A functional feature analysis on diverse protein–protein interactions: application for the prediction of binding affinity

Jiesi Luo · Yanzhi Guo · Yun Zhong ·
Duo Ma · Wenling Li · Menglong Li

Received: 20 January 2014 / Accepted: 26 April 2014 / Published online: 1 May 2014

Abstract Protein–protein interactions (PPIs) play crucial roles in diverse cellular processes. There are different types of PPIs based on the composition, affinity and whether the association is permanent or transient. Analyzing the diversity of PPIs at the atomic level is crucial for uncovering the key features governing the interactions involved in PPI. A systematic physico-chemical and conformational studies were implemented on interfaces involved in different PPIs, including crystal packing, weak transient heterodimers, weak transient homodimers, strong transient heterodimers and homodimers. The comparative analysis shows that the interfaces tend to be larger, less planar, and more tightly packed with the increase of the interaction strength. Meanwhile the strong interactions undergo greater conformational changes than the weak ones involving main chains as well as side chains. Finally, using 18 features derived from our analysis, we developed a support vector regression model to predict the binding affinity with a promising result, which further demonstrate the reliability of our studies. We believe this study will provide great help in more thorough understanding the mechanism of diverse PPIs.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-014-9746-y) contains supplementary material, which is available to authorized users.

J. Luo · Y. Guo (✉) · Y. Zhong · D. Ma · W. Li · M. Li (✉)
College of Chemistry, Sichuan University,
Chengdu 610064, Sichuan, People's Republic of China
e-mail: yzguo@scu.edu.cn

M. Li
e-mail: liml@scu.edu.cn

Keywords Protein–protein interactions · Interface properties · Conformational analysis · Regression model prediction

Introduction

Interactions between proteins play essential roles in all aspects of cellular physiology. The structural models of complexes formed by protein–protein interactions (PPIs) in the Protein Data Bank (PDB) provide an opportunity to understand the principles governing the interactions involved in protein–protein recognition. However, not all interactions observed in structural data determined by X-ray crystallography are biologically relevant. Many of them are artifacts of crystallization that would not appear in the physiological state or in solution. These interactions are called crystal packing contacts or non-specific interfaces as they are not biologically associated. Distinguishing biologically relevant interfaces from lattice contacts in protein crystals is still a fundamental problem in structural biology [1–5]. In addition, biological interactions can be further classified into diverse types based on their composition, affinity and life time [6–8]. These categories include the distinction between homo-oligomeric and hetero-oligomeric complexes, where the former is a PPI between identical chains, as opposed to hetero-oligomeric complexes. Another distinction is between obligate and non-obligate complexes. In an obligate complex, the protomers are unstable on their own in vivo, whereas the protomers of non-obligate complexes can dissociate from each other and exist as stable structures. Similarly, protein complexes can be divided into permanent and transient according to their lifetimes. Permanent complexes are very strong and irreversible. However, the transient ones readily undergo changes in the oligomeric state.

In recent years, several studies have examined the general interface properties of protein complexes in order to distinguish different types of PPIs. Jones and Thornton explored the interface properties of four different PPI types: homodimers, heterodimers, enzyme-inhibitor complexes and antibody-protein complexes. They concluded that homodimers are generally larger and more hydrophobic than the other three types of protein complexes [9]. Ponstingl et al. [4] used contact area and atom pair frequencies to differentiate homodimers and monomers. Nooren and Thornton [10] examined the interfaces of weak transient homodimers and strong transient heterodimers and found that weak transient homodimers tend to have smaller, more planar and polar interfaces than strong transient heterodimers. Bahadur et al. compared the physicochemical and geometrical properties of interfaces involved in specific and non-specific PPIs. Results suggest that non-specific interfaces are less hydrophobic and contain much fewer fully buried atoms than specific interfaces [1]. De et al. analyzed the interfaces of obligate and non-obligate complexes. They show that there is a clear trend for the obligate interfaces to be non-polar and involve stable secondary structural elements across the interfaces [11]. Zhu et al. [5] used support vector machine (SVM) to differentiate obligate, non-obligate and crystal packing interactions. Guharoy and Chakrabarti [12] presented a detailed secondary structural analysis on homodimeric and heterodimeric interfaces. Bernauer et al. [2] developed a new method of Voronoi tessellation to discriminate the interfaces between homodimers and crystal packing. Dey et al. [13] compared weak transient homodimers and permanent homodimers, indicating that weak homodimers are loosely packed which may contribute to their low stability. Recently, La et al. [14] constructed amino acid substitution models to predict permanent and transient interfaces.

Although these studies have give their interface distinctions, it is important to recognize that these distinctions are not entirely clear cut and a continuum does exist between the different types of interactions. Reviewing the previous studies, an analysis has not been implemented on a comprehensive PPI data including all types. In this article, PPIs are divided into five types, including crystal packing, weak transient heterodimers, weak transient homodimers, strong transient heterodimers and homodimers. We tried to collect data covering a widest range of PPIs from non-specific to specific, from weak to strong and from homos to heteros, so as to reveal the factors that influence the formation of PPIs of different types. We first investigated the differences of eight physicochemical properties among these interactions. Then further analysis on the conformational changes was carried out. We got a general rule that governs these different interactions and interesting findings were also obtained. Finally, based on the functional features derived from our analysis, a support vector regression (SVR) model was constructed to quantitatively represent the strength of PPIs. The results demonstrate that our analysis will be a useful complementary to existing method for analyzing PPIs, and thus contribute to better understand and eventually design diverse PPIs.

Materials and methods

Dataset

Here, five sets of protein–protein interfaces were collected. They are crystal packing, weak transient heterodimers, weak transient homodimers, strong transient heterodimers and homodimers, respectively. Crystal packing can be generated by rotating and translating the asymmetric units (ASU) according to symmetry operators provided for each crystal. The ASUs are the smallest unit of the crystal, and they interact with each other both within the ASU and among the adjacent ASUs to stabilize the crystal. The interactions adjacent ASUs are usually designated as crystal-packing as they do not have any biological specificity, and are abundantly available in the PDB as they govern the molecular packing of protein crystals [1–4]. In this work, 188 crystal packing dimers were retrieved from the work of Bahadur et al. [1]. The other four types of interfaces can be considered to be biologically relevant and they also can be discriminated as permanent interactions and transient interactions based on the lifetime or stability of the complex. The permanent interactions are usually very stable and irreversible whereas transient interactions associate and dissociate temporarily in vivo [6–8]. Most of the homodimers are permanent as they assemble tightly as soon as they are synthesized, and stay together over times longer than the life of a cell. So we used 117 homodimers representing permanent interactions from another work of Bahadur et al. [17].

The transient interactions, depending on their functional roles in the cell, have a wide range of affinities and lifetimes and hence can be further subdivided into weak and strong. Weak transient interactions show a fast bound–unbound equilibrium with the dissociation constant (K_d) values typically in the μM range. The strong transient interactions, triggered by binding of an effector molecule or conformational change, may last longer and have a continuum of K_d that exists between the weak and the more permanent interactions [6–8]. Here, 44 strong transient heterodimers were collected from the Affinities dataset [15] and the work of Nooren [10] as strong transient interactions with K_d ranging between 10^{-9} and 10^{-6} M. In previous studies, Nooren [10] and Dey [13] both surveyed the biochemical literature to identify homodimeric protein assemblies that have K_d in the μM range or higher. They marked these proteins "weak" to indicate that they dissociate easily. So 51 weak transient homodimers were retrieved from Nooren [10] and Dey [13] as weak transient interactions after removing duplicate structures. We also collected 54 weak transient heterodimers with the K_d values of higher than 1.0×10^{-6} M as weak transient interactions from the Affinities dataset [15] and PDBbind dataset [16], because PPI interfaces differ between homodimers and heterodimers [1, 9, 11]. Although both two categories of interfaces may represent weak interactions, there are still differences between weak transient homodimers and weak transient heterodimers. A summary of all the datasets is shown in the Supplementary Table S1 of Supporting information file S1.

Methods

Protein–protein interfaces have been defined according to the change in solvent accessible surface area (ΔASA) on binding. Here, ASAs were calculated using NACCESS [18], with a probe sphere of radius 1.4 Å. Residues with relative solvent accessible surface area (RSA) >25 % are identified as surface residues [19] and those are the interface residues if they have ASAs that decrease by >0.1 Å^2 upon binding [9]. Besides, residues with at least one atom that is fully buried interface (ASA = 0) are defined as core residues, while rim residues do not contain any fully buried atoms. Hot spots have been defined as those sites where alanine mutations cause a significant increase in the binding free energy of at least 2.0 kcal/mol [20].

Interface properties

Here, eight important interface properties were calculated to reveal the structural basis of different types of PPIs. They are the interface area (IA), the ratio of interface area/surface area ($R_{i/s}$), the non-polar area fraction (f_{np}), the fully buried atoms fraction (f_{bu}), core area fraction (f_{core}), residue propensity score (R_p), local density index (LD) and the number of residues on host spots (Nhs), respectively. IA is defined as the total decrease of ASA of the two proteins upon interaction and it reflects the size of the interfaces. $R_{i/s}$ is the ratio of the interface area to the rest area of the protein surface in the two subunits. The f_{np} is measured as the percentage of the interface area contributed by non-polar atoms and it reflects the hydrophobic property of the complexes. The f_{bu} denotes the fraction of interface atoms fully buried in the complex with zero ASA to the total number of interface atoms and it reflects the packing density of the interfaces. The f_{core} is defined as the percentage of the interface area contributed by core residues and it reflects the size of interface core region. According to the description by Bahadur et al. [1], the selection or exclusion of certain types of amino acids at an interface can be expressed as a set of propensities:

$$P_i = \ln(f_i/f_i^{\circ}) \tag{1}$$

where f_i is the number or area fraction contributed by residue type i to the interface, $f_i^{\circ}$is the equivalent fraction contributed to the protein surface. Summing the propensities of all the residues present in an interface yields the R_p score, which indicates whether its amino acid composition predicts the assembly to be stable ($R_p > 0$) or labile ($R_p < 0$). The shape complementation of the interfaces is measured by the LD index [1]. The LD index is defined as follow, for each interface atom a, the number n_a of the interface atoms are counted within an optimized distance 12 Å of atom a in the same subunit; then, n_a is averaged over all N interface atoms:

$$LD = 1/N \sum n_a \tag{2}$$

In this work, the IA, $R_{i/s}$, f_{np}, f_{bu}, f_{core}, R_p score and LD index were calculated using the ProFace program [21] and the hot spots are predicted by the computational interface alanine scanning server [20].

Conformational analysis

Protein structures undergo a variety of conformational changes upon interaction. The conformational changes may mediate signaling transduction or trigger allosteric effects and have an essential role in many biological processes. The observed conformational changes can be divided into different types of motion, including domain and loop motions, unfolding and refolding of regular secondary structure, disorder to order transitions, and other changes involving protein backbone and side chain [19, 22–24]. Here, the root mean squared deviation (RMSD) is used to measure the conformational differences between the unbound and bound structures. It is calculated based on a superposition of the two structures using all atoms concerned. The superpositions were made on all residues, surface residues and interface residues, respectively. The RMSD was computed by the ProFit program and the domain motions, hinge axes and hinge bending motion were predicted using DynDom program [25].

A detailed list of the totally properties used in this work to parameterize PPIs and the software and papers used to calculate these properties are shown in the Supporting information file S2 and the Supplementary Table S4 of Supporting information file S1, respectively.

Regression model prediction

Estimating the interaction strength is not only important for elucidating the molecular mechanism underlying interactions, but also essential for developing effective tools for protein–protein docking. Nowadays, the binding affinity prediction methods are mainly based on four strategies: physical-based force fields [26], knowledge-based statistical potentials [27, 28], empirical scoring functions [29] and hybrid scoring functions [30, 31]. Although these methods are successful in protein–protein docking evaluation and some of them have made a great progress in binding affinity prediction, there is still a space to improve the performance. As an alternative to modelling assumptions in scoring functions, the machine learning methods have already been used in the prediction of binding affinity of protein–ligand and protein–protein complexes. Ballester and Mitchell [32] presented the application of random forests (RFs) to predicting protein–ligand binding affinity and Li et al. [33] built a SVR model to predict protein–protein binding affinity. The machine learning methods can achieve satisfactory results without assuming any predefined model when used for binding affinity prediction. In this work, we employed a SVR model to establish the correlations between the functional features and binding affinity of the protein complexes. The fivefold cross-validation was used to train and test the stability of regression model. 48 heterodimers from the Affinities dataset in this work are randomly split into 5 groups with approximately equal sample size. Four groups are selected as the training set for developing regression model, the remaining group as the testing set for evaluating it. This process is repeated five times so that every group is used as a testing set once. An external test set of 20 heterodimers was collected for blindly evaluating the predictive power of the model. The performance of regression model can be measured by Pearson's correlation coefficient (R) and root mean squared error (RMSE):

$$R = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)(y_i^{pred} - \bar{y}_i^{pred})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \sum_{i=1}^{n}(y_i^{pred} - \bar{y}_i^{pred})^2}} \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^{pred})^2} \quad (4)$$

where n is the number of samples in the training and test sets, y_i and $y_i^{pred}$are the values of experimentally determined affinity and estimated affinity respectively, and $\bar{y}_i$ and $\bar{y}_i^{pred}$are the average of y_i and y_i^{pred}respectively.

Results and discussion

Interface properties

Table 1 gives the average values of the physiochemical properties in different types of interfaces and the distributions of these properties are shown in Fig. 1.

The size of the interfaces

Figure 1a shows that the IA increases from crystal packing to homodimers, although the average IA of the weak transient heterodimers is slightly smaller than that of crystal packing (Table 1). It should be noted that the average IA of crystal packing used in this work is slightly larger than that of other packing interfaces. A smaller average IA of 570 Å^2 has been obtained from a sample of 1,320 packing interfaces found in 152 crystals of monomeric proteins [34]. Since the areas of most crystal packing contacts are much smaller than those of biological interfaces, IA is deemed as the most important feature for distinguishing biological interactions from crystal packing in previous studies [2, 4, 5]. However, there are also cases with small biological interfaces found in this work. The weak transient heterodimers contain interfaces similar in size to those in the crystal packing, which indicates that the IA alone may be not sufficient to discriminate the interfaces between biological and crystal packing. On the other hand, it also implies that the weak transient heterodimers may be the lower IA limit for specific recognition.

The $R_{i/s}$ in Fig. 1b displays almost identical trend to the observation on IA. The average $R_{i/s}$ is smallest for crystal packing and the largest for homodimers. On average, the $R_{i/s}$ of homodimers is 1.5-fold larger than those of strong transient heterodimers and weak transient homodimers, and about twofold larger than those of weak transient heterodimers and crystal packing (Table 1). The high $R_{i/s}$ of homodimers may indicate that a more stable interface not only requires a larger interface area but also a greater fraction of surface area involving in the formation of interface.

Composition of the protein–protein interfaces

The chemical composition of the interfaces can be divided into two types: non-polar (carbon containing) and polar (N, O and S containing). From Table 1, the non-polar atoms contribute 58 % of the interface area in the crystal packing. The interfaces of weak transient heterodimers, weak transient homodimers and strong transient heterodimers have an average non-polar fraction between 59 and 63 %. Interfaces formed in homodimers tend to be most hydrophobic with about 65 % non-polar on average. Interestingly, the average f_{np} of weak transient homodimers is

Table 1 Average physicochemical properties of different types of protein–protein interactions

Interface	Weak protein–protein interaction		Strong protein–protein interaction		Nonspecific interfaces
	Weak transient heterodimers	Weak transient homodimers	Strong transient heterodimers	Homodimers	Crystal packing
Number	54	51	44	117	188
No. of interface residues	41 ± 10	43 ± 12	65 ± 22	102 ± 52	47 ± 15
Core residues	19 ± 6	24 ± 8	34 ± 12	63 ± 35	19 ± 8
Rim residues	22 ± 7	20 ± 10	31 ± 13	40 ± 23	28 ± 12
Core atoms	143 ± 34	159 ± 47	235 ± 81	390 ± 202	158 ± 52
No. of interface atoms	39 ± 16	50 ± 23	69 ± 28	146 ± 91	34 ± 18
Rim atoms	105 ± 26	110 ± 36	166 ± 62	244 ± 124	125 ± 42
Interface area (Å2)	677 ± 148	764 ± 224	1,132 ± 398	1,900 ± 1,003	743 ± 250
Composition (IA %)					
Core area	64.6 ± 14.4	73.0 ± 13.7	70.7 ± 9.3	77.5 ± 9.8	57.9 ± 15.8
Rim area	35.4 ± 14.4	27.0 ± 13.7	29.3 ± 9.3	22.5 ± 9.8	42.1 ± 15.8
Non-polar(fnp)	58.5 ± 7.2	62.9 ± 8.2	59.4 ± 5.7	65.3 ± 6.9	57.9 ± 8.4
Interface area ratio (%)	8.4 ± 3.3	10.1 ± 3.7	10.3 ± 3.4	16.1 ± 7.1	6.7 ± 2.7
Rp propensity score	0.12 ± 2.36	1.46 ± 2.34	−0.30 ± 2.36	4.30 ± 5.05	−1.13 ± 2.64
Atomic packing					
% buried atoms (fbu)	26.7 ± 8.0	30.9 ± 9.4	30.0 ± 7.6	36.5 ± 8.5	21.1 ± 8.2
Local density index (LD)	34.3 ± 5.2	35.2 ± 5.8	39.2 ± 4.9	45.0 ± 8.1	31.7 ± 5.8
No. of hot spot	8 ± 3	9 ± 4	14 ± 6	26 ± 16	6 ± 4

Data are expressed as mean ± SD

slightly larger than that of strong transient heterodimers in this work, as shown in Fig. 1c. One possible explanation is that the two types of interactions have different assembling modes. The subunits that form heterodimers are often, but not always, independently stable inside the cell and they interact with each other to carry out a specific function. On the other hand the subunits of homodimers are typically not found as stable structures inside the cell and the complex formation occurs simultaneously during the folding process. Before interaction, the high f_{np} on the protein surface may form large hydrophobic patches. Burying of large hydrophobic patches in homodimers favors the formation of a more stable assembly. However, the subunits of heterodimers are stable in the cell before the formation of the complex assembles. So the subunits surface that is buried at the interface of heterodimers is in contact with water until interaction. Large hydrophobic patches on the surface would increase the tendency of aggregation, which makes the subunits insoluble.

The average chemical composition of the interfaces of crystal packing is very close to that of the solvent-accessible protein surface [35]. It also can be further confirmed by the residue propensity score (R_p). The R_p shows the preference of all amino acid types at an interface versus the surface. The crystal packing has a negative average R_p, indicating the AAC of interfaces closely resemble those of protein surfaces. All other four types of interfaces have a positive R_p on average. Moreover, Fig. 1d demonstrates that like the f_{np}, the average R_p of the weak transient homodimers is higher than that of strong transient heterodimers. The relatively low R_p in weak and strong transient heterodimers and relatively high R_p in weak transient homodimers and homodimers may suggest that the difference between the composition of interface and solvent-accessible surface may be less marked in heterodimers but strongly marked in homodimers.

Shape of the interfaces

Compared with other parameters, i.e. planarity, circularity [9] and gap volume index [36] for characterizing the shape complementarity of the two protein surfaces in contact, LD is less sensitive to edge effects in the atomic positions of the X-ray structures [13]. It measures the packing density at each point of one interface. The average values of LD are 45 for the homodimers, 39 for strong transient heterodimers, 35 for weak transient homodimers, 34 for weak transient heterodimers and 32 for crystal packing, respectively (Table 1). The trend shown in Fig. 1e indicates that the interfaces are better packed from crystal packing to homodimers.

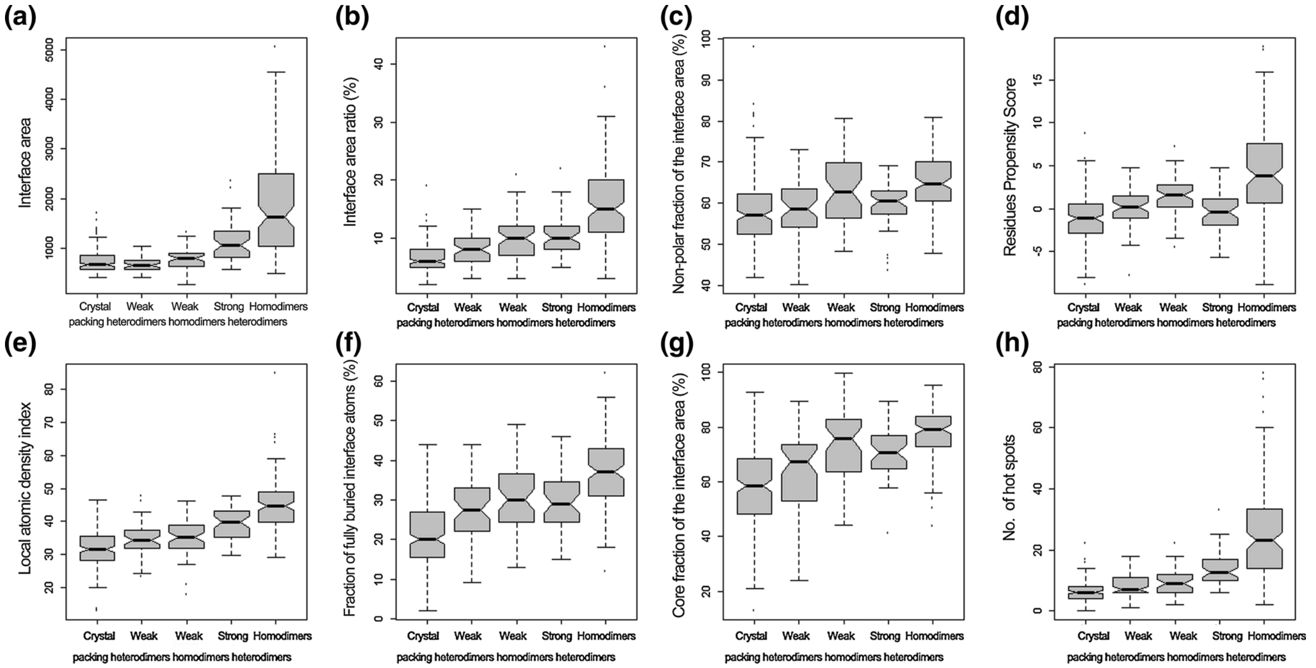

Fig. 1 Distribution of physiochemical properties in different types of protein–protein interactions. *Boxplots* for **a** interface area; **b** interface area ratio; **c** non-polar area fraction; **d** residues propensity score; **e** local atomic density index; **f** fully buried atoms fraction; **g** core area fraction; **h** hot spots

Buried interface atoms and core area fraction

For the specific interactions, about 27–37 % of the interface atoms are fully buried. In crystal packing, the f_{bu} is only 21 % (Table 1). The fraction of fully buried atoms can be affected by the packing density of interfaces. Tightly packed interfaces have more number of contacts and remove water from the core region of protein interface. Therefore, Fig. 1f indicates that more atoms are buried as the interfaces are better packed. The core residues are defined as those that contain buried interface atoms. Interface can be divided into two distinct regions, the core and rim region. The core region usually occupies the interface center and it is surrounded by the rim region. On average the core residues constitute 65–77 % of the interface area in specific interactions, but in crystal packing, the core area fraction is only 58 % (Table 1). Figure 1g shows the core fraction of interface area from crystal packing to homodimers.

Through further analysis, we plotted the core area fraction of each interface versus its interface area in Fig. 2. It shows that crystal packing has a wide distribution, from 13 to 93 %, but the specific interactions form a distinct cluster. Taking $f_{\mathrm{core}} = 65$ % as a cutoff between specific interactions and crystal packing, we found that 75 % of the former have $f_{\mathrm{core}} > 65$ %, 70 % of the latter have $f_{\mathrm{core}} < 65$ %. This result suggests that fraction of the interface area contributed by the core residues may be a key determinant of biological interfaces.

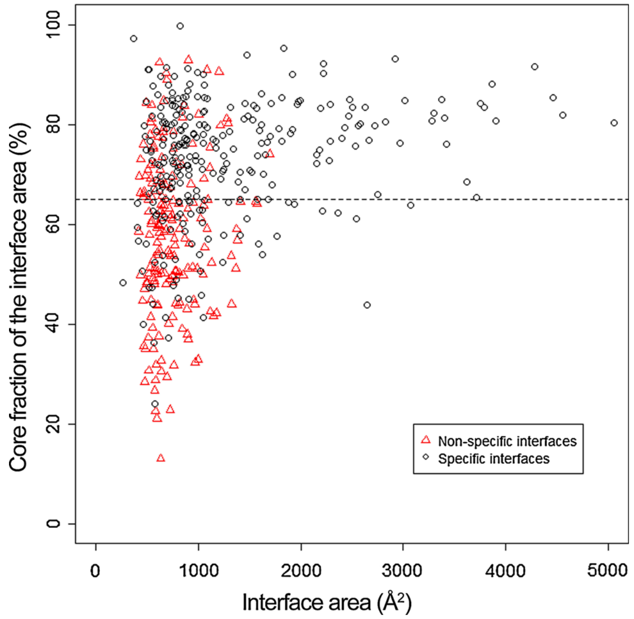

Fig. 2 Interface area and core area fraction of the interfaces in specific and non-specific interactions

Hot spots

On average Nhs is 26, 14, 9, 8 and 6 for homodimers, strong transient heterodimers, weak transient homodimers, weak transient heterodimers and crystal packing, respectively.

Figure 1h shows that there are more hot spots in the interfaces from crystal packing to homodimers, which can provide a proof that the core residues are very important in distinguishing biological interfaces. The core residues are fully buried and may contribute significantly to the binding energy. However, the partially buried residues are rarely found to contribute to the binding energy. So the core residues are a necessary but not sufficient condition for hot spots.

Conformational changes associated with protein–protein interactions

An essential feature of protein–protein interactions can well affect the conformation of the protein components. The extent of the conformational changes can be assessed when the X-ray structures of both unbound and bound forms are known independently. In this work, only 48 transient heterodimers collected from the Affinities dataset have high-resolution unbound and bound structures. However, these unbound and bound structures may be determined by different experimental methods or different experimental conditions. Therefore, there are experimental differences between the unbound and bound structures. To obtain an estimated value for the extent of conformational changes that can be expected from experimental differences in the determination of crystal structures, a control experimental dataset is needed. We searched 100 % identical sequences without non-water heteroatoms in PDB for all the unbound components of the 48 transient heterodimers. If more than two structures were found, only one with the best resolution was chosen. Finally we found 24 pairs of structures. The conformational changes values of the control set and 48 transient heterodimers are list in Supplementary Table S2 (a) and S2 (b) of Supporting information file S1, respectively.

Figure 3a shows a histogram of the C_a RMSDs of all residues. The average C_a RMSDs of all residues are 1.1 and 1.4 Å for the weak and strong transient heterodimers, respectively. There are 28 and 34 pairs having C_a RMSDs of all residues higher than the control limit for the two datasets, respectively. The largest C_a RMSD is 7.29 Å, observed in Thioredoxin reductase (1F6M) which consists of an FAD-binding domain and an NADPH-binding domain. The two domains are connected by two β-strands and loops that would undergo a large movement as they bind the substrate [37]. In this work, the domain motion of a protein was detected by the DynDom program [25]. Given two conformations of a protein, the DynDom program will analyze the conformational change in terms of dynamics domains, hinge axes, and hinge-bending regions. In the two datasets, 15 pairs were found to undergo a domain shift when binding the protein partner. The detailed analysis of these domain motions are shown in Table 2. In addition, there are 16 pairs where the C_a RMSD of all residues between unbound and bound protein partners is well over 2 Å in Fig. 3a and 8 of which are caused by the domain motion (Table 2). This result indicates that the significant domain motion may be an important factor for the large conformational changes in this work.

The histogram of C_a RMSDs of interface residues is shown in Fig. 3b. For the strong transient heterodimers, the C_a of interface residues tend to undergo larger conformational changes than the weak transient heterodimers. There are 26 pairs of strong transient heterodimers having C_a RMSDs of interface residues higher than the control limit. However, only 14 pairs of weak transient heterodimers move more than the control limit. The average C_a RMSDs of interface residues are 0.9 and 1.3 Å for the weak and strong transient heterodimers, respectively. In the Fig. 3c, 24 pairs of weak transient heterodimers and 32 pairs of strong transient heterodimers have C_a RMSDs of surface residues higher than the control limit, respectively. The result shows that significant movements take place in the C_a of surface residues upon interaction. Betts and Sternberg attributed the conformational changes of surface residues to the flexibility and disorder of these residues [38]. We believe that in addition to these factors, the significant domain motion may also affect the conformation of surface residues. The large domain motions in Table 2 are all observed to have high C_a RMSD values on surface residues.

The majority of protein complexes in the two datasets have small C_a RMSD between unbound and bound structures. Indeed, about 80 % of pairs used in this work have C_a RMSD < 2 Å for both interface residues and all residues. Thus, the conformational changes in the side-chain play an important role in PPIs in this work. Fig. 3b, d show that the side-chain RMSDs of interface residues are relatively large compared with the main chain RMSDs of interface residues. This result may further support above conclusion that conformational changes in the side-chain is more important for PPIs in this work. But it is worth noting that only C_a atom is used to represent main chain, however, all the side-chain atoms are used to represent side-chain. From this point of view, the side-chains have more atoms contributing to conformational changes than main chains. The number of pairs for strong transient heterodimers with side-chain RMSDs of interface residues higher than the control limit is more than the weak transient heterodimers in Fig. 3d. There are 24 and 34 pairs having side-chain RMSDs of interface residues higher than the control limit for the two datasets, respectively, which indicates that the strong transient heterodimers undergo greater conformational changes involving side-chain of interface residues

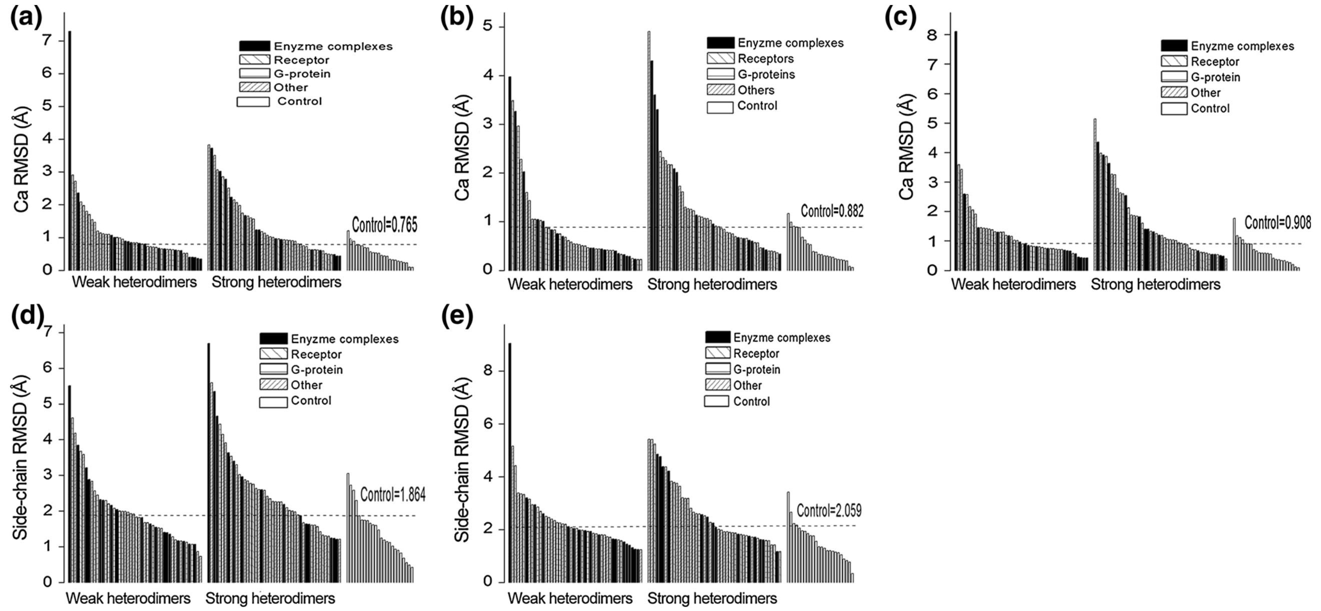

Fig. 3 RMSDs between the bound structures and unbound equivalents of the weak and strong heterodimers. The *dotted lines* show the values expected from experimental differences calculated from the corresponding residues of identical unbound proteins. **a** Ca RMSDs for all residues. **b** Ca RMSDs for interface residues. **c** Ca RMSDs for surface residues. **d** Side-chain RMSDs for interface residues. **e** Side-chain RMSDs for surface residues

Table 2 The domain motion analysis of weak and strong transient heterodimers

Bound	Unbound	Domain 1		Domain 2		RMSD of C_a atoms			Rotation angle	Types
		Size	C_a RMSD	Size	C_a RMSD	All	Surface	Interfaces		
2AQ3_B	1CK1_A	59	0.62	23	0.37	2.09	2.58	0.42	6	Enterotoxin
2FJU_A	1MH1_A	146	0.58	21	0.47	0.70	0.83	0.51	10.1	Rac GTPase
1KLU_A	1H15_AB	144	0.45	31	0.39	0.71	0.91	0.41	7.4	Transferase
1F6M_A	1CL0_A	180	0.43	132	1.12	7.29	8.10	3.98	65.7	Thioredoxin reductase
1E4K_C	1FNL_A	84	1.12	82	0.88	1.71	2.06	0.58	13.3	Immune system receptors
1E4K_B	2DTQ_B	102	0.61	100	2.12	2.91	3.59	3.49	16.8	Immune system
2MTA_H	2BBK_J	318	0.29	22	0.37	0.39	0.57	0.23	6.7	Methylamine dehydrogenase
1ZM4_A	1N0V_C	468	0.62	74	0.51	2.36	2.60	3.27	13.5	Elongation factor 2
1XD3_B	1YJ1_A	36	0.50	25	0.28	0.80	0.53	1.30	5.2	Ubiquitin
1JMO_H	2CN0_HL	198	0.36	20	0.46	0.43	0.51	0.62	5.5	Thrombin
1JMO_A	1JMJ_A	353	1.05	21	6.01	3.73	4.36	4.31	48.5	Heparin cofactor
2HQS_H	1OAP_A	79	0.37	24	0.57	0.59	0.39	0.66	7.5	Lipoprotein
2HQS_A	1CRZ_A	246	1.65	147	0.82	2.51	3.25	1.10	15.1	Toxin binding protein
1IJK_B	1FVU_A	101	0.5	28	0.42	0.97	1.19	0.96	6.9	Toxin
2AJF_E	2GHV_E	150	2.2	20	2.57	2.85	3.87	0.86	30.1	Receptor binding domain

than weak transient heterodimers. In the Fig. 3e, 23 pairs of weak transient heterodimers and 25 pairs of strong transient heterodimers have side-chain RMSDs of surface residues higher than the control limit, respectively. The number of surface residues with side-chain RMSDs higher than the control limit is slightly less than that of interface residues, which indicates that the side-chain of interface residues may move more than the side-chain of surface residues in this work. The reason may be that the changes in the side-chain of interface residues occur to form interactions, improve geometric complementarity or to avoid steric clash [39].

The extent of conformational change of different functional classes would also be different. Figure 3 shows that large movement is more common in the complexes involving enzymes or receptors than those involving G-proteins. Actually, most of enzymes or receptors undergo domain motions during ligand binding. As demonstrated above, substrate binding affects the relative orientation of two domains in thioredoxin reductase. The receptors' extracellular domains may also be affected by the interaction with ligands. For example, the domain rotations take place in the extracellular region of Eph receptor when it binds ephrins [40]. However, the G-proteins undergo secondary structure change or disorder-to-order transitions when they interact with other proteins [41].

Protein–protein binding affinity prediction

For protein complexes, the binding affinity translates in physical–chemical terms into an equilibrium dissociation constant (K_d), which can be measured at equilibrium or derived from the reaction kinetics, and the related Gibbs free energy of dissociation ΔG [42]. Nowadays, the experimental methods such as surface plasmon resonance (SPR), isothermal calorimetry (ITC), and titration by fluorescence or other spectroscopic methods all can be used to determine the K_d [15]. Since it is too time-consuming and expensive to experimentally perform protein–protein binding assay, developing reliable computational methods to predict protein–protein binding affinity is fundamentally important. The interactions of surface residues of two proteins decide the strength of this interaction, it is probable to predict the binding affinity based on the functional features derived from our analysis. Here, we employed a support vector regression (SVR) model to predict the binding affinity of transient heterodimers complexes. In order to evaluate the performance of the SVR model, and to compare with other regression methods, we have used other seven regression methods: partial least squares (PLS), Gaussian process (GP), multiple linear regression (MLR), multivariate adaptive regression splines (MARS), RF, radial basis function interpolation (RBF) and M5′ Regression Tree (M5′) [43]. The results are given in Table 3. As can be seen, the non-linear algorithms: SVR, MARS, RF, RBF, M5′ and GP performed better than the linear algorithms, PLS and MLR. The linear algorithms may be sensitivity to outliers, inability to account non-linear relationships and degradation in performance in high dimensions. However, the non-linear algorithms which are able to avoid fitting noise in the data while still detecting the signal may be better suited to the interaction affinity prediction. In the non-linear algorithms, SVR performed best as compared to others. The approach proposed by Moal et al. [43] is to construct a set of 200 molecular descriptors and feed these descriptors into the four machine learning algorithms. The 200 molecular descriptors contained redundant features which may be noisy in the models. The SVR exhibited an acceptable goodness-of-fit for the K_d of 48 heterodimers, with $R = 0.79$ and RMSE = 0.78. Meanwhile, the correlation plot of experimentally determined affinities versus SVR model-derived values is shown in Fig. 4a. According to the definition of Perkins et al. [8], the weak transient PPIs usually give a fast bound–unbound equilibrium with the dissociation constant (K_d) values of 1.0×10^{-6} or higher and the strong ones are of K_d values between 10^{-9} and 10^{-6}. If using 1.0×10^{-6} as the threshold to classify the weak and strong PPIs, almost all heterodimers are correctly predicted, except the thioredoxin in complex with thioredoxin reductase (1F6 M). It was wrongly classified as strong, which may be caused by the largest conformational changes in this complex.

Table 3 Comparison of the performances of different regression models

	R	RMSE
Support vector regression	0.79	0.78
Random forest	0.65	2.19
Radial basis function interpolation	0.60	1.91
Multivariate adaptive regression splines	0.60	2.63
M5′ regression tree	0.59	2.16
Gaussian process	0.56	0.99
Multiple linear regression	0.44	1.12
Partial least squares	0.34	1.61

The SVR model was further tested on an independent external test set. It is well known that the high fitting ability of the developed model does not guarantee its real-world predictive power, and the external validation is the only way to establish a reliable predictor. Therefore, the built SVR model based on the 48 heterodimers were further used to carry out prediction on 20 heterodimers not used for training. The SVR model predicts binding affinity for test heterodimers with $R = 0.84$ and RMSE = 0.6 in Fig. 4b, indicating the stability of our model. All predictive values for the training set and test set are shown in Supplementary Table S3 (a) and S3 (b) of Supporting information file S1, respectively. Overall, the result shows that the functional features proposed in this work are not only able to classify the types of PPIs, but also can characterize the protein–protein binding affinity. In addition, this work may have highlighted the need to take into account the type of PPIs when predicting the protein–protein binding affinity.

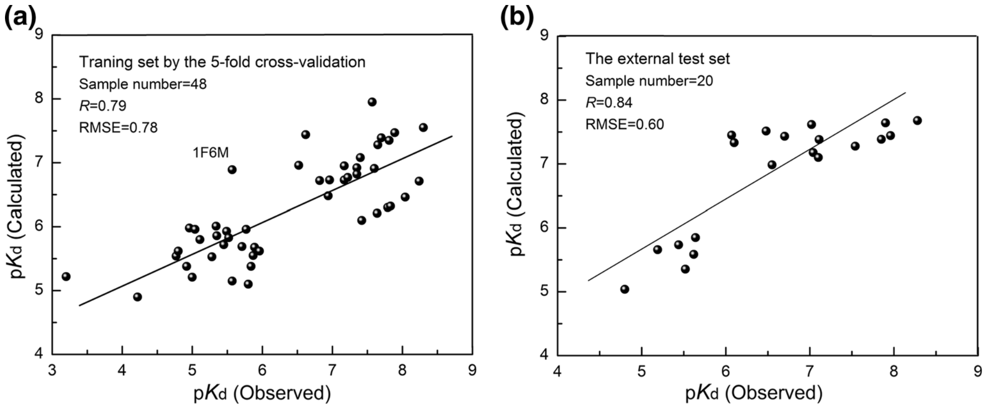

Fig. 4 Scatter plot of calculated versus experimental binding affinities for 48 heterodimers in training set (**a**) and for 20 heterodimers in test set (**b**) by using SVR

Conclusions

We have presented a functional feature analysis on five types of PPIs. The analysis results show that there is a clear trend for strong interactions to have larger, better packed and less planar interfaces than the weak interactions, and they also have larger interface movements involving main chains as well as side chains upon interaction. In addition, we have observed that the large core area fraction has very high occurrence in specific interactions. This indicates that the core area fraction appears to be fundamental determinants of biological interfaces. Finally, a combination of the functional features derived from our analysis was used to predict the K_d values and the prediction results demonstrate that these features are useful to estimate the protein–protein binding affinity and may play an important role in understanding the mechanism of PPIs.

Acknowledgments This work was funded by the National Natural Science Foundation of China (No. 20905054 and 21175095).

References

1. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2004) A dissection of specific and non-specific protein–protein interfaces. J Mol Biol 336:943–955
2. Bernauer J, Bahadur RP, Rodier F, Janin J, Poupon A (2008) DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein–protein interactions. Bioinformatics 24:652–658
3. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372:774–797
4. Ponstingl H, Henrick K, Thornton JM (2000) Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 41:47–57
5. Zhu H, Domingues FS, Sommer I, Lengauer T (2006) Noxclass: prediction of protein–protein interaction types. BMC Bioinformatics 7:27
6. Nooren IM, Thornton JM (2003) Diversity of protein–protein interactions. EMBO J 22:3486–3492
7. Ozbabacan SE, Engin HB, Gursoy A, Keskin O (2011) Transient protein–protein interactions. Protein Eng 24:635–648
8. Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C (2010) Transient protein–protein interactions: structural, functional, and network properties. Structure 18:1233–1243
9. Jones S, Thornton JM (1996) Principles of protein–protein interactions. Proc Natl Acad Sci USA 93:13–20
10. Nooren IM, Thornton JM (2003) Structural characterisation and functional significance of transient protein–protein interactions. J Mol Biol 325:991–1018
11. De S, Krishnadev O, Srinivasan N, Rekha N (2005) Interaction preferences across protein–protein interfaces of obligatory and non-obligatory components are different. BMC Struct Biol 5:15
12. Guharoy M, Chakrabarti P (2007) Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein–protein interactions. Bioinformatics 23:1909–1918
13. Dey S, Pal A, Chakrabarti P, Janin J (2010) The subunit interfaces of weakly associated homodimeric proteins. J Mol Biol 398:146–160
14. La D, Kong M, Hoffman W, Choi YI, Kihara D (2013) Predicting permanent and transient protein–protein interfaces. Proteins 81:805–818
15. Kastritis PL, Moal IH, Hwang H, Weng Z, Bates PA, Bonvin AM, Janin J (2011) A structure-based benchmark for protein–protein binding affinity. Protein Sci 20:482–491
16. Wang R, Fang X, Lu Y, Wang S (2004) The PDBbind database: collection of binding affinities for protein–ligand complexes with three-dimensional structures. J Med Chem 47:2977–2980
17. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2003) Dissecting subunit interfaces in homodimeric proteins. Proteins 53:708–719
18. Hubbard SJ, Thornton JM (1993) NACCESS: computer program. University College London, London
19. Ruvinsky AM, Kirys T, Tuzikov AV, Vakser IA (2011) Side-chain conformational changes upon protein–protein association. J Mol Biol 408:356–365
20. Kortemme T, Baker D (2002) A simple physical model for binding energy hot spots in protein–protein complexes. Proc Natl Acad Sci USA 99:14116–14121
21. Saha RP, Bahadur RP, Pal A, Mandal S, Chakrabarti P (2006) ProFace: a server for the analysis of the physicochemical features of protein–protein interfaces. BMC Struct Biol 6:11
22. Gutteridge A, Thornton J (2004) Conformational change in substrate binding, catalysis and product release: an open and shut case? FEBS Lett 567:67–73

23. Gutteridge A, Thornton J (2005) Conformational changes observed in enzyme crystal structures upon substrate binding. J Mol Biol 346:21–28
24. Echols N, Milburn D, Gerstein M (2003) MolMovDB: analysis and visualization of conformational change and structural flexibility. Nucleic Acids Res 31:478–482
25. Hayward S, Lee RA (2002) Improvements in the analysis of domain motions in proteins from conformational change: DynDom version 1.50. J Mol Graph Model 21:181–183
26. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc Chem Res 33:889–897
27. Melo F, Feytmans E (1997) Novel knowledge-based mean force potential at atomic level. J Mol Biol 267:207–222
28. Su Y, Zhou A, Xia X, Li W, Sun Z (2009) Quantitative prediction of protein–protein binding affinity with a potential of mean force considering volume correction. Protein Sci 18:2550–2558
29. Böhm HJ (1998) Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. J Comput Aided Mol Des 12:309–323
30. Englebienne P, Moitessier N (2009) Docking ligands into flexible and solvated macromolecules. 4. Are popular scoring functions accurate for this class of proteins? J Chem Inf Model 49:1568–1580
31. Oda A, Tsuchida K, Takakura T, Yamaotsu N, Hirono S (2006) Comparison of consensus scoring strategies for evaluating computational models of protein–ligand complexes. J Chem Inf Model 46:380–391
32. Ballester PJ, Mitchell JB (2010) A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics 26:1169–1175
33. Li XL, Zhu M, Li XL, Wang HQ, Wang SL (2012) Protein–protein binding affinity prediction based on an SVR ensemble. Intell Comput Technol 7389:145–151
34. Janin J, Rodier F (1995) Protein–protein interaction at crystal contacts. Proteins 23:580–587
35. Lo Conte L, Chothia C, Janin J (1999) The atomic structure of protein–protein recognition sites. J Mol Biol 285:2177–2198
36. Laskowski RA (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 13:323–330
37. Obiero J, Pittet V, Bonderoff SA, Sanders DA (2010) Thioredoxin system from *Deinococcus radiodurans*. J Bacteriol 192:494–501
38. Betts MJ, Sternberg MJ (1999) An analysis of conformational changes on protein–protein association: implications for predictive docking. Protein Eng 12:271–283
39. Janin J, Chothia C (1990) The structure of protein–protein recognition sites. J Biol Chem 265:16027–16030
40. Chrencik JE, Brooun A, Recht MI, Kraus ML, Koolpe M, Kolatkar AR, Bruce RH, Martiny-Baron G, Widmer H, Pasquale EB, Kuhn P (2006) Structure and thermodynamic characterization of the EphB4/Ephrin-B2 antagonist peptide complex reveals the determinants for receptor specificity. Structure 14:321–330
41. Grant BJ, Gorfe AA, McCammon JA (2010) Large conformational changes in proteins: signaling and other functions. Curr Opin Struct Biol 20:142–147
42. Zhou P, Wang C, Tian F, Ren Y, Yang C, Huang J (2013) Biomacromolecular quantitative structure–activity relationship (BioQSAR): a proof-of-concept study on the modeling, prediction and interpretation of protein–protein binding affinity. J Comput Aided Mol Des 27:67–78
43. Moal IH, Agius R, Bates PA (2011) Protein–protein binding affinity prediction on a diverse set of structures. Bioinformatics 27:3002–3009