P, Pilati T, Resnati G, Terraneo G (2011) Halogen bonding in halocarbon-protein complexes: a structural survey. Chem Soc Rev 40:2267–2278
16. Legon AC (2010) The halogen bond: an interim perspective. Phys Chem Chem Phys 12:7736–7747
17. Xu Z, Liu Z, Chen T, Wang Z, Tian G et al (2011) Utilization of halogen bond in lead optimization: a case study of rational design of potent phosphodiesterase type 5 (PDE5) inhibitors. J Med Chem 54:5607–5611
18. Murray JS, Riley KE, Politzer P, Clark T (2010) Directional weak intermolecular interactions: sigma-hole bonding. Aust J Chem 63:1598–1607
19. Murray JS, Lane P, Politzer P (2009) Expansion of the sigma-hole concept. J Mol Model 15:723–729
20. Politzer P, Murray JS, Concha MC (2007) Halogen bonding and the design of new materials: organic bromides, chlorides and perhaps even fluorides as donors. J Mol Model 13:643–650
21. Politzer P, Murray JS, Clark T (2010) Halogen bonding: an electrostatically-driven highly directional noncovalent interaction. Phys Chem Chem Phys 12:7748–7757
22. Politzer P, Lane P, Concha MC, Ma Y, Murray JS (2007) An overview of halogen bonding. J Mol Model 13:305–311
23. Bissantz C, Kuhn B, Stahl M (2010) A medicinal chemist's guide to molecular interactions. J Med Chem 53:5061–5084
24. Merino A, Bronowska AK, Jackson DB, Cahill DJ (2010) Drug profiling: knowing where it hits. Drug Discov Today 15:749–756
25. Kubota H, Avarbock MR, Brinster RL (2004) Growth factors essential for self-renewal and expansion of mouse spermatogonial stem cells. Proc Natl Acad Sci U S A 101:16489–16494
26. Hernandes MZ, Cavalcanti SM, Moreira DR, de Azevedo Junior WF, Leite AC (2010) Halogen atoms in the modern medicinal chemistry: hints for the drug design. Curr Drug Targets 11:303–314
27. Dobes P, Rezac J, Fanfrlik J, Otyepka M, Hobza P (2011) Semiempirical quantum mechanical method PM6-DH2X describes the geometry and energetics of CK2-inhibitor complexes involving halogen bonds well, while the empirical potential fails. J Phys Chem B 115:8581–8589
28. Ibrahim MA (2012) AMBER empirical potential describes the geometry and energy of noncovalent halogen interactions better than advanced semiempirical quantum mechanical method PM6-DH2X. J Phys Chem B 116:3659–3669
29. Ibrahim MAA (2012) Molecular mechanical perspective on halogen bonding. J Mol Model 18:4625–4638
30. Ibrahim MAA (2011) Molecular mechanical study of halogen bonding in drug discovery. J Comput Chem 32:2564–2574
31. Kolar M, Hobza P (2012) On extension of the current biomolecular empirical force field for the description of halogen bonds. J Chem Theory Comput 8:1325–1333
32. Ibrahim MAA (2011) Performance assessment of semiempirical molecular orbital methods in describing halogen bonding: quantum mechanical and quantum mechanical/molecular mechanical-molecular dynamics study. J Chem Inf Model 51:2549–2559
33. Jorgensen WL, Schyman P (2012) Treatment of halogen bonding in the OPLS-AA force field: application to potent anti-HIV agents. J Chem Theory Comput 8:3895–3901
34. Kolar M, Hobza P, Bronowska AK (2013) Plugging the explicit sigma-holes in molecular docking. Chem Commun 49:981–983
35. Carter M, Rappe AK, Ho PS (2012) Scalable anisotroplic shape and electrostatic models for biological bromine halogen bonds. J Chem Theory Comput 8:2461–2473
36. Lee MC, Duan Y (2004) Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized born solvent model. Proteins 55:620–634
37. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule-ligand interactions. J Mol Biol 161:269–288

38. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748
39. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE et al (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19:1639–1662
40. Jain AN (1996) Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. J Comput Aided Mol Des 10:427–440
41. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL et al (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem 47: 1750–1759
42. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ et al (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47: 1739–1749
43. Wang RX, Lai LH, Wang SM (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J Comput Aided Mol Des 16:11–26
44. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR et al (2006) Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. J Med Chem 49:6177–6196
45. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP (1997) 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 11:425–445
46. Bohm HJ (1994) 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 8:243–256
47. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Arthurs S et al (2000) Deciphering common failures in molecular docking of ligand-protein complexes. J Comput Aided Mol Des 14:731–751
48. Velec HF, Gohlke H, Klebe G (2005) 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 48:6296–6303
49. Muegge I, Martin YC (1999) A general and fast scoring function for protein-ligand interactions: a simplified potential approach. J Med Chem 42:791–804
50. Huang SY, Zou X (2006) An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. J Comput Chem 27:1866–1875
51. Huang SY, Zou X (2006) An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. J Comput Chem 27:1876–1882
52. Gohlke H, Hendlich M, Klebe G (2000) Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol 295:337–356
53. Gohlke H, Hendlich M, Klebe G (2000) Predicting binding modes, binding affinities and 'hot spots' for protein-ligand complexes using a knowledge-based scoring function. Perspect Drug Discov 20: 115–144
54. Clark T, Hennemann M, Murray JS, Politzer P (2007) Halogen bonding: the sigma-hole. Proceedings of "Modeling interactions in biomolecules II", Prague, September 5th-9th, 2005. J Mol Model 13: 291–296
55. Politzer P, Murray JS, Clark T (2013) Halogen bonding and other sigma-hole interactions: a perspective. Phys Chem Chem Phys 15: 11178–11189
56. Rendine S, Pieraccini S, Forni A, Sironi M (2011) Halogen bonding in ligand-receptor systems in the framework of classical force fields. Phys Chem Chem Phys 13:19508–19516
57. Zheng M, Xiong B, Luo C, Li S, Liu X et al (2011) 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 51:2994–3004
58. Thomas PD, Dill KA (1996) Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 257: 457–469
59. Thomas PD, Dill KA (1996) An iterative method for extracting energy-like quantities from protein structures. Proc Natl Acad Sci U S A 93:11628–11633
60. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN et al (2000) The protein data bank. Nucleic Acids Res 28:235–242
61. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T et al (2011) Open babel: an open chemical toolbox. J Cheminform 3:33
62. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK et al (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem 30:2785–2791
63. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C et al (1984) A new force-field for molecular mechanical simulation of nucleic-acids and proteins. J Am Chem Soc 106:765–784
64. Wang RX, Fang XL, Lu YP, Wang SM (2004) The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J Med Chem 47:2977–2980
65. Wang RX, Fang XL, Lu YP, Yang CY, Wang SM (2005) The PDBbind database: methodologies and updates. J Med Chem 48: 4111–4119
66. Ewing TJ, Makino S, Skillman AG, Kuntz ID (2001) DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des 15:411–428
67. Lang PT, Brozell SR, Mukherjee S, Pettersen EF, Meng EC et al (2009) DOCK 6: combining techniques to model RNA-small molecule complexes. Rna 15:1219–1230
68. Cheng TJ, Li X, Li Y, Liu ZH, Wang RX (2009) Comparative assessment of scoring functions on a diverse test set. J Chem Inf Model 49:1079–1093
69. Krammer A, Kirchhoff PD, Jiang X, Venkatachalam CM, Waldman M (2005) LigScore: a novel scoring function for predicting binding affinities. J Mol Graph Model 23:395–407
70. Muegge IA (2006) PMF scoring revisited. J Med Chem 49:5895–5902