Computational and Structural Biotechnology Journal 19 (2021) 6291–6300

ELSEVIER

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Review

Structure-based protein–ligand interaction fingerprints for binding affinity prediction

Debby D. Wang [a],*, Moon-Tong Chan [b], Hong Yan [c]

[a] School of Health Science and Engineering, University of Shanghai for Science and Technology, 516 Jungong Rd, Shanghai 200093, China
[b] School of Science and Technology, Hong Kong Metropolitan University, 30 Good Shepherd St, Ho Man Tin, Hong Kong
[c] Department of Electrical Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 30 August 2021
Received in revised form 9 November 2021
Accepted 13 November 2021
Available online 25 November 2021

Keywords:
Interaction fingerprint
Protein–ligand binding affinity
Scoring function
Machine learning
Computer-aided drug design

ABSTRACT

Binding affinity prediction (BAP) using protein–ligand complex structures is crucial to computer-aided drug design, but remains a challenging problem. To achieve efficient and accurate BAP, machine-learning scoring functions (SFs) based on a wide range of descriptors have been developed. Among those descriptors, protein–ligand interaction fingerprints (IFPs) are competitive due to their simple representations, elaborate profiles of key interactions and easy collaborations with machine-learning algorithms. In this paper, we have adopted a building-block-based taxonomy to review a broad range of IFP models, and compared representative IFP-based SFs in target-specific and generic scoring tasks. Atom-pair-counts-based and substructure-based IFPs show great potential in these tasks.

1. Introduction

Attributed to the advances in high-resolution structure determination [1] and computational methodologies for structural analysis and molecular design [2,3], structure-based drug design (SBDD) has developed into a robust and promising technique for drug discovery [4,5]. As an important participant in SBDD, molecular docking has proven to be a valuable tool for identifying novel hit compounds from a large chemical library for a particular target [6]. Binding affinity prediction (BAP) for putative protein–ligand complexes is essential in this process, and is generally achieved through scoring functions (SFs) [7–11]. An SF can be evaluated according to its performance in multiple tasks, including docking (identifying near-native binding modes), screening (distinguishing active binders from decoys), ranking (correctly ranking the binding affinities of the ligands for a given target) and scoring (achieving a linear correlation between the predicted binding scores and experimental binding data). Different applications of SFs emphasize one or more of these tasks, such as virtual screening (docking and screening) and lead optimization (ranking and scoring). Classical SFs generally prioritize the rapid screening speed over accurate prediction of binding affinity, and thus hardly perform well in scoring and ranking tasks [12–14]. Improving the scoring and ranking powers is a requisite to the development of SFs, but remains to be a challenge in SBDD.

In recent years, the extensive use of machine learning has been refocused from quantitative structure–activity relationship (QSAR) studies [15] onto structure-based predictive modeling [16–18]. The increasingly available structural and binding-affinity data of protein–ligand complexes, which allow the training of BAP models, have led to a surge in machine-learning SFs [19]. These SFs can handle a large volume of structural data, and have been demonstrated to outperform classical SFs in scoring works [20,19]. Feature engineering is crucial to the construction of a machine-learning SF. This process translates a complex structure into a series of descriptors, and is often guided by the knowledge of biologically relevant interactions such as hydrogen bonds, hydrophobic contacts, ionic interactions (salt bridges), π-stacking and π-cation interactions [21]. Recently, interaction fingerprints (IFPs) have become a study focus of SF descriptors, due to the simple representations and elaborate profiles of key interactions. Given a protein–ligand complex structure (Fig. 1A), IFPs are generally defined based on protein–ligand interacting atoms (Fig. 1A), and stored as 1-dimensional vectors or matrices of Booleans, integers or floating-point numbers. In earlier works, structural IFPs have only been classified roughly, such as keyed fingerprints vs. feature collections [22] and ligand descriptors vs. protein descriptors [20]. Moreover, the heterogeneity in benchmark data, preprocessing procedures, implementations and evaluation metrics

* Corresponding author.
E-mail address: d.wang@usst.edu.cn (D.D. Wang).

https://doi.org/10.1016/j.csbj.2021.11.018
2001-0370/© 2021 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

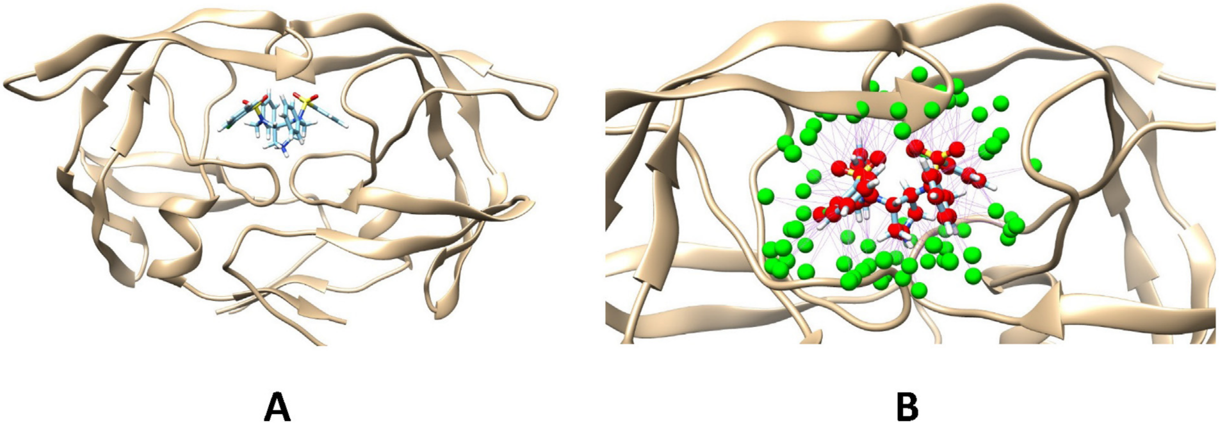

Fig. 1. Example of a protein–ligand complex and the interacting atoms. (A) Complex of HIV-1 protease and its inhibitor (PDB ID:2QNQ). (B) Protein–ligand interacting atoms defined by a distance threshold ($4.5\mathring{A}$).

results in the incomparability among IFP-based scoring works. Accordingly, in this paper we have reviewed a broad range of IFP models according to a building-block-based taxonomy, and compared the SFs incorporating representative IFP models in several scoring tasks (target-specific or generic). SFs based on other descriptors or IFP applications in tasks other than scoring (such as screening) are out of scope of this review and can be found in previous reviews [23,24].

2. Protein–ligand IFPs

Building blocks vary from one type of IFPs to another. They primarily include structural elements (protein residues, protein atoms and molecular substructures) and intermolecular interactions (atom pairs and interaction pairs/triplets). Accordingly, we classify a variety of IFP models based on their building blocks, and review them as follows.

2.1. IFPs based on protein residues

Structural interaction fingerprint (SIFt) was a pioneer study in representing and analyzing 3D protein–ligand binding interactions [25]. Each SIFt is a simple 1D binary string generated by identifying a common panel of binding-site residues, using seven bits to represent each residue in the panel, and concatenating the bit strings of all residues (Fig. 2A). The bit string for each residue covers different types of interactions, including (1) contact with the ligand, (2) binding with any protein main-chain atom, (3) binding with any protein side-chain atom, (4) polar interaction, (5) nonpolar interaction, (6) hydrogen bond with acceptor in protein, and (7) hydrogen bond with donor in protein. Accordingly, a SIFt can be denoted as

$$S = (S^1_{R_1}, \ldots, S^7_{R_1}, S^1_{R_2}, \ldots, S^7_{R_2}, \ldots, S^1_{R_n}, \ldots, S^7_{R_n}) \tag{1}$$

where $S^j_{R_i}$ is 0 or 1, and denotes whether an interaction of type j exists between protein residue R_i and the ligand. Based on SIFt, a number of extensions have been developed. PyPLIF adopts an analogous list of protein–ligand interactions (apolar, aromatic, hydrogen bond and electrostatic) to generate IFPs [26]. In r-SIFt, each bit indicates whether a specific R group or core fragment of the ligand interacts with a specific protein residue [27]. Three additional interactions (hydrophobic, aromatic and charged) to those of SIFt were considered in [28]. Marcou et al. defined 11 types of protein–ligand interactions based on a list of atom flags and geometric criteria, including hydrophobic, aromatic (face-to-face), aromatic (edge-to-face), hydrogen bond (acceptor in ligand), hydrogen bond (donor in ligand), ionic bond with ligand negatively charged, ionic bond with ligand positively charged, weak hydrogen bond (acceptor in ligand), weak hydrogen bond (donor in ligand), π-cation and metal complexation [29]. By default, their IFP model only considers the first seven interactions (most frequent) for each protein residue, but the remaining interactions (weak hydrogen bonds, π-cation interactions and metal complexation) can be easily incorporated. Weighted SIFt (w-SIFt) introduces a weight to each interaction bit to capture the relative importance of each bit for binding, with the weights determined by stochastic optimization techniques or as simple averages at each bit position [30]. A w-SIFt model has been proposed in [31], which assigns the weights for electrostatic interactions (positive charge, negative charge and metal-binding interactions) as 2 and the weights for the rest (hydrogen-bond donor/acceptor, $\pi-\pi$ stacking and hydrophobic interactions) as 1. Wojcikowski et al. filled each IFP position with continuous features, ranging from van der Waals potential (Opls2005 force field), hydrogen bonds and halogen bonds (donor-hydrogen distance and donor/acceptor counts), salt bridges, π-interactions and π-cation interactions between a protein residue and the ligand [32].

Aside from the aforementioned models, another type of residue-based IFPs employ molecular interaction energy components (MIECs) for each position. By combining the molecular dynamics (MD) simulation techniques and MM-GB/SA free energy decomposition approach, Sun et al. developed an MIEC-based IFP [33]. *Autodock* was used in their work to produce the initial protein–ligand complex structures, which were further optimized through explicit-solvent MD simulations (three stages, total 5 picoseconds). Based on each optimized complex structure, the MM-GB/SA approach was employed to calculate residue-ligand binding free energy and its components (MIECs, Eq. 2).

$$\Delta G^{residue-ligand}_{bind} = \Delta G^{vdW}_{bind} + \Delta G^{ele}_{bind} + \Delta G^{GB}_{bind} + \Delta G^{SA}_{bind} \tag{2}$$

where ΔG^{vdW}_{bind}, ΔG^{ele}_{bind}, ΔG^{GB}_{bind} and ΔG^{SA}_{bind} denote van der Waals interaction, electrostatic interaction, and the polar/non-polar parts of solvation free energy, respectively. Ligand-binding residues were selected according to the ranking of average $\Delta G^{residue-ligand}_{bind}$, and different MIECs of these residues constitute the IFPs, such as

$$S^{MIEC} = (\Delta G^{vdW}_{R_1}, \Delta G^{ele}_{R_1}, \Delta G^{GB}_{R_1}, \Delta G^{SA}_{R_1}, \ldots, \Delta G^{vdW}_{R_n}, \Delta G^{ele}_{R_n}, \Delta G^{GB}_{R_n}, \Delta G^{SA}_{R_n}) \tag{3}$$

Later, the same group proposed a similar IFP model, which combines *Glide* docking, implicit-solvent MD simulations, MM-GB (PB)/SA approach for energy decomposition and threshold-based identification of ligand-binding residues [34]. Protein–ligand Empirical Interaction Components (PLEIC) method [35] identifies

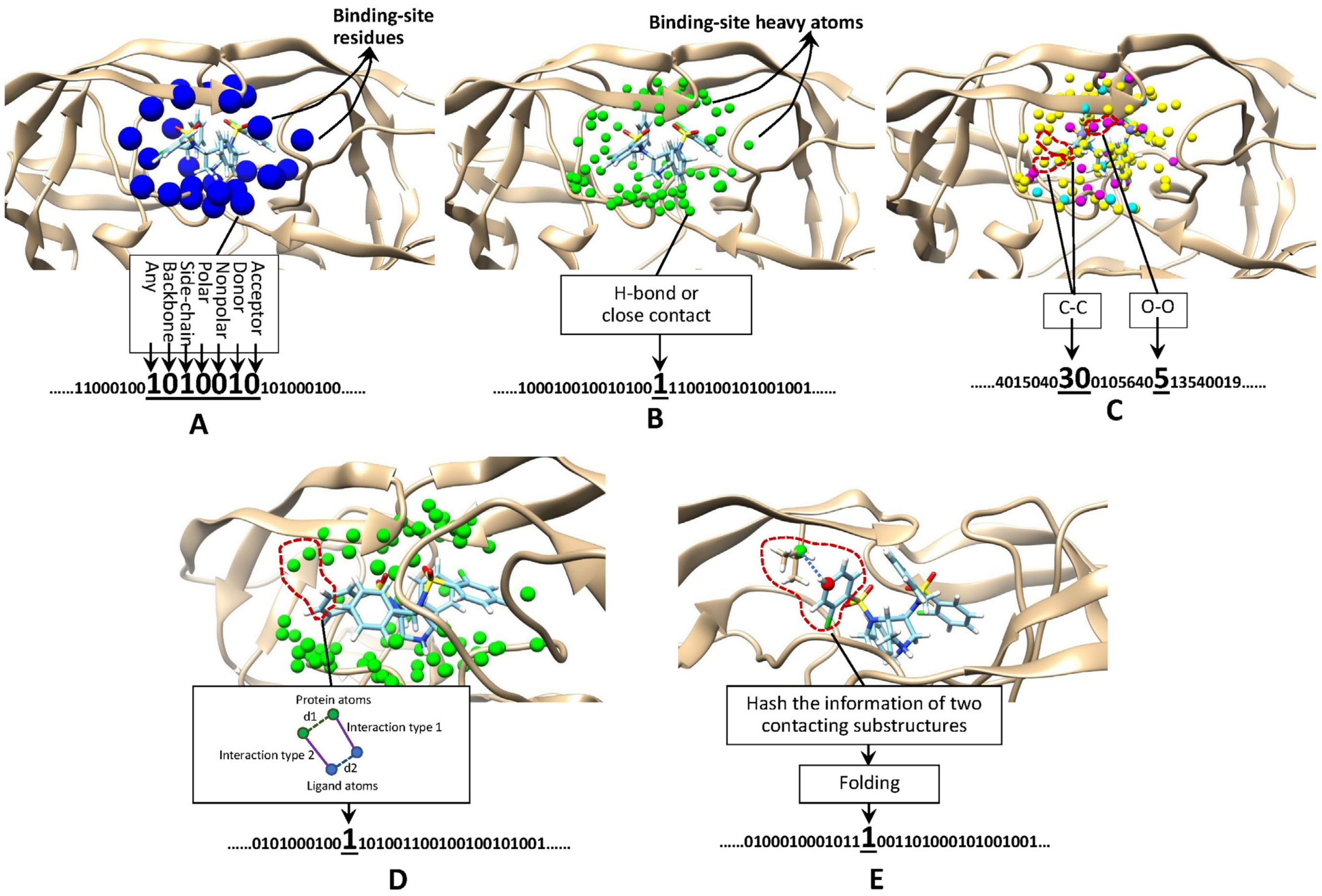

Fig. 2. Interpretations of the construction processes of representative IFPs. (A) SIFt. (B) CHIF. (C) Atom-pair counts used by RF-Score. (D) APIF. (E) SPLIF.

a panel of residues according to three types of interactions with the ligand (van der Waals interaction, hydrophobic contact and hydrogen bond), and empirically calculated such MIECs as follows.

$$\begin{cases} VDW_i = \sum_{j}^{ligand}\sum_{k}^{res}[(\frac{d_0}{d_{jk}})^8 - 2\times(\frac{d_0}{d_{jk}})^4] \\ HC_i = \sum_{j}^{ligand}\sum_{k}^{res} f(d_{jk}) \qquad \text{with} \quad f(d) = \begin{cases} 1.0, d \leqslant d_0 + 0.5\mathring{A} \\ \frac{d_0+2.0-d}{1.5}, d_0 + 0.5\mathring{A} < d \leqslant d_0 + 2.0\mathring{A} \\ 0, d > d_0 + 2.0\mathring{A} \end{cases} \\ HB_i = \sum_{j}^{ligand}\sum_{k}^{res}(\frac{1}{1+(d_{jk}/2.6)^6}/0.58) \end{cases} \tag{4}$$

where VDW_i, HC_i and HB_i represent the three types of interaction energies between residue i and the ligand, d_{jk} is the distance between atom j in the ligand and atom k in residue i, d_0 denotes the sum of atomic radii of j and k, and only hydrogen-bond donors and acceptors in the ligand and residue i are considered when calculating HB_i. Similarly, *Glide* was adopted for docking each ligand to the target and estimating the MIECs between the ligand and protein residues (within $12\mathring{A}$ from the ligand center) in Yasuo et al.'s work [36], where van der Waals interactions, electrostatic interactions and hydrogen bonds were considered as MIECs. Ji et al. adopted different protocols of MD simulations for pre-processing each *Glide*-docked complex structure, and constructed IFPs (Eq. 3) based on MM-GB/SA free-energy-decomposition calculations for the key residues ($\Delta G_{bind}^{residue-ligand} < -0.1 kcal/mol$) [37].

2.2. IFPs based on protein atoms

Analogous to SIFt, knowledge-based IFP (KB-IFP) was proposed as a similarity-search tool for reference-based scoring [38]. It is a bit string as long as the number of binding-site heavy atoms, and generated based on the interactions detected by pairwise interatomic parameters (distance d and Hydrogen-bond angle a). Interactions of hydrogen bonds ($d \leqslant 3\mathring{A}$ and $a < 90^\circ$) and close contacts ($d <$ sum of vdW radii) are identified first, and a bit in KB-IFP is set to 1 if the corresponding heavy atom (A_i) forms an interaction with an atom in the ligand (Eq. 5).

$$S = (S_{A_1}, S_{A_2}, \ldots, S_{A_n}) \tag{5}$$

These IFPs are named as HIF (hydrogen bond-based), CIF (close contact-based) or CHIF (considering both types of interactions, Fig. 2B). A similar atom-based IFP model considers the properties of hydrogen bonds within a binding site (strength, accessibility of hydrogen-bond groups and geometric arrangement) [39]. In addition, some atom-based IFPs consider the energy terms related to each binding-site atom. A protein per atom score contribution derived interaction fingerprint (PADIF) for protein–ligand complexes [40] characterizes the protein binding-site atoms using the per atom contributions of the GOLD Score, which involves a hydrogen-bonding term and multiple linear potentials to model van der Waals and repulsive terms [41].

2.3. IFPs based on atom-pair counts

Atom-pair counts (intermolecular interaction features) were used to construct RF-Score, which is arguably the first machine-learning SF [42]. Among protein–ligand atom pairs that interact within a certain distance range ($< 12\AA$), the counts of specific pairs, classified by the atom types, are the main SF-descriptors. Specifically, atom types of $\{C, N, O, S\}$ for protein atoms and $\{C, N, O, F, P, S, Cl, Br, I\}$ for ligand atoms were considered, constituting 36 descriptors (e.g. $C - C, C - N, \ldots, S - I$). This list of descriptors can also be regarded as an integer IFP (Fig. 2B) and expressed as

$$S = (S_1, S_2, \ldots, S_{36}) \tag{6}$$

where S_i represents the number of interactions of type i (e.g. $C - C, C - N, \ldots, S - I$). These simple descriptors can be easily calculated and often lead to fast scoring works.

CScore was proposed later [43], to further subdivide the interactions into repulsive and attractive types by introducing two fuzzy membership functions. Atom type of H was additionally considered in this work, yielding the following integer IFP,

$$S = (S_1^{repulsive}, S_1^{attractive}, S_2^{repulsive}, S_2^{attractive}, \ldots, S_{50}^{repulsive}, S_{50}^{attractive}) \tag{7}$$

OnionNet is a recent SF that employs atom-pair counts as descriptors and uses deep-learning models for affinity prediction [44]. This method starts from each ligand atom, and defines a series of distance bins for the atom ($bin_1 \sim bin_N$: $(0, d_0), [d_0, d_0 + \delta), [d_0 + \delta, d_0 + 2\delta), \ldots, [d_0 + (N-2)\delta, d_0 + (N-1)\delta)$). Considering eight element types ($\{C, N, O, H, P, S, HAX, DU\}$, HAX: halogen atoms $\{F, Cl, Br, I\}$, DU: remaining types) for protein and ligand atoms, atom-pair counts within the N distance bins form a 2D integer IFP as

$$S = \begin{bmatrix} S_1^{bin_1} & \ldots & S_1^{bin_N} \\ S_2^{bin_1} & \ldots & S_2^{bin_N} \\ \vdots & \vdots & \vdots \\ S_{64}^{bin_1} & \ldots & S_{64}^{bin_N} \end{bmatrix} \tag{8}$$

where $S_i^{bin_j}$ counts the atom pairs of type i (e.g. $C - C, C - N, \ldots, DU - DU$) and within bin_j. Such IFPs (Onionnet setting: $N = 60, d_0 = 1\AA, \delta = 0.5\AA$) were extracted and fed into deep convolutional neural networks (CNNs) for affinity prediction. Recently, the same group has proposed the OnionNet-2 model, which modifies the original atom pairs into pairs of a protein residue and a ligand atom [45]. Specifically, 21 residue types (20 standard types and an expanded type) were considered, and the residue-atom distance was calculated as the distance between the ligand atom and the nearest heavy atom in the residue.

As another extension of RF-Score, extended connectivity interaction features (ECIFs) model employs six atomic properties (atom type, explicit valence, connections to heavy atoms, connections to hydrogens, aromaticity and ring membership) to define the types of atom pairs [46]. For example, ‘$O;2;1;0;0;0$’indicates an oxygen atom with an explicit valence of 2, connected to 1 heavy atom, and having no aromaticity nor a ring membership. Accordingly, ‘$O;2;1;0;0;0$’-‘$N;3;2;1;0;0$’represents a specific type of atom pairs. These descriptors lead to an integer IFP with a length of 1540. In this study (GBDT-based), a series of distance cutoffs ($4.0\AA$ to $15.0\AA$ with an interval of $0.5\AA$) were used to generated ECIFs, and the cutoff of $6\AA$ was reported to result in the best predictions.

2.4. IFPs based on pairs or triplets of interactions

Atom-pair-based interaction fingerprints (APIF) consider interactions between pairs of protein and pairs of ligand atoms[47]. An APIF is generated in four steps: identification of the active site (threshold of $10\AA$), detection of protein–ligand interactions (hydrogen bonds and hydrophobic contacts), classification of pairwise interactions and construction of the final IFP (Fig. 2D). Six types of pairwise protein–ligand interactions can be detected [47]. For each pairwise interaction, the distance between the two protein atoms (d_1) and that between the two ligand atoms (d_2) are each mapped into 7 distance bins ($bin_1 \sim bin_7 : 0 - 2.5\AA; 2.5 - 4\AA; 4 - 6\AA; 6 - 9\AA; 9 - 13\AA; 13 - 18\AA$ and $> 18\AA$). The final APIF is a string of $6 \times 7 \times 7 = 294$ bits as expressed in Eq. 9,

$$S = (S_1^{bin_1 - bin_1}, S_1^{bin_1 - bin_2}, \ldots, S_1^{bin_7 - bin_7}, \ldots, S_6^{bin_1 - bin_1}, S_6^{bin_1 - bin_2}, \ldots, S_6^{bin_7 - bin_7}) \tag{9}$$

where $S_i^{bin_j - bin_k}$ indicate the occurrence of type-i pairwise interactions with d_1 in bin_j and d_2 in bin_k.

Pharmacophore-based interaction fingerprints (Pharm-IF) are similar to APIF, but characterize each pairwise interaction using the distance and pharmacophore features of the two ligand atoms [48]. To generate Pharm-IF, protein–ligand interactions (hydrogen bond, ionic and hydrophobic) are identified. Pairwise interactions are formed by combining these individual interactions, such as *hydrophobic – hydrophobic*. Then the Pharm-IF of a protein–ligand complex is constructed as the matrix,

$$H_{tp,br} = \sum_{p \in I_{tp}} A_{br}(p) \tag{10}$$

where tp is the type of pairwise interaction, p is a pairwise interaction of tp, br indicates the boundaries of distance bins for the ligand atoms in p (e.g. $1\AA, 2\AA, 3\AA, \ldots$), and $A_{br}(p)$ is defined as:

$$A_{br}(p) = \begin{cases} 0, & if\ |br - d_p| \geqslant 1 \\ 1 - |br - d_p|, & otherwise \end{cases} \tag{11}$$

where d_p is the distance between the two ligand atoms in p.

Triplets of interactions to construct IFPs (TIFPs) [20] were identified based on specific pharmacophoric properties (hydrophobic, aromatic, hydrogen-bond donor/acceptor, ionic and metal) and geometric criteria, with each interaction represented by an interaction pseudoatom (IPA). Then triplets of IPAs are characterized by the pharmacophoric properties of the IPAs and their related distances (binned into $0 - 4\AA; 4 - 6\AA; 6 - 8\AA; 9 - 13\AA; 13 - 17\AA; > 17\AA$). These triplets can be pruned into 210 integers (a TIFP string) by removing redundancy and validating the geometry, with each integer registering the count of IPA triplets occurring at the binned distances.

2.5. IFPs based on molecular substructures

Molecular substructures or fragments are frequently used to cluster compounds and for ligand-based virtual screening [49,50]. Structural fingerprints like the extended connectivity fingerprints (ECFPs) [51] are representatives who employ these descriptors. The ECFP of a molecule is a folded string of integer identifiers, which are iteratively assigned to circular atom environments (substructures) up to a pre-defined bond radius (R). Initially, each heavy atom ($r = 0$) is assigned with an identifier based on a group of properties (e.g. mass, charge or connections), and such identifiers are iteratively updated to cover larger atom environments ($r = 1, \ldots, R$). For a heavy atom $\mathbf{a}$ in a molecule, suppose $\{\mathbf{n}_i | i = 1, \ldots, n\}$ are its neighboring atoms, $\{b_i | i = 1, \ldots, n\}$ are the bonds connecting **a** and its neighbors, and $idf_{\mathbf{x}}^{r-1}$ is the identifier of the environment centered at x and with a radius of $r-1$. Then the identifier of the atom environment with center **a** and radius r is derived by combining $idf_{\mathbf{a}}^{r-1}, \{b_i | i = 1, \ldots, n\}$ and $idf_{\mathbf{n}_i}^{r-1}$ through a hash procedure. Most substructure-based IFPs are based on ECFPs, and can be classified as ligand-centric or interaction-centric.

As an early ligand-centric, substructure-based IFP, the Interaction Annotated Structural Features (IASF) method [22] calculates atom-based energy scores based on the FlexX scoring function, which considers neutral hydrogen bonds, salt bridges, aromatic interactions, lipophilic contacts and van der Waals contacts [52]. The substructures (circular atom environments) at the binding-site region are generated by the ECFP$_4$ ($R = 2$) algorithm and annotated with cumulative atom-based energy scores. The scores of the substructures (U_i) can be organized as an IFP (Eq. 12) or accumulated as the full score for the host-ligand interaction.

$$S = (S_{U_1}, S_{U_2}, \ldots, S_{U_n}) \tag{12}$$

Different types of molecular fragments, including interacting fragments (IFs), random fragments (RFs) and interaction-compromised fragments (ICFs), have been defined for protein–ligand complexes [53] and can be mapped to fingerprints according to MACCS structural keys [54]. Here the IFs include ligand atoms involved in hydrogen bonds, ionic interactions or van der Waals interactions with the protein [53]. As an extension of IFs, atom-centered interaction fragments (AIFs) have been used to construct IFPs for similarity searches [55]. The AIF of each atom in an IF can be formed by combining it with its direct neighbors based on a pre-defined bond radius (2, 4 or 6 in [55]), analogous to ECFP substructures. AIFs form a library of unique descriptors, which can be used to encode a binary or integer fingerprint.

Differing from IFPs that are based on a specific list of interaction types, interaction-centric, substructure-based IFPs implicitly account for all types of local interactions and do not require pre-defined geometric criteria to identify interactions [56]. Structural protein–ligand interaction fingerprints (SPLIF), a pioneer study, explicitly encode interacting substructures of a protein–ligand complex [57]. A SPLIF can be generated by identifying interacting atoms (distance within 4.5Å), extracting circular substructures and folding the information of pairwise substructures (Fig. 2E). These substructures are centered on interacting atoms and detected by the ECFP$_2$ ($R = 1$) algorithm. Each position in a SPLIF (S_i) indicates the existence or occurrence of a specific pair of substructures. Protein–ligand extended connectivity fingerprints (PLEC FP) consider protein–ligand contact substructures with mul-

Table 1
An overview of different IFP models and the related scoring tasks.

Category	IFP	Ref	Format			Target-specific scoring			Generic scoring (evaluated on *PDBbind* Core Set)		
			Binary	Integer	Floating number	Target	Machine learning algorithm [a]	Evaluation [b]	Version	Machine learning algorithm[a]	Evaluation [b]
Residue-based	SIFt, r-SIFt	[25,29,28,26,27]	√			–	–	–	–	–	–
	w-SIFt	[30,31]			√	p38α [30]	–	$R = 0.604$ [30]	–	–	–
	Continuous IFP	[32]			√	HIV-1 protease	GBDT	$R = 0.77 + 0.007/$ $RMSE = 1.48 + 0.02$	–	–	–
	MIEC-IFP	[33–37]			√	5HT2AR, CB1, M1R, VEGFR2, ERK2, A2AR [37]	multiple	$\bar{R}^2 = 0.53/$ $\bar{RMSE} = 1.40$ [37]	–	–	–
Atom-based	KB-IFP	[38]	√			–	–	–	–	–	–
	PADIF	[40]			√	–	–	–	–	–	–
Atom-pair-counts -based[c]	APC (RF-Score)	[42]		√		–	–	–	v2007	RF	$R = 0.776/$ $SD = 1.58$
	EAPC (CScore)	[43]		√		subset of *PDBbind* (v2009)	NN	$R = 0.8237/$ $RMSE = 1.0872$	v2009	NN	$R = 0.7768/$ $RMSE = 1.4540$
	APCiDB (OnionNet)	[44]		√		–	–	–	v2016 v2013	CNN	$R = 0.816/$ $RMSE = 1.278$ $R = 0.78/$ $SD = 1.45$
	RAPCiDB (OnionNet-2)	[45]		√		–	–	–	v2016 v2013	CNN	$R = 0.864/$ $RMSE = 1.164$ $R = 0.821/$ $RMSE = 1.357$
	ECIF	[46]		√		–	–	–	V2016	GBDT	$R = 0.857/$ $RMSE = 1.193$
Multi-interaction -based	APIF	[47]		√		–	–	–	–	–	–
	Pharm-IF	[48]		√		–	–	–	–	–	–
	TIFP	[20]			√	–	–	–	–	–	–
Substructure -based	IASF	[22]			√	–	–	–	–	–	–
	SPLIF	[57]	√	√		–	–	–	–	–	–
	PLEC FP	[58]	√	√		–	–	–	v2016 v2013	NN	$R = 0.82$ $R = 0.77$
	PrtCmm IFP	[59]	√	√		–	–	–	v2019	RF	$R = 0.793/$ $RMSE = 2.014$

[a] GBDT: gradient boosting decision tree, RF: random forest, NN: neural network, CNN: convolutional neural network.
[b] R: Pearson's correlation between predicted and experimental affinities, RMSE: root-mean-square error, SD: standard deviation.
[b] APC: atom-pair counts, EAPC: evolved APC, APCiDB: atom-pair counts in distance bins, RAPCiDB: residue-atom-pair counts in distance bins.

tiple pairs of radii or depths (e.g. $R_{protein} = 5$ and $R_{ligand} = 1$), and hash these substructure pairs to specific fingerprint positions [58]. Proteo-chemometric IFPs (PrtCmm IFPs) [59] can be generated by separately encoding interacting substructures (produced by the ECFP algorithm) of the protein and ligand, and concatenating the two fingerprints.

3. Comparison of IFP Scores

Machine-learning SFs that absorb IFPs as descriptors, IFP Scores for short, play an important role in BAP problems. A scoring task in BAP can be either target-specific (multiple ligands for the same target) or generic (multiple targets). An overview of the aforementioned IFPs and the related IFP Scores is presented in Table 1. Here the IFPs for RF-Score, CScore, OnionNet and OnionNet-2 are named as atom-pair counts (APC), evolved atom-pair counts (EAPC), atom-pair counts in distance bins (APCiDB) and residue-atom-pair counts in distance bins (RAPCiDB), respectively. The reported performances of the IFP Scores, mostly on well-acknowledged benchmarks (*PDBbind* Core Sets), are also listed. Unfortunately, these scoring works are hardly comparable, due to the disagreement in training/benchmark data (different databases or different versions of the same database), data pre-processing procedures, machine-learning models, implementation details and evaluation metrics. Meanwhile, only few toolkits have been released to offer the use of IFPs (Table 2), the majority of which are the extensions of SIFt. This makes the comparison among IFPs in BAP more difficult. In this work, we used Python to re-program representative IFP models and attempted to provide a comparison among the related SFs, using a uniform setting. Specifically, we use the raw data (no further processing) in *PDBbind* v2019 [60] for training and validating SFs. The *PDBbind* database, which covers the structural and affinity data of a variety of protein–ligand complexes (around 18,000), has been extensively used in BAP works. The developers have further filtered these complexes by multistep quality control (Refined Set and Core Set), which kept the ligands of interest and maximumly guaranteed the diversity of ligands for each target. Several classic machine-learning models (RFs, GBDTs and regression trees) were adopted in SF construction, and Pearson's correlation coefficient ($corr = \frac{\sum_{i=1}^{n}(y_i^{pred}-\bar{y}^{pred})(y_i^{exp}-\bar{y}^{exp})}{\sqrt{\sum_{i=1}^{n}(y_i^{pred}-\bar{y}^{pred})^2}\sqrt{\sum_{i=1}^{n}(y_i^{exp}-\bar{y}^{exp})^2}}$) and root-mean-square error ($RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i^{pred}-y_i^{exp})^2}{n}}$) between the predicted and experimental binding affinities were used as the evaluation metrics.

3.1. Target-specific scoring

Several frequently studied targets (Table 3) were selected from *PDBbind* for target-specific scoring (details of data sets in Supplementary file). IFP Scores were constructed through the consociation between IFP models and machine-learning algorithms.

Table 2
Several toolkits offering the use of IFPs.

Toolkit	Online address	IFP type	Ref	BAP works
IChem	http://bioinfo-pharma.u-strasbg.fr/labwebsite/download.html	Residue-based	[61]	-
OEChem	https://www.eyesopen.com/oechem-tk	Residue-based	[29]	-
PyPLIF	http://code.google.com/p/pyplif	Residue-based	[26]	-
MOE	https://www.chemcomp.com/Products.htm	Residue-based	[62]	-
ODDT	https://github.com/oddt/oddt	Substructure-based (PLEC FP)	[63]	[58]

1. Representative IFPs from the five classes (Section 2) were investigated. Each type of IFPs was generated for protein–ligand complexes in each target-specific scoring task. When it comes to generating IFPs, interacting atoms are first detected by a distance threshold (t_{int}), and different thresholds were used in the original IFP works (Table 4). Then pharmacophoric properties of these interacting atoms and specific geometric criteria were used to identify key interactions for IFP construction. Residue-based and atom-based IFPs were generated as binary strings, while the others as integer vectors.
2. The above IFPs were fed into three classic machine-learning algorithms (RFs [64], GBDTs [65] and regression trees [66]) to build IFP Scores. For each target, the corresponding dataset was randomly partitioned into training, test and validation sets (70%:15%:15%). Training and test sets were used for model training and parameter tuning, and the best model was further evaluated on the validation set. Due to the random partitions, this process was repeated five times and the average performance (*corr* and *RMSE*) was reported. In the model-training stage, key parameters were tuned as follows.
   - Substructure-based IFPs: the length of the fingerprint 2^l from 2^3 to 2^{12} ($l = 3, 4, \ldots, 12$),
   - RFs: the number of tree members n_{tree} from 5 to 100 at a step of 5,
   - GBDTs: the boosting stages n_{stage} from 5 to 100 at a step of 5,
   - Regression trees: the maximum depth d from 2 to 50 at a step of 2.

These parameters led to models not too complex for the small target-specific datasets.

The performances of different IFP Scores in the three target-specific tasks are displayed in Fig. 3. The scoring performances vary with different target proteins. Better performances have been observed on the ‘BETA-SECRETASE 1’ dataset, and worse performances on the ‘BROMODOMAIN-CONTAINING PROTEIN 4’ dataset. This implies the dependence of machine-learning modeling on sufficient training data, as the ‘BETA-SECRETASE 1’ dataset has more protein–ligand complexes and the ‘BROMODOMAIN-CONTAINING PROTEIN 4’ dataset less complexes. On the other hand, the performances vary with different machine-learning methods. RFs and

Table 3
Datasets in *PDBbind* database for scoring tasks.

Task	Target protein	Number of protein–ligand complexes	Affinity range (-logKd/Ki)
Target-specific scoring	HIV-1 protease	301	[3.9, 12.7]
	BETA-SECRETASE 1	326	[2.4, 10.77]
	BROMODOMAIN-CONTAINING PROTEIN 4	176	[2.22, 9.15]
Generic scoring	Multiple (refined set)	4852	[2.0, 11.92]
	Multiple (Core Set)	285	[2.07, 11.82]

Table 4
IFPs for constructing target-specific scoring functions.

IFP	$t_{int}(\mathring{A})$	Key interactions	pharmacophoric properties & geometric criteria
SIFt1	4.5	Contact, main-chain atom, side-chain atom, polar, nonpolar, hydrogen-bond donor/acceptor	[20,25]
SIFt2	4.5	Hydrogen-bond donor/acceptor, hydrophobic, polar, nonpolar, aromatic (face-to-face), aromatic (edge to face), metal-acceptor	[20,25]
HIF	10	Hydrogen bonds	[20,38]
CIF	10	Close contacts	[20,38]
CHIF	10	Hydrogen bonds and close contacts	[20,38]
APC	12	Atom-pair counts	[42]
APCiDB	30.5	Atom-pair counts in distance bins	[44]
ECIF	6	Extended connectivity interaction features	[46]
APIF	10	Pairwise interactions (hydrophobic, hydrogen-bond donor/acceptor)	[20,47]
SPLIF	4.5	Implicitly encodes all possible local interactions ($R_{protin} = R_{ligand} = 1$)	[57]
PLEC FP	4.5	Implicitly encodes all possible local interactions ($R_{protin} = 5$, $R_{ligand} = 1$)	[58]
PrtCmm IFP	4.5	Implicitly encodes all possible local interactions ($R_{protin} = R_{ligand} = 1$)	[59]

GBDTs perform better than regression trees, and this partly explains the extensive use of these two methods in earlier scoring works [32,37,42,46,59].

As shown in Fig. 3, given a target protein and a machine-learning method, the performances of IFP Scores largely depend on the IFP models. Generally, IFP Scores originating from atom-pair-counts-based and substructure-based IFPs perform better in these tasks. By averaging the results for the three tasks, we derived Fig. 4. Here the atom-pair-counts-based and substructure-based IFP Scores perform the best, followed by the APIF-based, SIFts-based and atom-based IFP Scores. Compared via Pearson's correlations among each type of machine-learning models, PLEC FP-RF Score ($\bar{corr} = 0.68, \bar{RMSE} = 1.10$), ECIF-GBDT Score ($\bar{corr} = 0.67, \bar{RMSE} = 1.05$) and APC-TREE Score ($\bar{corr} = 0.57, \bar{RMSE} = 1.28$) perform the best.

3.2. Generic scoring

Residue-based and atom-based IFPs are not applicable to generic scoring. Representative IFPs from the other three classes were adopted in this task, and these IFPs were generated as integer vectors. *PDBbind* Refined Set was partitioned into training and test sets (90%:10%) for model training and parameter-tuning, and the best model was evaluated on the Core Set (Table 4, details of data sets in Supplementary file). Overlapping complexes with the Core Set were removed from the refined set to provide a fair validation. Analogous to target-specific scoring tasks, RFs, GBDTs and regression trees were employed to construct IFP Scores. However, due to the increased data, the parameters of these models were tuned in broader ranges as follows.

- RFs: the number of tree members n_{tree} from 300 to 700 at a step of 100,
- GBDTs: the boosting stages n_{stage} from 300 to 700 at a step of 100,
- Regression trees: the maximum depth d from 10 to 100 at a step of 5.

The results are displayed in Fig. 5. As expected, IFP Scores based on regression trees underperform those based on RFs or GBDTs. Given a machine-learning method, the IFP Scores originating from different IFP models behave similarly as in the target-specific scoring tasks. Atom-pair-counts-based and substructure-based IFP Scores perform better than APIF Score, and the best performances of these two classes belong to APCiDB-RF Score ($corr = 0.81$, $RMSE = 1.49$) and PrtCmm-RF Score ($corr = 0.80, RMSE = 1.40$).

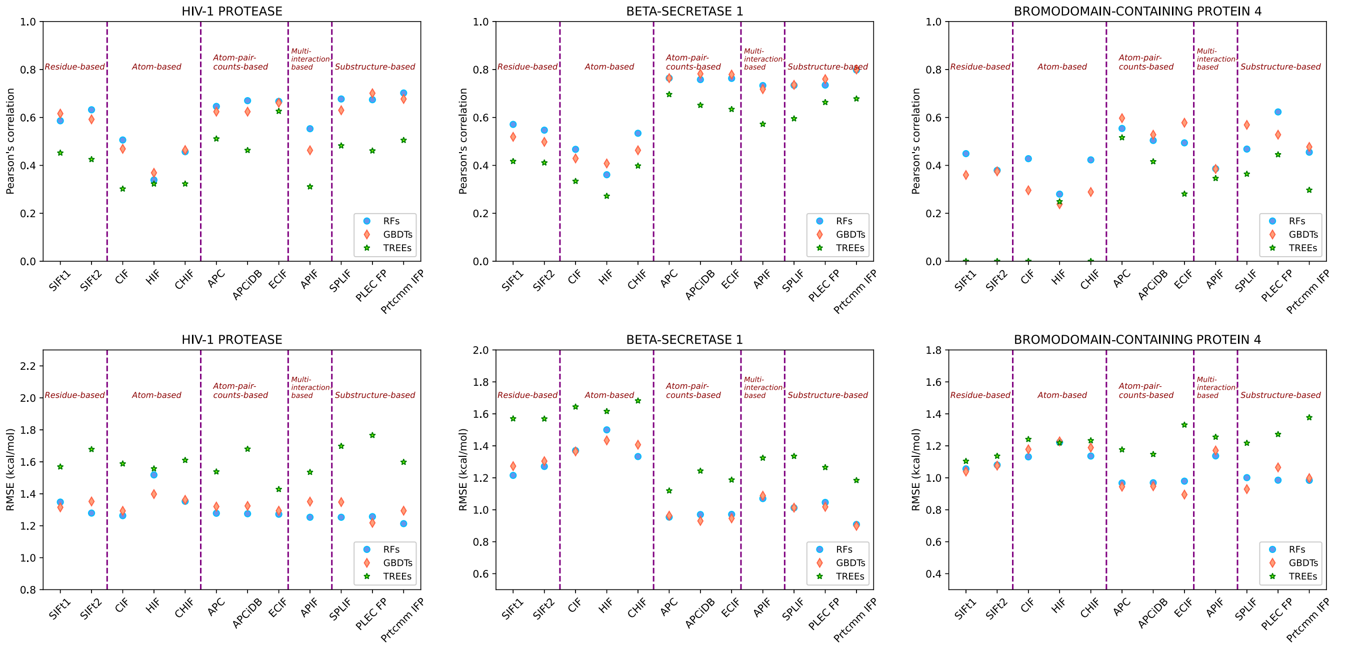

Fig. 3. Performances of IFP Scores in three target-specific tasks (targets: HIV-1 protease, BETA-SECRETASE 1, BROMODOMAIN-CONTAINING PROTEIN 4). These IFP Scores were constructed by associating an IFP model (SIFt1, SIFt2, CIF, HIF, CHIF, APC, APCiDB, ECIF, APIF, SPLIF, PLEC FP or PrtCmm IFP) and a machine-learning method (RFs, GBDTs or regression trees). Performances are evaluated using Pearson's correlation (upper panels) and RMSE (lower panels) between the predicted and experimental affinities.

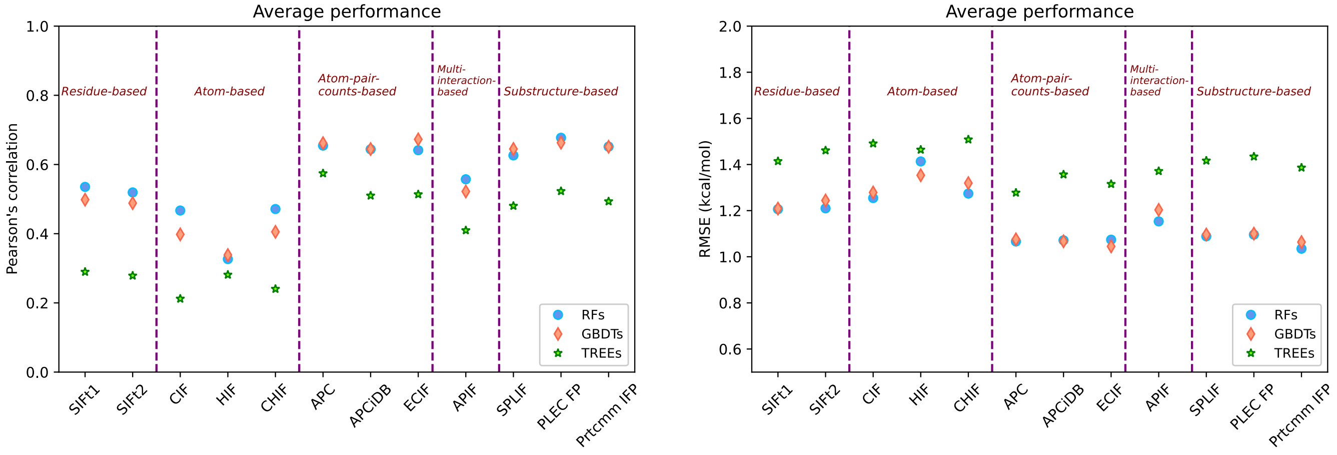

Fig. 4. Average performances of IFP Scores in three target-specific tasks (targets: HIV-1 protease, BETA-SECRETASE 1, BROMODOMAIN-CONTAINING PROTEIN 4). Pearson's correlations and RMSEs between the predicted and experimental affinities are presented in the left and right panels, respectively.

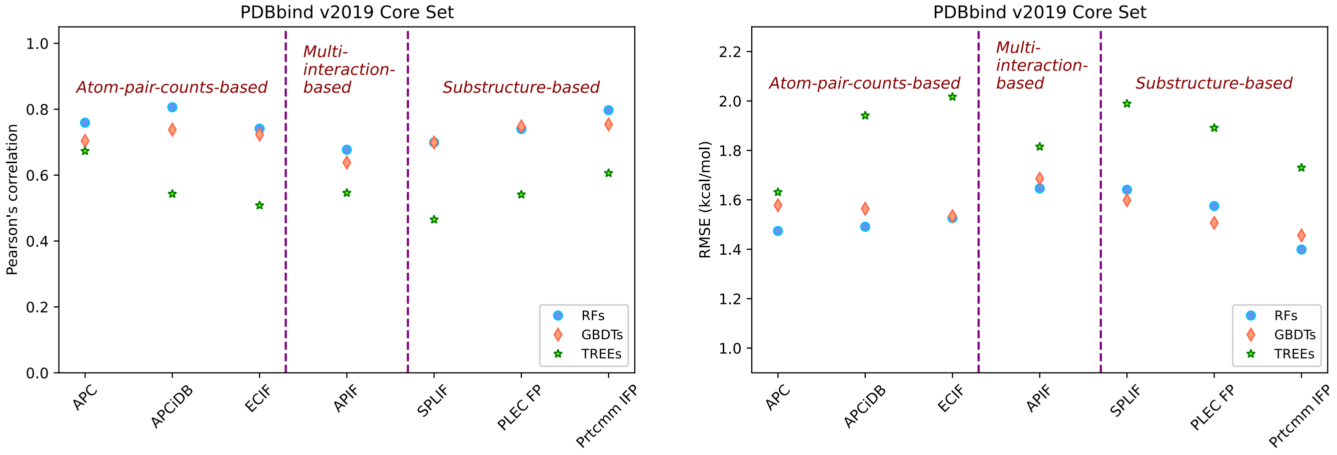

Fig. 5. Performances of IFP Scores in the generic scoring task (multiple targets). The performances are evaluated on *PDBbind* v2019 Core Set, with Pearson's correlations and RMSEs between the predicted and experimental affinities presented in the left and right panels.

4. Summary and outlook

BAP for protein–ligand complexes from their X-ray crystal structures is an important but challenging problem in computer-aided drug design. Plentiful descriptors, which differ in granularity and representation, have been developed and fed into machine-learning algorithms to form BAP-oriented SFs. Recently, the extensive use of deep-learning techniques in various areas also boosted the developments of deep-learning SFs. However, these SFs can only improve the scoring performance marginally, let alone the increasingly complex representations of the descriptors they use [46]. Contrarily, IFPs are simple-format descriptors that carry key protein–ligand interactions, and they can collaborate with classic machine-learning algorithms to form competitive SFs for BAP. In this paper, we reviewed a wide range of IFPs following a building-block-based taxonomy, and compared representative IFP Scores in target-specific and generic scoring tasks. Validated on *PDBbind* v2019 datasets, atom-pair-counts-based and substructure-based IFP Scores performed well, demonstrating their potential in BAP problems.

One limitation of IFP methods is that they rely on the availability of protein–ligand complex structures [21]. However, with the structure-determination techniques (experimental or in silico) becoming more mature, abundant structures have been produced to support the development of IFP methods. Early IFP methods (e.g. SIFts, KB-IFPs and APIFs) use pre-defined interaction types and empirical geometric criteria to detect protein–ligand interactions, which restricts their ability to cover more interactions. Substructure-based IFPs can implicitly account for all types of local interactions, which makes them more competitive in scoring works. Meanwhile, the encoding step in IFP methods has evolved from the interaction-type-based form to a more manageable form, with the introduction of a hash procedure. Altering the basic atom properties, substructure definition and/or encoding ways in these models can potentially improve the scoring performances further. Atom-pair-counts-based IFPs are arguably the simplest feature representations in BAP. They can be easily and rapidly generated, and often lead to good scoring performances. However, these models rely on predefined interaction types (e.g. C–C, C-O, C-S), which may need to be adapted for different training data. Besides, refining the interaction types may promote them to be more promising in scoring works. At last, the performances of IFP Scores (machine-learning-based) are subject to the types and parameters of the involved machine-learning models. Altering the model types or parameters will probably results in fluctuations of the scoring performances. Fine-tuning the parameters in model-training stage using a wider range of values may potentially improve the model performance.

Funding

This work was support by Hong Kong Innovation and Technology Commission and Hong Kong Research Grants Council (Projects 11200818 and UGC/FDS16/M08/18).

CRediT authorship contribution statement

Debby D. Wang: Conceptualization, Methodology, Software, Writing - original draft. **Moon-Tong Chan:** Software, Investigation, Validation. **Hong Yan:** Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.csbj.2021.11.018.

References

[1] Massa Werner. Crystal structure determination. Springer Science & Business Media; 2013.

[2] Tom L Blundell, Bancinyane L Sibanda, Rinaldo Wander Montalvão, Suzanne Brewerton, Vijayalakshmi Chelliah, Catherine L Worth, Nicholas J Harmer, Owen Davies, and David Burke. Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. Philosophical Transactions of the Royal Society B: Biological Sciences, 361(1467):413–423, 2006..

[3] Maurizio Pellecchia, Ivano Bertini, David Cowburn, Claudio Dalvit, Ernest Giralt, Wolfgang Jahnke, Thomas L James, Steve W Homans, Horst Kessler, Claudio Luchinat, et al. Perspectives on nmr in drug discovery: a technique comes of age. Nature reviews Drug discovery, 7(9):738–745, 2008..

[4] Kuhn Bernd, Guba Wolfgang, Hert Jerome, Banner David, Bissantz Caterina, Ceccarelli Simona, Haap Wolfgang, Korner Matthias, Kuglstatter Andreas, Lerner Christian, et al. A real-world perspective on molecular design: miniperspective. J Med Chem 2016;59(9):4087–102.

[5] Jerome De Ruyck, Guillaume Brysbaert, Ralf Blossey, and Marc F Lensink. Molecular docking as a popular tool in drug design, an in silico travel. Advances and applications in bioinformatics and chemistry: AABC, 9:1, 2016..

[6] Lionta Evanthia, Spyrou George, Vassilatis Demetrios K, Cournia Zoe. Structure-based virtual screening for drug discovery: principles, applications and recent advances. Current Topics Med Chem 2014;14(16):1923–38.

[7] Huang Niu, Kalyanaraman Chakrapani, Bernacki Katarzyna, Jacobson Matthew P. Molecular mechanics methods for predicting protein–ligand binding. PCCP 2006;8(44):5166–77.

[8] Mooij Wijnand TM, Verdonk Marcel L. General and targeted statistical potentials for protein–ligand interactions. Proteins: Struct, Funct, Bioinf 2005;61(2):272–87.

[9] Krammer André, Kirchhoff Paul D, Jiang X, Venkatachalam CM, Waldman Marvin. Ligscore: a novel scoring function for predicting binding affinities. J Mol Graph Model 2005;23(5):395–407.

[10] Jain Ajay N. Scoring functions for protein-ligand docking. Curr Protein Pept Sci 2006;7(5):407–20.

[11] Gregory L Warren, C Webster Andrews, Anna-Maria Capelli, Brian Clarke, Judith LaLonde, Millard H Lambert, Mika Lindvall, Neysa Nevins, Simon F Semus, Stefan Senger, et al. A critical assessment of docking programs and scoring functions. J Med Chem, 49(20):5912–5931, 2006..

[12] Li Yan, Minyi Su, Liu Zhihai, Li Jie, Liu Jie, Han Li, Wang Renxiao. Assessing protein–ligand interaction scoring functions with the casf-2013 benchmark. Nature Protocols 2018;13(4):666–80.

[13] Minyi Su, Qifan Yang YuDu, Feng Guoqin, Liu Zhihai, Li Yan, Wang Renxiao. Comparative assessment of scoring functions: the casf-2016 update. J Chem Inform Modeling 2018;59(2):895–913.

[14] Waszkowycz Bohdan, Clark David E, Gancia Emanuela. Outstanding challenges in protein–ligand docking and structure-based virtual screening. Wiley Interdisciplinary Reviews: Computational Molecular. Science 2011;1 (2):229–59.

[15] Dudek Arkadiusz Z, Arodz Tomasz, Gálvez Jorge. Computational methods in developing quantitative structure-activity relationships (qsar): a review. Combinatorial Chem High Throughput Screening 2006;9(3):213–28.

[16] Jiménez José, Skalic Miha, Martinez-Rosell Gerard, De Fabritiis Gianni. K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. J Chem Inform Modeling 2018;58(2):287–96.

[17] Evan N Feinberg, Debnil Sur, Zhenqin Wu, Brooke E Husic, Huanghao Mai, Yang Li, Saisai Sun, Jianyi Yang, Bharath Ramsundar, and Vijay S Pande. Potentialnet for molecular property prediction. ACS central science, 4(11):1520–1530, 2018..

[18] Duc Duy Nguyen, Zixuan Cang, Kedi Wu, Menglun Wang, Yin Cao, and Guo-Wei Wei. Mathematical deep learning for pose and binding affinity prediction and ranking in d3r grand challenges. Journal of computer-aided molecular design, 33(1):71–82, 2019..

[19] Ain Qurrat U, Aleksandrova Antoniya, Roessler Florian D, Ballester Pedro J. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. Wiley Interdisciplinary Reviews: Computational Molecular. Science 2015;5(6):405–24.

[20] Desaphy Jeremy, Raimbaud Eric, Ducrot Pierre, Rognan Didier. Encoding protein–ligand interaction patterns in fingerprints and graphs. J Chem Inform Modeling 2013;53(3):623–37.

[21] Sebastian Salentin, V Joachim Haupt, Simone Daminelli, and Michael Schroeder. Polypharmacology rescored: Protein–ligand interaction profiles for remote binding site similarity assessment. Progress in biophysics and molecular biology, 116(2–3):174–186, 2014..

[22] Crisman Thomas J, Sisay Mihiret T, Bajorath Jurgen. Ligand-target interaction-based weighting of substructures for virtual screening. J Chem Inform Modeling 2008;48(10):1955–64.

[23] Hongjian Li, Kam-Heung Sze, Gang Lu, and Pedro J Ballester. Machine-learning scoring functions for structure-based drug lead optimization. Wiley Interdisciplinary Reviews: Computational Molecular Science, 10(5):e1465, 2020..

[24] Shen Chao, Ding Junjie, Wang Zhe, Cao Dongsheng, Ding Xiaoqin, Hou Tingjun. From machine learning to deep learning: Advances in scoring functions for protein–ligand docking. Wiley Interdisciplinary Rev: Comput Mol Sci 2020;10 (1):e1429.

[25] Deng Zhan, Chuaqui Claudio, Singh Juswinder. Structural interaction fingerprint (sift): a novel method for analyzing three-dimensional protein-ligand binding interactions. J Med Chem 2004;47(2):337–44.

[26] Radifar Muhammad, Yuniarti Nunung, Istyastono Enade Perdana. Pyplif: Python-based protein-ligand interaction fingerprinting. Bioinformation 2013;9(6):325.

[27] Deng Zhan, Chuaqui Claudio, Singh Juswinder. Knowledge-based design of target-focused libraries using protein- ligand interaction constraints. J Med Chem 2006;49(2):490–500.

[28] Mordalski Stefan, Kosciolek Tomasz, Kristiansen Kurt, Sylte Ingebrigt, Bojarski Andrzej J. Protein binding site analysis by means of structural interaction fingerprint patterns. Bioorganic Med Chem Letters 2011;21(22):6816–9.

[29] Marcou Gilles, Rognan Didier. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. J Chem Inform Modeling 2007;47 (1):195–207.

[30] Nandigam Ravi K, Kim Sangtae, Singh Juswinder, Chuaqui Claudio. Position specific interaction dependent scoring technique for virtual screening based on weighted protein- ligand interaction fingerprint profiles. J Chem Inform Modeling 2009;49(5):1185–92.

[31] Guo-Bo Li, Zhu-Jun Yu, Sha Liu, Lu-Yi Huang, Ling-Ling Yang, Christopher T Lohans, and Sheng-Yong Yang. Ifptarget: a customized virtual target identification method based on protein–ligand interaction fingerprinting analyses. J Chem Inform Modeling, 57(7):1640–1651, 2017..

[32] Leidner Florian, Yilmaz Nese Kurt, Schiffer Celia A. Target-specific prediction of ligand affinity with structure-based interaction fingerprints. J Chem Inform Modeling 2019;59(9):3679–91.

[33] Sun Huiyong, Pan Peichen, Tian Sheng, Lei Xu, Kong Xiaotian, Li Youyong, Li Dan, Hou Tingjun. Constructing and validating high-performance miec-svm models in virtual screening for kinases: a better way for actives discovery. Sci Rep 2016;6(1):1–12.

[34] Chen Fu, Sun Huiyong, Liu Hui, Li Dan, Li Youyong, Hou Tingjun. Prediction of luciferase inhibitors by the high-performance miec-gbdt approach based on interaction energetic patterns. PCCP 2017;19(15):10163–76.

[35] Yan Yuna, Wang Weijun, Sun Zhaoxi, Zhang John ZH, Ji Changge. Protein–ligand empirical interaction components for virtual screening. J Chem Inform Modeling 2017;57(8):1793–806.

[36] Yasuo Nobuaki, Sekijima Masakazu. Improved method of structure-based virtual screening via interaction-energy-based learning. J Chem Inform Modeling 2019;59(3):1050–61.

[37] Beihong Ji, Xibing He, Jingchen Zhai, Yuzhao Zhang, Viet Hoang Man, and Junmei Wang. Machine learning on ligand-residue interaction profiles to significantly improve binding affinity prediction. Briefings in Bioinformatics, 2021..

[38] Mpamhanga Chidochangu P, Chen Beining, McLay Iain M, Willett Peter. Knowledge-based interaction fingerprint scoring: a simple method for improving the effectiveness of fast scoring functions. J Chem Inform Modeling 2006;46(2):686–98.

[39] Kelly Matthew D, Mancera Ricardo L. Expanded interaction fingerprint method for analyzing ligand binding modes in docking and structure-based drug design. J Chem Inform Computer Sci 2004;44(6):1942–51.

[40] Jasper Julia B, Humbeck Lina, Brinkjost Tobias, Koch Oliver. A novel interaction fingerprint derived from per atom score contributions: exhaustive evaluation of interaction fingerprint performance in docking based virtual screening. J Cheminformatics 2018;10(1):1–13.

[41] Korb Oliver, Stutzle Thomas, Exner Thomas E. Empirical scoring functions for advanced protein- ligand docking with plants. J Chem Inform Modeling 2009;49(1):84–96.

[42] Ballester Pedro J, Mitchell John BO. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics 2010;26(9):1169–75.

[43] Ouyang Xuchang, Handoko Stephanus Daniel, Kwoh Chee Keong. Cscore: a simple yet effective scoring function for protein–ligand binding affinity prediction using modified cmac learning architecture. J Bioinform Comput Biol 2011;9(supp01):1–14.

[44] Zheng Liangzhen, Fan Jingrong, Yuguang Mu. Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. ACS Omega 2019;4(14):15956–65.

[45] Zechen Wang, Liangzhen Zheng, Yang Liu, Yuanyuan Qu, Yong-Qiang Li, Mingwen Zhao, Yuguang Mu, and Weifeng Li. Onionnet-2: A convolutional neural network model for predicting protein-ligand binding affinity based on residue-atom contacting shells. arXiv preprint arXiv:2103.11664, 2021..

[46] Sánchez-Cruz Norberto, Medina-Franco José L, Mestres Jordi, Barril Xavier. Extended connectivity interaction features: Improving binding affinity prediction through chemical description. Bioinformatics 2021;37(10):1376–82.

[47] Pérez-Nueno Violeta I, Rabal Obdulia, Borrell José I, Teixidó Jordi. Apif: a new interaction fingerprint based on atom pairs and its application to virtual screening. J Chem Inform Modeling 2009;49(5):1245–60.

[48] Sato Tomohiro, Honma Teruki, Yokoyama Shigeyuki. Combining machine learning and pharmacophore-based interaction fingerprint for in silico screening. J Chem Inform Modeling 2010;50(1):170–85.

[49] Xue Ling, Bajorath Jurgen. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. Comb Chem High Throughput Screening 2000;3(5):363–72.

[50] Hajduk Philip J, Greer Jonathan. A decade of fragment-based drug design: strategic advances and lessons learned. Nature Reviews Drug discovery 2007;6(3):211–9.

[51] Rogers David, Hahn Mathew. Extended-connectivity fingerprints. J Chem Inform Modeling 2010;50(5):742–54.

[52] Rarey Matthias, Kramer Bernd, Lengauer Thomas, Klebe Gerhard. A fast flexible docking method using an incremental construction algorithm. J Mol Biol 1996;261(3):470–89.

[53] Tan Lu, Lounkine Eugen, Bajorath Jurgen. Similarity searching using fingerprints of molecular fragments involved in protein- ligand interactions. J Chem Inform Modeling 2008;48(12):2308–12.

[54] Durant Joseph L, Leland Burton A, Henry Douglas R, Nourse James G. Reoptimization of mdl keys for use in drug discovery. J Chem Inform Computer Sci 2002;42(6):1273–80.

[55] José Batista, Lu Tan, and Jurgen Bajorath. Atom-centered interacting fragments and similarity search applications. J Chem Inform Modeling, 50(1):79–86, 2010..

[56] Vass Márton, Kooistra Albert J, Ritschel Tina, Leurs Rob, de Esch Iwan JP, de Graaf Chris. Molecular interaction fingerprint approaches for gpcr drug discovery. Current Opinion Pharmacol 2016;30:59–68.

[57] Da C, Kireev D. Structural protein–ligand interaction fingerprints (splif) for structure-based virtual screening: method and benchmark study. J Chem Inform Modeling 2014;54(9):2555–61.

[58] Wójcikowski Maciej, Kukiełka Michał, Stepniewska-Dziubinska Marta M, Siedlecki Pawel. Development of a protein–ligand extended connectivity (plec) fingerprint and its application for binding affinity predictions. Bioinformatics 2019;35(8):1334–41.

[59] Wang Debby D, Xie Haoran, Yan Hong. Proteo-chemometrics interaction fingerprints of protein–ligand complexes predict binding affinity. Bioinformatics 2021.

[60] Wang Renxiao, Fang Xueliang, Yipin Lu, Yang Chao-Yie, Wang Shaomeng. The pdbbind database: methodologies and updates. J Med Chem 2005;48(12):4111–9.

[61] Da Silva Franck, Desaphy Jeremy, Rognan Didier. Ichem: a versatile toolkit for detecting, comparing, and predicting protein–ligand interactions. ChemMedChem 2018;13(6):507.

[62] Chemical Computing Group Inc. Molecular operating environment (moe), 2016..

[63] Wójcikowski Maciej, Zielenkiewicz Piotr, Siedlecki Pawel. Open drug discovery toolkit (oddt): a new open-source player in the drug discovery field. J Cheminformatics 2015;7(1):1–6.

[64] Mark R Segal. Machine learning benchmarks and random forest regression. 2004..

[65] Peter Prettenhofer and Gilles Louppe. Gradient boosted regression trees in scikit-learn. 2014..

[66] Roger J Lewis. An introduction to classification and regression tree (cart) analysis. In Annual meeting of the society for academic emergency medicine in San Francisco, California, volume 14, 2000..

Debby D. Wang is conducting computational predictions of protein-ligand binding affinity and mutation-induced affinity changes, and developing bioinformatics and health-informatics tools. She is an Assistant Professor in the School of Health Science and Engineering, University of Shanghai for Science and Technology.

Moon-Tong Chan's research interests include regression analysis, generalized linear mixed models and multilevel statistical models. He is currently an Assistant Professor in School of Science and Technology, Hong Kong Metropolitan University.

Hong Yan's research interests include biomolecular pattern recognition, image processing and high-performance computing. He is Chair Professor of Computer Engineering and Wong Chung Hong Professor of Data Engineering at City University of Hong Kong.