A structural database of chain-chain and domain-domain interfaces of proteins

Neeladri Sen[1,2] and M.S. Madhusudhan[1,*]

[1]Indian Institute of Science Education and Research, Pune, India 411008

[2]Institute of Structural and Molecular Biology, University College London, London WC1E 6BT, UK

[*] To whom the correspondence should be made

madhusudhan@iiserpune.ac.in

Abstract

In this study, we mined the PDB and created a structural library of 178,465 interfaces that mediate protein-protein/domain-domain interactions. Interfaces involving the same CATH fold(s) were clustered together. Our analysis of the library reveals similarities between chain-chain and domain-domain interactions. The library also illustrates how a single protein fold can interact with multiple folds using similar interfaces. The library is hence a useful resource to study the types of interactions between protein folds. Analysing the data in the library reveals various interesting aspects of protein-protein and domain-domain interactions such as how proteins belonging to folds that interact with many other folds also have high EC values. These data could be utilized to seek potential binding partners. It can also be utilized to investigate the different ways in which two or more folds interact with one another structurally. We constructed a statistical potential of pair preferences of amino acids across the interface for chain-chain and domain-domain interactions separately. They are quite similar further lending credence to the notion that domain-domain interfaces could be used to study chain-chain interactions. We analysed protein complexes modelled by AlphaFold2 and RoseTTAFold and noticed that some of the modes of interaction involve folds and interfaces that have not been observed to bind in the PDB. Lastly and importantly, the library includes predicted small molecule binding sites at protein-protein interfaces. This has applications as interfaces containing small molecule binding sites can be easily targeted to prevent the interaction and perhaps form a part of a therapeutic strategy.

Keywords

protein chain, protein domain, interaction interface, ligand binding site, protein fold

1. Introduction

Protein-protein interactions are vital for several biological processes [1,2]. Identifying and characterizing such interactions could help explain the functioning of proteins and the basis of various diseases [3–5]. Various databases such as Database of Interacting Proteins (DIP) [6], Biomolecular Interaction Network Database (BIND) [7,8], Molecular Interaction Database (MINT) [9], Interactome3D [10], STRING [11] etc. list experimentally validated interactions. Among these databases, only very few such as Interactome3D contain structural information about the interacting partners. Various other databases (such as ProtCID [12], PIBASE [13] etc which we have discussed later) contain only information about protein complexes in the Protein DataBank (PDB).

The 3D structure of complexes of interacting proteins helps explain the mechanism of interaction, which in turn shed light on the functioning of cellular pathways [14–17]. 3D structures of these complexes can be determined by X-ray crystallography, NMR spectroscopy, cryo-EM, etc. Though the number of 3D structures of heteromeric protein complexes is steadily increasing, these experiments are expensive, labour intensive and technically challenging [18,19], necessitating computational approaches [19–25]. With the advent of high accuracy deep learning driven methods such as AlphaFold2 [26] and RoseTTAFold [27] for predicting the 3D structures of proteins, the next frontier is to make accurate predictions of the structures of heteromeric protein complexes. One of the key ingredients of the success of these prediction methods was the availability of over 100,000 structures and over 200 million sequences of proteins. There are however far fewer heteromeric protein complexes in the PDB (around 42,000). The same deep learning techniques have been adapted for the prediction of the structures of protein complexes [28,29]. However, to improve these techniques, we would need to either wait for more data on 3D structures of complexes to accumulate or leverage information from known structures, paying particular attention to domain-domain interfaces. Our study, which primarily focuses on binary interactions, is an effort in the latter direction. An argument can be made that protein cores and protein interfaces show similar physico-chemical properties (such as amino acid composition, contact preferences etc) and so monomeric structures could help build structures of complexes. However, cores and interfaces differ in structural packing and composition [30]. Hence the vast repertoire of protein structures that are not in complexes cannot be used to model protein complexes and we need to rely on data from protein interfaces.

Large proteins contain multiple domains, which are defined as independent folding, evolving and structural units in proteins. Two protein chains can fuse (gene fusion) leading to the formation of two protein domains in a single protein. Conversely, two (or more) domains of the same protein can split and evolve into two (or more) independent chains [31,32]. During these fusion/fission events, a chain-chain interface can convert to a domain-domain interface or vice versa. Hence interfaces on domains can be structurally similar to that between chains. The domain definitions/boundaries of individual proteins

have been characterized in SCOP/SCOPe [33,34], CATH [35–37], Pfam [38] and Ecod [39] which can be utilized to identify interfaces between different protein domains.

Multiple libraries have been developed in the past to characterize protein-protein/domain-domain interfaces such as 3DID [40–42], PIBASE [13], SCOPPI [43], SNAPPI-DB [44], SCOWLP [45,46], ProtCID [12], PrePPI [47] etc. Many of these databases classify and cluster the interfaces to show similarities between different interfaces or study specific properties of the interfaces such as conservation, the importance of water etc. Methods such as PRISM [20] and InterComp [48] utilizes the interfaces as templates to model protein complexes [49]. Databases such as PrePPI [47] in addition to experimentally determined structures also contain predicted protein complexes.

Protein complexes can be modeled either by docking one protein onto another or by comparative modeling using a template protein complex. Template-based modeling of protein complexes has been shown to be more accurate in comparison to docking [50–52]. In addition, recent literature indicates that the structural repertoire of protein interfaces is degenerate and close to complete [53,54] and nature reuses similar interfaces across different proteins. Hence a library of such observed protein-protein/domain-domain interfaces will be useful in understanding and modeling protein complexes. A combined domain-domain and chain-chain interface library might be useful and may account for gene fusion/fission events. Hence, a composite library might provide a better sampling of the structural space of the protein interfaces.

We have created a library of all known interfaces between different proteins (chain-chain interface) and also separately catalogued intra-chain domain-domain interfaces. Libraries as such can be useful in studying how protein folds structurally interact with one another. Thus, we structurally clustered interfaces belonging to the same fold, to identify the various modes of interactions between proteins belonging to the same fold. Using the interface library, we showed how domain-domain and chain-chain interfaces and non-homologous protein complexes (belonging to the same/different folds) could have structurally similar interfaces. In addition to structurally characterizing the interfaces, we compared the amino-acid pair preferences between domain-domain and chain-chain interfaces. We also showed that the interfaces are more conserved as compared to the whole protein. The library also contains predicted small molecule binding sites that could be targeted to prevent protein complex formation, with possible therapeutic applications. We have also compared modelled protein complexes to our interface library to identify interactions not observed in the PDB. Regardless of how proteins interact, be it homologous pairs interacting similarly [55,56] or unrelated protein pairs (or proteins and peptides) using similar interfaces [48] [49], our interface library could be used to predict the 3D structures of the complexes.

2. Results

2.1. Library of interfaces

We extracted 112,043 binary chain-chain interfaces and 66,442 binary domain-domain interfaces (domain definitions based on CATHv4.2) from 42,259 PDB structures (table 1). As mentioned in the methods section, we call these fold combinations. Note that only 89,993 out of 112,043 fold combinations have an associated CATH identifier. This has implications on how these fold combinations are clustered (see section 2.3). The 22,110-fold combinations that are not annotated by CATH ids come from 6712 PDB entries. These interfaces are still a part of the library, albeit without being clustered (see section 2.3).

Binary interfaces are made of both homo- and heteromeric interactions. Among chain-chain interfaces, 62% of fold combinations were homomers, whereas this was only 28% of the domain-domain interfaces. Homomeric associations are more abundant in interactions between protein chains than interactions between protein domains.

2.2. Fold related properties of the interface library

2.2.1. Fold Combination

The binary fold combinations in our library contain records of CATH ids interacting with one another. There are 524 CATH ids that interact with only one other fold and on the other extreme, the CATH id 3.40.50 (Rossman fold) interacts with 238 other folds. The number of fold combinations that a CATH id is a part of has a correlation coefficient of 0.82 with the number of its Enzyme Commission (EC) terms (Figure S1). Generally, the higher the number of EC terms associated with a fold, the greater the number of folds it would interact with. Of the 1391 CATH folds, only 1109 folds interact with one another. Of these, 1092 folds interact with <30 other folds. The Rossman fold (CATH ID- 3.40.50) has the highest number of fold combinations, 238, to go along with over 1,000 EC terms (Table2). CATH id 1.20.5 (Single alpha-helices interacting in coiled-coils or other helix-helix interfaces) interacts with 90 different folds, perhaps because of the diversity of the sequences that can take up this particular fold [57]. The list of 15 folds that are a part of >30 fold combinations (table 2) contains some of the most abundantly populated CATH ids. The fact that some folds such as Rossmann fold, TIM barrels and Jelly rolls have more EC terms (1087, 317 and 132 respectively) than they appear in fold combinations (238, 66 and 57 respectively) could imply that several fold combinations are yet to structurally explored. The higher the number of EC terms in a fold, the more varied the functions of the proteins in the fold, hence the larger the number of folds they interact with.

2.2.2. Number of interfaces per fold combination

The 1109 folds that form chain-chain and domain-domain interfaces feature in 3,065 unique fold combinations. 585 fold combinations have only one known interface (1 unique chain-chain or domain-domain interface in 1 fold combination). 2,502 (82%) fold combinations had <=30 interfaces (Figure S2). An interface here refers to a particular configuration of a fold combination. A single fold combination could be observed in the PDB to have multiple interfaces. A homo-oligomeric structure could have various instances of the same interface within a single PDB entry.
Homo-dimeric interactions between Rossmann folds (14,844 interfaces), Immunoglobulin-like folds (6,528 interfaces) and Glutamine Phosphoribosylpyrophosphate-subunit 1-domain 1 folds (6,318 interfaces) are the three largest populations of fold combinations. As was shown earlier (Table 2), the Rossmann fold and the Immunoglobulin-like fold are associated with 1,087 and 111 EC Terms respectively. In addition, the Rossman Fold is the most diversified and prevalent fold of ancient evolutionary origin and accounts for about 15% of the human proteome [58]. A large number of instances (6,318) of Glutamine Phosphoribosylpyrophosphate-subunit-1-domain-1 fold homo-dimers are found in only 277 PDB entries, as most/all of these entries contain homo-multimers of the same protein. This fold has only 14 EC terms. This indicates the low diversity in the sequences and functions of proteins adopting these folds.

2.3. Clustering of the interface library

Given the number of instances the same fold combinations appear in different interfaces, we clustered (within each fold combination) the 156,375 interfaces ((89,933 chain-chain and 66,442 domain-domain) according to structural similarity. To find general structural patterns in the ways folds and domains interact with one another, we grouped them into structural clusters within each fold combination. This clustering can also help identify if non-homologous protein pairs can use the same interface geometry and if chain-chain and domain-domain interfaces are structurally similar. The cluster representatives can also serve as structurally non-redundant templates for the comparative modelling of protein complexes or multidomain proteins. The interfaces were grouped into 27,885 clusters that have structure overlap >=80% and RMSD <=1.5 Å with respect to that of the representative PDB. 13,039 (~47%) clusters only contain 1 PDB (Figure S3), which might be a result of the stringent RMSD and structure overlap criterion used during clustering. 26,401 (~95%) clusters contain less than 20 PDB per cluster (Figure S3).

Superfamily based investigation of interface clusters
Each CATH fold is further subdivided into homologous superfamily. We checked if the structural clusters formed as described above were within a superfamily. Out of the 27,885 clusters, 27,741 clusters (99.5%) had all interfaces belonging to the same

superfamily combination. Only 144 clusters had interfaces belonging to different superfamily combinations. Out of these 144 clusters, 130 clusters had interfaces from a combination of 2 superfamilies. Of the remaining 14 clusters having more than 2 superfamily combination in the same cluster, 9 clusters belonged to the fold 1.20.5 (Single alpha helix involved in coiled-coil or other helix-helix interactions). This is because the single alpha helix fold is a secondary structural component that could easily interact with other helices from various superfamilies, leading to the clustering together of multiple superfamilies. Even though our study was done at a fold level, 99.5% of the clusters were formed superfamily wise, indicating that interfaces are structurally conserved within superfamilies. Had we used less stringent criteria as cut-offs for clustering, we might have had more clusters across superfamilies.

2.3.1. Number of interface clusters per fold combinations

The number of clusters per fold combination ranges from 1 to 8,004. 1,626 fold combinations (out of the 3,065 observed combinations) had only 1 cluster (585 fold combinations had only 1 interface and hence only had 1 cluster). 2,942 (96%) fold combinations had <20 clusters (Figure S4). 3043 (99%) fold combinations were clustered into less than 100 clusters.

The top few folds having a large of self-interactions are those involving the Rossmann fold, immunoglobulin-like fold and TIM barrel, with 14,844, 6,528 and 3,598 interfaces respectively. These interfaces were grouped into 7,809, 1,675 and 984 clusters respectively. The high number of clusters in these fold combinations was because multiple non-homologous proteins take up these folds [59] (as indicated by high EC numbers for the fold - Table 2). This results in multiple modes of interactions depending on the protein type. Most of the interfaces (73%) of self-interactions between Rossman folds have high RMSD (>2 Å) and low structure overlap (<70%) when the different protein interfaces are structurally superimposed on the cluster representatives (Figure S5). However, self-interactions in domains such as Glutamine Phosphoribosylpyrophosphate-subunit 1-domain 1 folds containing 6,318 interfaces, have only 202 clusters because of the low number of EC terms (14) for the CATH fold, indicating low diversity of the sequences taking up the fold.

In certain cases, we have split PDB entries belonging to the same fold combination into multiple clusters because of the stringency in our clustering criterion. Certain folds such as the hemagglutinin-ectodomain chain B had 484 out of 567 interfaces clustered together with cluster representative 4gxx_BD. The other interfaces were grouped into different clusters because they had a structural overlap ranging between 70-80% and RMSD between 1.5 Å – 2.4 Å with 4gxx_BD and hence were not clustered together. It is clear that the stringency of clustering modulates the number of clusters. In addition, these stringent clusters can help in modeling protein interfaces by providing better resolved clusters (Figure S6).

2.3.2. Similarity between domain-domain and chain-chain interfaces

Over the course of evolution, protein domains can split into chains or protein chains can come together to form domains. The interface library could be used to find evidence of structurally similar chain-chain and domain-domain interfaces. Hence, a composite library such as the one described here can serve as a source of templates to model both protein complexes and multidomain proteins.

514 fold combinations contain both chain-chain and domain-domain interfaces. From these fold combinations, 102 clusters contain both chain-chain and domain-domain interfaces clustered together (Supplementary Text 1). These clusters had a median of 57% of domain-domain interface and 43% chain-chain interface. Clusters, as such, highlight the fact that nature reuses the same geometry across different types of interfaces.

Here are a few interesting examples that illustrate the usage of the same interface in domain-domain and chain-chain interactions, even when the proteins with those folds/domains are unrelated to one another. The domain-domain interface of Giardia dicer is superimposed onto a chain-chain interface of the Nuclease domain of ribonuclease 3; with a structure overlap of 86% and an RMSD of 1.33 Å (Figure S7A). The two proteins share a sequence identity of 23% and belong to the Ribonuclease iii N terminal endonuclease domain, Chain A fold. In the other example, the chain-chain interface of AVA_4353 protein superimposes onto the domain-domain interface of PhuS protein with a structure overlap of 91% and RMSD of 1.28 Å (Figure S7B). The two proteins belong to the heme utilizing iron like fold and share no significant sequence similarity. Curiously, within a complex of a protein carboxysome shell protein CcmP (with two domains of the same fold Alpha Beta Plaits), we see structurally similar chain-chain and domain-domain interface (Figure 1). The two domains within a monomer however share no significant sequence similarity.

2.3.3. Sequence conservation at the interface compared to that of the whole protein/domain

In this section, we are comparing the conservation of residues on the interface to the conservation of residues throughout the protein/domain. The sequence identity of the interface was computed as the number of identical residues between the structurally aligned positions of the cluster representative and the members of the cluster. The structural comparison was done using the structural alignment tool CLICK [60]. The sequence identity of the full protein/domain was computed as the number of identical residues computed from a BLAST2seq alignment [61] of the cluster representative with the members of the cluster. Structural alignments were not used while computing the identity of the full protein as the proteins might be structurally dissimilar even though they have structurally similar interfaces. In the case of chain-chain interfaces, the full protein

was used for the alignment, however for domain-domain interfaces, only the domains under consideration were used. In instances where there are multiple copies of the same interaction in a single PDB, such as homo-multimeric complexes, only 1 of the interfaces was considered for the analysis of sequence conservation, to avoid redundancy.
Our analysis shows that in a majority of cases (54%), the interface residues were more conserved than the entire protein chains within a cluster. In 23% the protein chains showed higher conservation than the interfaces. In the remaining 23% of cases the conservation in the protein chain and the interface were about the same (Figure S8). We investigated the extreme cases where the difference in the identities at the interface and the full protein sequence was > 30%. Most of these were a consequence of non-significant short sequence alignments, non-rigidity of the interface, structurally similar chain-chain and domain-domain interfaces in the same protein and structurally similar chains in heterooligomer (Supplementary Text 2).

Clustering of sequentially unrelated interfaces

The interface library shows how some sequentially unrelated/distantly related proteins could have similar interface structures. This library, can in turn, be useful to identify templates (based on structural similarity) to model protein-protein complexes or build structures of multidomain proteins. Around 2% of the interfaces clustered together have an identity of <30% while 10% of them have an identity of <40%. L2- Haloacid dehalogenase from X. autotrophicus (Figure S9A) (PDB ID – 1QQ5) and the hypothetical 2 haloalonoic acid dehalogenase S. tokodaii (Figure S9A) (PDB ID – 2w43) contains the 1.10.150-3.40.50 fold combination. The two proteins are 31% identical to each other. However, the two interfaces superimpose on each other with a structure overlap of 89% and RMSD of 1.47 Å. The Human Psoriasin (Figure S9B) and the Bovine protein SC0067 contain the 1.10.238 fold but are only 27% identical to each other. They have a similar structure and interact using the same geometry with the interface structure overlap of 94.6% and RMSD of 1.0 Å.

2.4. Structurally similar protein-protein interfaces from different folds

A fold interacting with different folds using the same geometry

The clustering of the interface library was limited to proteins belonging to the same fold combinations, as an all against all comparison of all the interfaces irrespective of their folds is computationally expensive and out of the scope of our computational resources. However, we compared a few interfaces across different folds to check if there exists structural similarity of the interface irrespective of the fold the protein chains/domain belong to.

The NusG (Transcription antitermination protein) and the Transcription elongation factor SPT5 belong to the same fold of alpha-beta plaits and are 34% sequentially identical to each other. The NusG protein interacts with DNA dependent RNA polymerase E, which belongs to Ruberythrin Domain 2-fold whereas the SPT5 interacts with the Transcription elongation factor SPT4 belongs to Herpes Virus 1 fold. Even though, the interacting proteins (DNA dependent RNA polymerase and SPT4) to the two proteins (NusG and SPT5 respectively) belong to different folds and are only 29% identical sequentially the interacting interface is similar with a structure overlap of 93% and RMSD of 1.72 Å (Figure 2). We also found an example of interfaces belonging to different folds that share the same geometry which has been shown in Supplementary Text 3.
We believe that while we have not systematically categorized such similarities across fold combinations, these data could easily be mined from our library for individual fold combinations.

2.5. Pair preference of the amino acid residues at chain-chain vs domain-domain interfaces

The overall trends for the amino acid pair preferences at a chain-chain interface and domain-domain interface look similar (Figure S17). Here, favourable scores are positive numbers, and unfavourable scores are negative. The side chain interactions between Ala with all other amino acids are unfavourable (except Trp, Tyr, Phe which though favourable, have low scores). Another small amino acid, Cys, also has few favourable interactions. Except for the favourable Cys-Cys interaction (mostly disulphide bridges), its only other favourable interactions (for Met, Asn, Tyr, Trp and Phe) all have low scores. The interactions between the aromatic Tyr, Trp and Phe with all amino acids are favourable at both domain-domain and chain-chain interfaces. However, the interactions of other hydrophobic amino acids - Val, Ile and Leu are usually unfavourable (except for low favourable scores with Tyr, Trp and Phe). The interactions between negatively charged amino acids Asp and Glu and positively charged amino acids such as His, Lys, Arg, Gln are favourable, as should be expected. Pi-pi interactions (sp^2 containing side chains) are found among Tyr, Trp, Phe, His, Arg, Asp, Glu, Asn and Gln [62]. Most of these pairs show high favourability of interaction (potential of >1) except Phe-Asp, Phe-Glu, Asp-Asp, Asp-Glu, Glu-Glu pairs indicating the probable preference for pi-pi interactions at the interface. Self-pairs are preferred in chain-chain interfaces as compared to domain-domain interfaces. This can be because ~62% of the chain-chain interfaces are homo-oligomers as compared to ~28% of the domain-domain interfaces being homo-oligomeric (Results Section 2.2).

2.6. Small molecule binding site at protein-protein interfaces

Proteins function via interacting with other protein molecules. In disease conditions, these interfaces become important drug targets [63–65] and are invaluable in therapeutic discovery strategies. Hence, the prediction of small molecule binding sites at the protein-protein interfaces can be the first step toward inhibiting protein complex formation.
The overlap of the residues constituting the binding site and the interface was calculated (Figure S10). Of the 112,043 chain-chain interfaces, 61,367 and 74,849 interfaces had at least one chain with a minimum of 50% or 30% overlap respectively between the interface residues and the predicted small molecule binding site. If we consider both chains of the interface having at least 50% or 30% overlap with the binding site, the number of interfaces reduces to 30,525 and 61,667 respectively. If both the chains have been predicted as small molecule binding site, there is a higher probability that the interface could be targeted by small molecules. Depending on the stringency, the user can modulate the overlap percentage between the predicted binding site and interface. The predicted binding site (using DEPTH) could be used to dock/predict small molecules that could bind at the interface, hence disrupting the formation of the complex [66–69].
One example is that of the inhibition by a small molecule of the interaction between XIAP protein and caspase-9, which is a caspase involved in mitochondrial cell death [70] (Figure 3) [71]. The binding site on the Nipah virus glycoprotein had a 30% overlap with that of the interface residues with the ephrin B2 receptor of humans (PDB – 2VSM). Autodock [72] and DOCK [73] were used to predict the small drug-like molecule (ZINC63411510) that would go and bind the predicted binding site of the glycoprotein, hence preventing its interactions with the ephrin-B2 receptor [74].

2.7. Comparison of modelled protein complexes to the interface library

To explore the feasibility and ramification of using modelled/predicted structures, we used the 1106 core eukaryotic binary protein complexes modelled using AlphaFold2 and RoseTTAFold [75]. We ran these models by our interface library to check for new fold combinations. These would be binding modes not observed in the PDB, and hence not found in our database. We filtered these models to leave out those with disordered regions or regions with poor reliability scores at the interface (pLDDT score < 70) and ensured that the occluded surface area was 400 $Å^2$ or greater and contained at least 50 residues at the interface. This left us with 547 binary complexes. We next assigned CATH domains to the proteins in these complexes scanning their UniProt IDs against CATH-FunFams v4.3 (212,872 HMMs) using HMMER3 [76] with an e-value cut-off of 1e-03. Domain boundaries were resolved using CATH-resolve-hits [77] with a bit score cut-off of 25 and coverage of 80%. With this method, we could assign CATH domains to both chains in 308 of the 547 binary complexes. These 308 interfaces belonged to 233 fold

combinations, of which our database had records on 131. The remaining 102 fold combinations (from 105 interfaces) have not been seen in the PDB and were hence new binding modes identified by the deep learning complex modelling techniques. Even among the 131 known fold combinations, we detected new modes of binding. We clustered these fold combinations with the cluster representatives of the interfaces. Of the 203 interfaces (belonging to 131 fold combinations), only 53 interfaces had a structure overlap of >80% and RMSD <1.5 Å. However, given these are models and there can be regions of the interface which are not modelled well, the structural comparison may be inexact. Hence, we reduced the structure overlap cut-off to >70% and RMSD to <3 Å. With this more permissive comparison criteria, 174 interfaces matched one of the cluster representatives (Figure S19). Only 29 interfaces (belonging to 26 fold combinations) had a lower structure overlap indicating modes of interactions not seen in the PDB.

3. Discussions

Typically, interface libraries contain known associations between fold types. We believe that these associations should be viewed at the level of domains, which could be thought of as a unit of evolutionary conservation. Ideally, we want to annotate interaction patches on domains, but we found that in some complexes where 2 chains interact with one another multiple domains (more than 2) are involved. In such cases, we are unable to determine if the constituent binary domain associations could interact in isolation. Hence, we have organized the data into chain-chain and domain-domain interfaces. Our interface library consists of 112,043 pairs of interacting protein chains and 66,442 pairs of interacting domains taken from crystal structures deposited in the PDB. In our study, an interface is simplistically detected if the solvent accessible surface area occluded on dimer formation is greater than 400 $Å^2$, an approach similar to one espoused by the PQS server [78]. It is possible that using such a simple criterion to identify interfaces may lead to false positives/negatives. One of the future improvements we can make to our database is to use a more nuanced method of detecting interfaces. Change in the occluded surface area could be used in conjunction with various other information such as solvation energy, dissociation entropy, interface packing, surface complementarity, interface hydrophobicity, pair frequency, covariation, conservation, amino acid composition, surface conservation vs interface conservation etc [79].

In our interface library, we have used an 8 Å cut-off to identify interface residues. We and others had previously shown that an 8 Å cut-off, though permissive, is best at depicting the first shell of residues at the interface for the development of statistical potential [80,81]. We assume the 8 Å cut-off provides structural context to the residues that are important in maintaining the interface structure. This somewhat permissive cut-off could also help account for water-mediated interactions at the interface.

Binary chain-chain interactions are typically dominated by homomeric interactions, where the interacting partners have the same CATH id. In our library, 62% of all chain-chain

pairs and 28% of domain-domain interactions are homomeric. We conjecture that the other 72% of domain-domain interactions provide us with plausible templates for different types of interactions, many of which have not yet been deposited in the PDB.

In addition to its utility as a plausible template library for modeling protein-protein interactions, our library provides us with several pieces of useful data. The PDB contains 1,391 different CATH folds. Of these, 1109 folds have interacting partners either with itself or another fold. 15 of these folds interact with over 30 other folds. The Rossmann fold, for instance, interacts with 238 other folds. In general, there is a strong correlation (correlation coefficient > 0.8) between the number of EC terms associated with a fold and the number of folds it interacts with. A mismatch between the number of EC terms related to a fold and its number of known interacting partners gives one the basis to search for plausible new interacting partners or EC terms.

One reason for a high number of interfaces for certain folds (folds such as Rossmann fold, Immunoglobulin fold, α helices fold has >1000 interfaces) could be because of the large number of non-homologous proteins that populate these folds (maybe because of convergent evolution). It could also be because of the homo-oligomeric nature of certain folds. Of the 282 folds for which we have no evidence (yet) of interaction with other folds. ~67% of them belong to the orthogonal bundle (CATH ID – 1.10), irregular architecture (CATH ID – 4.10), 2-layer sandwich (CATH ID – 3.10), alpha-beta complex (CATH ID – 3.90) and up-down bundle (CATH ID -1.20) architecture. Though we cannot conclude this based on the data, we speculate that proteins belonging to these superfamilies do not interact with proteins of other folds.

This library now gives us a platform for examining the nature of interactions between one-fold and its multiple partners. Do date and party hubs use different types of interfaces [82], how could we categorize these, etc. We also examined the extent to which domains and chains use the same interfaces.

Our library has catalogued 3065 fold combinations involving 1109 folds of a possible 1391 folds. Speculatively, even if we assumed that only these 1109 folds were capable of interactions, we could have as many as 614,386 (966,745 fold combinations involving 1391 folds) fold combinations. Clearly, there is a significant mismatch between what is possible and what has been observed, PDB sampling bias and under-representation of complex structures notwithstanding. The extent of the mismatch implies that not all fold combinations are observed in nature. A close examination and analysis of the domain-domain associations in our library may be useful in guiding the construction of interactomes/networks.

The 155,375 interfaces (88,933 chain-chain interfaces+66,442 domain-domain interfaces) with an assigned domain definition clustered into 27,885 clusters based on the structural similarity of the interface. 99.5% of these clusters contained interfaces belonging to the same superfamily within a fold. If the clustering conditions were to be relaxed, more of the interface clusters would contain members from different

superfamilies. This indicates that structurally similar interfaces will mostly be found within a superfamily. In turn, this could assist in identifying interfaces on proteins within a superfamily. These interfaces could be homo- or heteromeric. Some of the often-recurring folds such as the Rossmann fold, TIM barrel and Immunoglobulin folds have >900 clusters of homomeric and heteromeric interfaces. This indicates the richness (diversity of sequences taking on the fold) of the fold in the way it explores interaction diversity, which in turn explains the high correlation with EC terms.

102 clusters had both chain-chain and domain-domain interfaces together, irrespective of the sequence similarity. This can indicate gene fusion leading to the formation of a domain-domain interface from a chain-chain interface or gene splitting leading to the formation of a chain-chain interface from the domain-domain interface. An interesting example is that of the CcmP protein, which has structurally similar chain-chain and domain-domain interfaces. Because domain-domain interfaces and chain-chain interfaces are sometimes structurally similar, our library can provide an increased number of templates to model multi-domain proteins whose individual domains may have been crystallized separately.

Our database is a redundant set of all the interfaces recorded in the PDB. The dataset has been made structurally non-redundant by clustering interfaces and choosing one representative per cluster. We did not remove redundant chains/domains from our analysis to capture the different modes in which identical sequences would interact with one another because of the flexibility of the interfaces. This could improve the predictive abilities of the library to model protein complexes allowing it to sample even small variations in otherwise similar conformations.

We observed that the interfaces have higher sequence similarity as compared to that of the whole protein. The sequence similarity and structure similarity go hand and hand however exceptions are noted. We noticed that ~2% of the interfaces which were clustered together (same fold combination) were structurally similar (structure overlap>80% and RMSD<1.5) but were not sequentially similar (<30% identity). These show that irrespective of the sequence identity this library can be used to search templates for protein complex modeling.

As stated earlier, our primary intent in creating the library was to accumulate a large number of interface templates to model protein-protein interactions. Though we have not explicitly done so in this study, our interface library when combined with a structure comparison tool, such as CLICK [60], could help us see similarities between interfaces from unrelated folds. Previous studies have also reported how dissimilar folds can use the same geometry at the interface to interact with a certain protein fold. Our interface library also contains many instances of homologous proteins interacting with each other using structurally different interfaces, such as in lectins, bacterial chemotaxis proteins, ASPP proteins etc. Previous studies have also pointed toward proteins utilizing similar geometry at the interfaces conjecturing that the structural repertoire of interfaces is close

to complete [53]. Hence, the interface library presented in this study can serve as a useful resource to model protein complexes using a topology-independent structural match to identify template interfaces for the same.

We also computed the amino acid pair preference at chain-chain and domain-domain interfaces to calculate amino acid substitution scores. Overall, the trends of what amino acid pairs are favoured/unfavoured are similar in both the chain-chain/domain-domain interfaces. However self-amino acid pairs were generally favoured at the chain-chain interface to a greater degree when compared to domain-domain interfaces. This could be because ~62% of chain-chain interfaces are homo-oligomers whereas only ~28% of the domain-domain interfaces are homo-oligomers. The similarity between the two pair preferences shows that domain-domain interfaces could supplement chain-chain interfaces and aid in the study of protein-protein interactions.

A significant predictive aspect of our library is the detection of plausible small molecule binding sites on interfaces. About 54% of protein-protein interfaces had at least 50% overlap between the interface residues and predicted small molecule binding site residues. This could serve as a useful resource in studying/inhibiting interactions with possible therapeutic applications. With these data, we could possibly analyse the interface structures to determine the most appropriate small molecule that could affect a known interaction. These small molecules could variously be carbohydrates, cofactors, drugs etc. However, all these predicted binding sites may not be druggable. Various methods such as PockDrug [83], Bitenet [84], DrugPred [85], DeepDrug3D [86] etc. have been developed to predict if a biding pocket is druggable. They use binding site properties such as hydrophobicity, cavity volume, physicochemical properties, polarities, compactness, hydrogen bonding abilities, Voronoi tessellations, voxel properties etc. These tools can hence be used in conjunction with our prediction depending on the requirements of the user to identify if the predicted binding pockets are druggable.

We envisage that future contributions to our database could be from modelled structures, given the recent success of deep modelling techniques. With this in mind, we analysed 308 (233 fold combinations) eukaryotic protein complex interfaces modelled using AlphaFold2 and RoseTTAFold. Of these, there were 102 protein fold combinations that have not (yet) been recorded in the PDB. Even among the 131 fold combinations already recorded in our database, 26 fold combinations had new interaction modes, as gleaned from computing structure overlap and RMSD to cluster representatives. These models have not yet been included in our dataset as some work needs to be done on quantifying the accuracies of the modelled complexes. But the preliminary results are encouraging and our database could contribute to self-consistent learning in these techniques and benefit future modelling of protein complexes.

In this study, we have laid the foundation for future protein-protein and domain-domain interactions studies/predictions/design. The interfaces here could prove useful in constructing the structures of protein complexes or even building a whole protein structure

from individually solved domains. Our library could also be used in conjugation with fragment-based interface design algorithms such as nanohedra [87]. With the rapid growth of structures resolved using cryo-EM, a library such as ours could also prove useful in refining such structures. The information presented in this study would soon be available as a queryable relational database on a webserver (work in progress).

4. Methods

4.1. Library of interfaces

All multi-chain and multi-domain (based on CATHv4.2 domain definition) complexes were extracted from the PDB. The accessible surface area for the individual protein chains/domains and all possible binary protein chain/domain complexes were calculated using MODELLER [88]. Only binary complexes with greater than 400 $Å^2$ change in solvent accessible surface area after interface formation were retained. This cut-off was used to filter crystallographic artifacts from biologically relevant interfaces in line with the PQS server [78,80,89,90]. Despite this cut-off, some of the interfaces could be artifacts of crystallization. However, we believe that these could still serve as viable templates and the scoring scheme employed for modeling could discern between actual interactions and artifacts [91].

Our library contains a list of interacting amino acid residues. Interacting residues are those that have at least one atom within 8 Å of another atom from a different chain/domain. These interacting residues constitute the interacting interface. The interface library contains dimeric chain-chain or domain-domain interfaces (domain-domain interfaces were between residues of the same chain). Oligomeric interfaces, involving more than 2 chains, are represented in the database by the constituent dimeric interfaces (subject to the same selection described above). All interfaces are also labelled by the CATH folds of their constituent chains/domains and are referred to as a ‘fold combination’ in this study.

4.2. Clustering of interfaces

All interfaces with the same fold combination were hierarchically clustered together such that the representative interface is the one with the highest resolution. All interfaces within a cluster were compared to the representative interface using CLICK [92] (a topology independent structural superimposition tool) with C^{α} and C^{β} as representative atoms for superimposition. An interface was clustered with the representative interface if the structure overlap was >80% and RMSD was <1.5 Å (these values were empirically chosen). A new representative interface was chosen from the remaining interfaces and the same procedure was repeated to find plausible clusters for all interfaces (Figure S11).

21 fold combinations with more than 1000 instances in PDBs were broken into two smaller sets of a maximum of 600 interfaces each (empirically chosen to ensure that the smaller subset had at least 400 members). This reduced the number of structural comparisons for clustering.

4.3. Pair preference of amino acids at domain-domain and chain-chain interfaces

A residue-residue interaction profile was calculated for all side chain-side chain interactions using the same statistical potential as described in the method PIZSA [93,94]. The scoring scheme is the ratio of the observed probability to the expected probability of an interface residue pair. It is a statistical potential that computes the propensity of pairs of amino acids to occur across an interface. Interactions across the interface were computed as atomic contacts with a cut-off distance of 4 Å. The PIZSA potential also calibrates this propensity by the number of atomic contacts mediating the interaction between the pair of amino acids. To prevent the overrepresentation of certain sequences while computing the statistical potential, the PDB entries were culled using PISCES [95] such that the maximum sequence identity was 40%.

4.4 Prediction of small molecule binding sites

The small molecule binding site of the individual chains that form the protein-protein interfaces was predicted with the software DEPTH [96] using default options. DEPTH was earlier compared to other state of art binding site prediction software such as MetaPocket2.0 [97] and Concavity [98] and was shown to be better or at par with these methods [96].

5. Data availability

The interface library can be downloaded from here.

Supplementary Material Description

Interface_library_Supplementary_Information.pdf – File containing all the supplementary figures and texts.

Conflict of interest

None

Acknowledgement

MSM would like to acknowledge the Wellcome trust-DBT India alliance for senior fellowship. NS would like to acknowledge the CSIR-SPMF fellowship. The authors would like to thank COSPI lab members for their insightful discussions. We would like to thank Minh N Nguyen for their initial work towards interface library creation. We would like to thank Gulzar Singh for the initial work towards the development of the webserver for the interface library. We would like to thank Mukundan S. for their technical help.

Bibliography

1. Yamada T, Bork P: **Evolution of biomolecular networks lessons from metabolic and protein interactions**. *Nat Rev Mol Cell Biol* 2009, doi:10.1038/nrm2787.

2. Alberts B: **The cell as a collection of protein machines: Preparing the next generation of molecular biologists**. *Cell* 1998, doi:10.1016/S0092-8674(00)80922-8.

3. Ryan DP, Matthews JM: **Protein-protein interactions in human disease**. *Curr Opin Struct Biol* 2005, doi:10.1016/j.sbi.2005.06.001.

4. Kuzmanov U, Emili A: **Protein-protein interaction networks: Probing disease mechanisms using model systems**. *Genome Med* 2013, doi:10.1186/gm441.

5. Sen N, Anishchenko I, Bordin N, Sillitoe I, Velankar S, Baker D, Orengo C: **Characterizing and explaining the impact of disease-associated mutations in proteins without known structures or structural homologs**. *Brief Bioinform* 2022, doi:10.1093/BIB/BBAC187.

6. Xenarios I: **DIP: the Database of Interacting Proteins**. *Nucleic Acids Res* 2000, doi:10.1093/nar/28.1.289.

7. Isserlin R, El-Badrawi RA, Badery GD: **The biomolecular interaction network database in PSI-MI 2.5**. *Database* 2011, doi:10.1093/database/baq037.

8. Bader GD, Betel D, Hogue CWV: **BIND: The Biomolecular Interaction Network Database**. *Nucleic Acids Res* 2003, doi:10.1093/nar/gkg056.

9. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G: **MINT: The Molecular INTeraction database**. *Nucleic Acids Res* 2007, doi:10.1093/nar/gkl950.

10. Mosca R, Céol A, Aloy P: **Interactome3D: Adding structural details to protein networks**. *Nat Methods* 2013, doi:10.1038/nmeth.2289.

11. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, et al.: **The STRING database in 2011: Functional interaction networks of proteins, globally integrated and scored**. *Nucleic Acids Res* 2011, doi:10.1093/nar/gkq973.

12. Xu Q, Dunbrack RL: **The protein common interface database (ProtCID)-A comprehensive database of interactions of homologous proteins in multiple crystal forms**. *Nucleic Acids Res* 2011, doi:10.1093/nar/gkq1059.

13. Davis FP, Sali A: **PIBASE: A comprehensive database of structurally defined protein interfaces**. *Bioinformatics* 2005, doi:10.1093/bioinformatics/bti277.

14. Kiel C, Beltrao P, Serrano L: **Analyzing Protein Interaction Networks Using Structural Information**. *Annu Rev Biochem* 2008,

doi:10.1146/annurev.biochem.77.062706.133317.

15. Aloy P, Russell RB: **Structural systems biology: Modelling protein interactions**. *Nat Rev Mol Cell Biol* 2006, doi:10.1038/nrm1859.

16. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, et al.: **Structure-based prediction of protein-protein interactions on a genome-wide scale**. *Nature* 2012, doi:10.1038/nature11503.

17. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, et al.: **A human protein-protein interaction network: A resource for annotating the proteome**. *Cell* 2005, doi:10.1016/j.cell.2005.08.029.

18. Shoemaker BA, Panchenko AR: **Deciphering protein-protein interactions. Part I. Experimental techniques and databases**. *PLoS Comput Biol* 2007, doi:10.1371/journal.pcbi.0030042.

19. Soni N, Madhusudhan MS: **Computational modeling of protein assemblies**. *Curr Opin Struct Biol* 2017, doi:10.1016/j.sbi.2017.04.006.

20. Baspinar A, Cukuroglu E, Nussinov R, Keskin O, Gursoy A: **PRISM: A web server and repository for prediction of protein-protein interactions and modeling their 3D complexes**. *Nucleic Acids Res* 2014, doi:10.1093/nar/gku397.

21. Guerler A, Govindarajoo B, Zhang Y: **Mapping monomeric threading to protein-protein structure prediction**. *J Chem Inf Model* 2013, doi:10.1021/ci300579r.

22. Hosur R, Peng J, Vinayagam A, Stelzl U, Xu J, Perrimon N, Bienkowska J, Berger B: **A computational framework for boosting confidence in high-throughput protein-protein interaction datasets**. *Genome Biol* 2012, doi:10.1186/gb-2012-13-8-r76.

23. Hosur R, Xu J, Bienkowska J, Berger B: **IWRAP: An interface threading approach with application to prediction of cancer-related protein-protein interactions**. *J Mol Biol* 2011, doi:10.1016/j.jmb.2010.11.025.

24. Vangone A, Oliva R, Cavallo L, Bonvin AMJJ: **Prediction of biomolecular complexes**. In *From Protein Structure to Function with Bioinformatics: Second Edition*. . 2017.

25. Rodrigues JPGLM, Bonvin AMJJ: **Integrative computational modeling of protein interactions**. *FEBS J* 2014, doi:10.1111/febs.12771.

26. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al.: **Highly accurate protein structure prediction with AlphaFold**. *Nat 2021 5967873* 2021, **596**:583–589.

27. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Dustin Schaeffer R, et al.: **Accurate prediction of protein structures and interactions using a three-track neural network**. *Science (80-)* 2021, **373**:871–876.

28. Humphreys I, Pei J, Baek M, Krishnakumar A, Anishchenko I, Ovchinnikov S, Zhang J, Ness TJ, Banjade S, Bagde SR, et al.: **Computed structures of core eukaryotic protein complexes**. *Science (80-)* 2021, **374**.

29. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, Žídek A, Bates R, Blackwell S, Yim J, et al.: **Protein complex prediction with AlphaFold-Multimer**. *bioRxiv* 2022, doi:10.1101/2021.10.04.463034.

30. Hadarovich A, Chakravarty D, Tuzikov A V., Ben-Tal N, Kundrotas PJ, Vakser IA: **Structural motifs in protein cores and at protein–protein interfaces are different**. *Protein Sci* 2021, **30**:381–390.

31. Enright AJ, Illopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events**. *Nature* 1999, doi:10.1038/47056.

32. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting Protein Function and Protein-Protein Interactions from Genome Sequences**. *Science (80-)* 1999, **285**:751–753.

33. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: A structural classification of proteins database for the investigation of sequences and structures**. *J Mol Biol* 1995, doi:10.1016/S0022-2836(05)80134-2.

34. Fox NK, Brenner SE, Chandonia JM: **SCOPe: Structural Classification of Proteins - Extended, integrating SCOP and ASTRAL data and classification of new structures**. *Nucleic Acids Res* 2014, doi:10.1093/nar/gkt1240.

35. Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, Orengo CA, Sillitoe I: **CATH: An expanded resource to predict protein function through structure and sequence**. *Nucleic Acids Res* 2017, doi:10.1093/nar/gkw1098.

36. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, et al.: **The CATH domain structure database: New protocols and classification levels give a more comprehensive resource for exploring evolution**. *Nucleic Acids Res* 2007, doi:10.1093/nar/gkl959.

37. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH - A hierarchic classification of protein domain structures**. *Structure* 1997, doi:10.1016/s0969-2126(97)00260-8.

38. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al.: **Pfam: The protein families database in 2021**. *Nucleic Acids Res* 2021, **49**:D412–D419.

39. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim BH, Grishin N V.: **ECOD: An Evolutionary Classification of Protein Domains**. *PLoS Comput Biol* 2014, **10**.

40. Mosca R, Céol A, Stein A, Olivella R, Aloy P: **3did: A catalog of domain-based interactions of known three-dimensional structure**. *Nucleic Acids Res* 2014, doi:10.1093/nar/gkt887.

41. Stein A, Russell RB, Aloy P: **3did: Interacting protein domains of known three-dimensional structure**. *Nucleic Acids Res* 2005, doi:10.1093/nar/gki037.

42. Stein A, Céol A, Aloy P: **3did: Identification and classification of domain-based interactions of known three-dimensional structure**. *Nucleic Acids Res* 2011, doi:10.1093/nar/gkq962.

43. Winter C: **SCOPPI: a structural classification of protein-protein interfaces**. *Nucleic Acids Res* 2006, doi:10.1093/nar/gkj099.

44. Jefferson ER, Walsh TP, Roberts TJ, Barton GJ: **SNAPPI-DB: A database and API of structures, iNterfaces and Alignments for Protein-Protein Interactions**. *Nucleic Acids Res* 2007, doi:10.1093/nar/gkl836.

45. Teyra J, Doms A, Schroeder M, Pisabarro MT: **SCOWLP: A web-based database for detailed characterization and visualization of protein interfaces**. *BMC Bioinformatics* 2006, doi:10.1186/1471-2105-7-104.

46. Teyra J, Samsonov SA, Schreiber S, Pisabarro MT: **SCOWLP update: 3D classification of protein-protein, -peptide, -saccharide and -nucleic acid interactions, and structure-based binding inferences across folds**. *BMC Bioinformatics* 2011, doi:10.1186/1471-2105-12-398.

47. Zhang QC, Petrey D, Garzón JI, Deng L, Honig B: **PrePPI: a structure-informed database of protein-protein interactions**. *Nucleic Acids Res* 2013, **41**.

48. Mirabello C, Wallner B: **Topology independent structural matching discovers novel templates for protein interfaces**. In *Bioinformatics*. . 2018.

49. Kanitkar TR, Sen N, Nair S, Soni N, Amritkar K, Ramtirtha Y, Madhusudhan MS: **Methods for Molecular Modelling of Protein Complexes**. *Methods Mol Biol* 2021, **2305**:53–80.

50. Kundrotas PJ, Zhu Z, Janin J, Vakser IA: **Templates are available to model nearly all complexes of structurally characterized proteins**. *Proc Natl Acad Sci U S A* 2012, doi:10.1073/pnas.1200678109.

51. Tuncbag N, Keskin O, Nussinov R, Gursoy A: **Fast and accurate modeling of protein-protein interactions by combining template-interface-based docking with flexible refinement**. *Proteins Struct Funct Bioinforma* 2012, doi:10.1002/prot.24022.

52. Kuzu G, Gursoy A, Nussinov R, Keskin O: **Exploiting conformational ensembles in modeling protein-protein interactions on the proteome scale**. *J Proteome Res* 2013, doi:10.1021/pr400006k.

53. Gao M, Skolnick J: **Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected**. *Proc Natl Acad Sci U S A* 2010, doi:10.1073/pnas.1012820107.

54. Verma R, Pandit SB: **Unraveling the structural landscape of intra-chain domain interfaces: Implication in the evolution of domain-domain interactions**. *PLoS One* 2019, doi:10.1371/journal.pone.0220336.

55. Aloy P, Ceulemans H, Stark A, Russell RB: **The relationship between sequence and interaction divergence in proteins**. *J Mol Biol* 2003, doi:10.1016/j.jmb.2003.07.006.

56. Aloy P, Russell RB: **Interrogating protein interaction networks through structural biology**. *Proc Natl Acad Sci U S A* 2002, **99**:5896–5901.

57. Bhattacharyya M, Upadhyay R, Vishveshwara S: **Interaction Signatures Stabilizing the NAD(P)-Binding Rossmann Fold: A Structure Network Approach**. *PLoS One* 2012, doi:10.1371/journal.pone.0051676.

58. Medvedev KE, Kinch LN, Dustin Schaeffer R, Pei J, Grishin N V.: **A Fifth of the Protein World: Rossmann-like Proteins as an Evolutionarily Successful Structural unit**. *J Mol Biol* 2021, **433**:166788.

59. Nagano N, Orengo CA, Thornton JM: **One fold with many functions: The evolutionary relationships between TIM barrel families based on their sequences, structures and functions**. *J Mol Biol* 2002, doi:10.1016/S0022-2836(02)00649-6.

60. Nguyen MN, Madhusudhan MS: **Biological insights from topology independent comparison of protein 3D structures**. *Nucleic Acids Res* 2011, **39**.

61. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389–402.

62. Vernon RMC, Chong PA, Tsang B, Kim TH, Bah A, Farber P, Lin H, Forman-Kay JD: **Pi-Pi contacts are an overlooked protein feature relevant to phase separation**. *Elife* 2018, doi:10.7554/eLife.31486.

63. Morelli X, Bourgeas R, Roche P: **Chemical and structural lessons from recent successes in protein-protein interaction inhibition (2P2I)**. *Curr Opin Chem Biol* 2011, doi:10.1016/j.cbpa.2011.05.024.

64. Mullard A: **Proteing-protein interaction inhibitors get into the groove**. *Nat Rev Drug Discov* 2012, doi:10.1038/nrd3680.

65. Nguyen MN, Sen N, Lin M, Joseph TL, Vaz C, Tanavde V, Way L, Hupp T, Verma CS, Madhusudhan MS: **Discovering Putative Protein Targets of Small Molecules: A Study of the p53 Activator Nutlin**. *J Chem Inf Model* 2019, doi:10.1021/acs.jcim.8b00762.

66. Arkin MR, Tang Y, Wells JA: **Small-molecule inhibitors of protein-protein interactions: Progressing toward the reality**. *Chem Biol* 2014, doi:10.1016/j.chembiol.2014.09.001.

67. Corbi-Verge C, Kim PM: **Motif mediated protein-protein interactions as drug targets**. *Cell Commun Signal* 2016, doi:10.1186/s12964-016-0131-4.

68. Skwarczynska M, Ottmann C: **Protein-protein interactions as drug targets**. *Future Med Chem* 2015, doi:10.4155/fmc.15.138.

69. Higueruelo AP, Jubb H, Blundell TL: **Protein-protein interactions as druggable targets: Recent technological advances**. *Curr Opin Pharmacol* 2013, doi:10.1016/j.coph.2013.05.009.

70. Shiozaki EN, Chai J, Rigotti DJ, Riedl SJ, Li P, Srinivasula SM, Alnemri ES, Fairman R, Shi Y: **Mechanism of XIAP-mediated inhibition of caspase-9**. *Mol Cell* 2003, doi:10.1016/S1097-2765(03)00054-6.

71. Wist AD, Gu L, Riedl SJ, Shi Y, McLendon GL: **Structure-activity based study of the Smac-binding pocket within the BIR3 domain of XIAP**. *Bioorganic Med Chem* 2007, doi:10.1016/j.bmc.2007.02.010.

72. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ: **AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.** *J Comput Chem* 2009, **30**:2785–91.

73. Lang PT, Brozell SR, Mukherjee S, Pettersen EF, Meng EC, Thomas V, Rizzo RC, Case DA, James TL, Kuntz ID: **DOCK 6: combining techniques to model RNA-small molecule complexes.** *RNA* 2009, **15**:1219–30.

74. Sen N, Kanitkar TR, Roy AA, Soni N, Amritkar K, Supekar S, Nair S, Singh G, Madhusudhan MS: **Predicting and designing therapeutics against the Nipah virus**. *PLoS Negl Trop Dis* 2019, doi:10.1371/journal.pntd.0007419.

75. Humphreys I, Pei J, Baek M, Krishnakumar A, Anishchenko I, Ovchinnikov S, Zhang J, Ness TJ, Banjade S, Bagde SR, et al.: **Computed structures of core eukaryotic protein complexes**. *Science (80-)* 2021, **374**.

76. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M: **Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions**. *Nucleic Acids Res* 2013, **41**.

77. Lewis TE, Sillitoe I, Lees JG: **cath-resolve-hits: a new tool that resolves domain matches suspiciously quickly**. *Bioinformatics* 2019, **35**:1766–1767.

78. Henrick K, Thornton JM: **PQS: A protein quaternary structure file server**. *Trends Biochem Sci* 1998, doi:10.1016/S0968-0004(98)01253-5.

79. Elez K, Bonvin AMJJ, Vangone A: **Biological vs. Crystallographic Protein Interfaces: An Overview of Computational Approaches for Their Classification**. *Cryst 2020, Vol 10, Page 114* 2020, **10**:114.

80. Davis FP, Braberg H, Shen MY, Pieper U, Sali A, Madhusudhan MS: **Protein complex compositions predicted by structural similarity**. *Nucleic Acids Res* 2006, doi:10.1093/nar/gkl353.

81. **Discriminating between homodimeric and monomeric proteins in the crystalline state - Ponstingl - 2000 - Proteins: Structure, Function, and Bioinformatics - Wiley Online Library**. [date unknown],

82. Han JDJ, Berlin N, Hao T, Goldberg DS, Berriz GF, Zhang L V., Dupuy D, Walhout AJM, Cusick ME, Roth FP, et al.: **Evidence for dynamically organized modularity in the yeast protein–protein interaction network**. *Nat 2004 4306995* 2004, **430**:88–93.

83. Borrel A, Regad L, Xhaard H, Petitjean M, Camproux AC: **PockDrug: A model for predicting pocket druggability that overcomes pocket estimation uncertainties**. *J Chem Inf Model* 2015, **55**:882–895.

84. Kozlovskii I, Popov P: **Spatiotemporal identification of druggable binding sites using deep learning**. *Commun Biol 2020 31* 2020, **3**:1–12.

85. Krasowski A, Muthas D, Sarkar A, Schmitt S, Brenk R: **DrugPred: A structure-based approach to predict protein druggability developed using an extensive nonredundant data set**. *J Chem Inf Model* 2011, **51**:2829–2842.

86. Pu L, Govindaraj RG, Lemoine JM, Wu HC, Brylinski M: **DeepDrug3D: Classification of ligand-binding pockets in proteins with a convolutional neural network**. *PLOS Comput Biol* 2019, **15**:e1006718.

87. Laniado J, Meador K, Yeates TO: **A fragment-based protein interface design algorithm for symmetric assemblies**. *Protein Eng Des Sel* 2021, **34**.

88. Šali A, Blundell TL: **Comparative protein modelling by satisfaction of spatial restraints**. *J Mol Biol* 1993, **234**:779–815.

89. Chen J, Sawyer N, Regan L: **Protein-protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area**. *Protein Sci* 2013, doi:10.1002/pro.2230.

90. Yan C, Wu F, Jernigan RL, Dobbs D, Honavar V: **Characterization of protein-protein interfaces**. *Protein J* 2008, doi:10.1007/s10930-007-9108-x.

91. Dhawanjewar AS, Roy AA, Madhusudhan MS: **A knowledge-based scoring function to assess quaternary associations of proteins**. *Bioinformatics* 2020,

doi:10.1093/bioinformatics/btaa207.

92. Nguyen MN, Tan KP, Madhusudhan MS: **CLICK--topology-independent comparison of biomolecular 3D structures.** *Nucleic Acids Res* 2011, **39**:W24-8.

93. Roy AA, Dhawanjewar AS, Sharma P, Singh G, Madhusudhan MS: **Protein Interaction Z Score Assessment (PIZSA): an empirical scoring scheme for evaluation of protein-protein interactions**. *Nucleic Acids Res* 2019, doi:10.1093/nar/gkz368.

94. Dhawanjewar AS, Roy AA, Madhusudhan MS: **A knowledge-based scoring function to assess the stability of quaternary protein assemblies**. *bioRxiv* 2019, doi:10.1101/562520.

95. Wang G, Dunbrack RL: **PISCES: A protein sequence culling server**. *Bioinformatics* 2003, **19**:1589–1591.

96. Tan KP, Nguyen TB, Patel S, Varadarajan R, Madhusudhan MS: **Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins.** *Nucleic Acids Res* 2013, **41**.

97. Huang B: **MetaPocket: a meta approach to improve protein ligand binding site prediction**. *OMICS* 2009, **13**:325–330.

98. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA: **Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure**. *PLoS Comput Biol* 2009, doi:10.1371/journal.pcbi.1000585.