Probing Protein Interaction Networks by Combining MS-Based Proteomics and Structural Data Integration

Guillaume Postic, Julien Marcoux, Victor Reys, Jessica Andreani, Yves Vandenbrouck, Marie-Pierre Bousquet, Emmanuelle Mouton-Barbosa, Sarah Cianférani, Odile Burlet-Schiltz, Raphael Guerois, et al.

Probing protein interaction networks by combining MS-based proteomics and structural data integration

Journal:	*Journal of Proteome Research*
Manuscript ID	pr-2020-000662.R2
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Postic, Guillaume; Université de Paris Marcoux, Julien; IPBS Reys, Victor; Centre de Biochimie Structurale Andreani, Jessica; CEA, VandenBrouck, Yves; Université Grenoble Alpes Bousquet, Marie-Pierre; Institut de Pharmacologie et de Biologie Structurale, Mouton-Barbosa, Emmanuelle; cnrs, ipbs Cianférani, Sarah; Institut Pluridisciplinaire Hubert Curien (IPHC), Laboratoire de Spectrométrie de Masse Bio-Organique (LSMBO) Burlet-Schiltz, Odile; cnrs, ipbs Guerois, Raphael; Institute for Integrative Cell Biology Labesse, Gilles; CNRS, Centre de Biochimie Structurale Tuffery, Pierre; Universite Paris Diderot, BFA, CNRS UMR8251, INSERM U1133

Probing protein interaction networks by combining MS-based proteomics and structural data integration

Guillaume Postic[1,2], Julien Marcoux[3], Victor Reys[5], Jessica Andreani[6], Yves Vandenbrouck[4], Marie-Pierre Bousquet[3], Emmanuelle Mouton-Barbosa[3], Sarah Cianférani[7], Odile Burlet-Schiltz[3], Raphael Guerois[6], Gilles Labesse*[5], Pierre Tufféry*[1]

[1]Université de Paris, BFA, UMR 8251, CNRS, ERL U1133, Inserm, RPBS, F-75013 Paris, France

[2]Institut Français de Bioinformatique (IFB), UMS 3601-CNRS, Universite Paris-Saclay, Orsay, France

[3]Institut de Pharmacologie et de Biologie Structurale, IPBS, Université de Toulouse, CNRS, UPS, Toulouse, France

[4]Univ. Grenoble Alpes, INSERM, CEA, IRIG-BGE, U1038, 38000, Grenoble, France

[5]CBS, Univ. Montpellier, CNRS, INSERM, Montpellier, France

[6]Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France

[7]Laboratoire de Spectrométrie de Masse BioOrganique, Université de Strasbourg, CNRS, IPHC UMR 7178, 67000 Strasbourg, France

Corresponding authors: pierre.tuffery@u-paris.fr; labesse@cbs.cnrs.fr

ABSTRACT

Protein–protein interactions play a major role in the molecular machinery of life and various techniques such as AP-MS are dedicated to their identification. However, those techniques return lists of proteins devoid of organizational structure, not detailing which proteins interact with which others. Proposing a hierarchical view of the interactions between the members of the flat list becomes highly tedious for large datasets when done by hand.

To help hierarchize this data, we introduce a new bioinformatics protocol that integrates information of the multimeric protein 3D structures available in the Protein Data Bank using remote homology detection, as well as information related to Short Linear Motifs and interaction data from the BioGrid. We illustrate on two unrelated use-cases of different complexity how our approach can be useful to decipher the network of interactions hidden in the list of input proteins, and how it provides added value compared to state-of-the-art resources such as Interactome3D or STRING. Particularly, we show the added value of using homology detection to distinguish between orthologs and paralogs, and to distinguish between core obligate and more facultative interactions. We also demonstrate the potential of considering interactions occurring through Short Linear Motifs.

KEYWORDS: Mass spectrometry, proteomics, protein interaction networks, protein complexes, computational proteomics, structural bioinformatics, remote homology, short linear motifs

1. INTRODUCTION

The vast majority of cellular functions are carried out by molecular machines made up of non-covalent protein–protein interactions (PPIs) and called protein complexes. However, PPI characterization is still far from being complete. Estimates are that between 120,000 and over one million such interactions occur in *Homo sapiens* [1], but only on the order of 50,000 high quality PPIs are currently identified [2]. Turning to the structural level, estimates are that in the human proteome (restricted to proteins in the human interactome) two thirds of the proteins have either a structure experimentally resolved (~17,000 structures less than 95% identical in the PDB) or at least a part accessible as structural models. But again, moving to the human interactome, only 12% of interactions have a structure/model (https://interactome3d.irbbarcelona.org/statistics.php) and currently ~2,200 heteromeric complexes are structurally characterized. This has motivated a large effort of the community to identify ways to increase our knowledge on PPIs.

Screening techniques to detect interacting partners fall into two categories: (i) genetic approaches, which measure mostly pairwise interactions and (ii) biochemical methods, which measure the co-complex interactions among multiple proteins. Both types of techniques have matured to the point where they now allow from focused to proteome-wide PPI network analyses. The genetic approaches consist of a test between two fusion proteins, whose interaction either activates the transcription of a reporter gene or restores the activity of a reporter enzyme. Examples of such binary techniques include the yeast-two hybrid system [3], bacterial-two hybrid system [4], MAmmalian Protein–Protein Interaction Trap [5] and protein complementation assays [6]. Large-scale biochemical methods mainly use enrichment strategies to capture a bait protein under conditions aiming at preserving native associations to its protein partners. Then, the bait is eluted with its potential interacting partners (preys), and subsequently

identified and quantified using liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS). Typical co-complex techniques include the affinity purification-mass spectrometry (AP-MS) [7,8], which involves tagging bait proteins with an epitope before purification of both bait and interacting partners using antibody affinity. An alternative method called proximity-dependant biotinylation consists in fusing an enzyme to the bait in order to modify its interacting partners [9]. Such techniques return a "flat" list of identified and quantified proteins that are all supposed to interact, directly or indirectly, with the bait protein. Although it can be associated with an observance number quantifying the number of occurrences of each peptide, at this stage, one issue is the absence of an organizational structure in the output list, to see which proteins do interact with which others. Proposing a hierarchical view of the interactions between the members of the flat list is extremely useful for biological interpretation, but tedious when done by hand. Moreover, one has also to consider that the list itself should be questioned: it may suffer from (i) the presence of identified PPIs that were wrongly assigned or do not occur physiologically and (ii) the non-detection of genuine associations—that we referred to hereafter as "false positives" and "false negatives", respectively (see Figure 1).

False positives in proteomics data may arise from various causes [8]. For example, when the bait protein is a member of different complexes in the cell, it will capture a set of prey proteins in which some never occur in the same complex (see Figure 1). Technically, for AP-MS experiments, non-specific binding of proteins to the antibody or the solid matrix may also contribute to the purification of proteins that do not specifically interact with the protein of interest. Experimental results are not either exempt from false negatives, as the integrity of the detected protein complexes may be compromised by different factors. For example, in biochemical co-complex methods, the purification step hinders transient and weak interactions

from being detected. Reversible [10] or irreversible [11] chemical cross-linking has been proposed to overcome this problem, but at the price of introducing a bias in the pull-down data. Finally, a low number of proteolytic peptides may lead to either false positives (identification of wrong proteoforms, e.g. splice variants) or false negatives (signal below the detection threshold). With regard to genetic binary approaches, transient complexes are considered, but all non-tagged proteins are excluded from the analysis.

One way to address these questions is to integrate the information of the multimeric protein 3D structures available in the Protein Data Bank (PDB) (http://www.rcsb.org/) [12]. Indeed, the experimental resolution of the structure of a complex (by methods such as X-ray, RMN, or cryo-EM) supposes a high enough binding affinity between the partners and is, therefore, a direct proof of the interaction. Moreover, the purification performed prior to the determination of the protein structure avoids the presence of artifactual partners (*i.e.* false positives) in the resulting 3D data. To date, this source of information encompasses the resolution of over 30,000 homomeric and 7,700 heteromeric non redundant complex structures. Furthermore, it is now well established that interactions tend to be preserved throughout evolution and interface binding modes are structurally conserved [13–15], even for complexes involving partners with as little as 20% sequence identity [16]. This provides a means for the large-scale identification of complexes, even for species in which very few structures are available, through the search of interologs (i.e. homologous complexes across species). Importantly, the oligomeric stability required by protein structure determination techniques guarantees the integrity of the "obligate" complex—which is the smallest set of subunits forming a stable complex, as observed among various species—and, thus, the absence of false negatives. Therefore, this also provides a means to control that all members of the obligate complex are observed in the interactome data, and

possibly, to identify false negatives. Finally, the knowledge of the structure directly provides information about the residues at the interface, providing the means to design further experiments to validate or modulate the interactions.

Turning to weaker or transient interactions, more and more evidence highlights the role of Short Linear Motifs (SLiMs), which correspond to short segments of the proteins mostly located within intrinsically disordered regions (IDRs). Many are documented in the PDB but are more difficult to detect by usual sequence comparison algorithms or fold recognition due to the small size of these SLiMs. In these cases, data accumulated in the Eukaryotic Linear Motif resource (ELM) (http://elm.eu.org/) [17] could prove helpful to decipher the potential interactions occurring between proteins of the "flat" list.

Although the whole process of integrating structural and interactomics data can be done manually for small datasets, it requires advanced skills in searching structure information. Moreover, the task becomes highly tedious and error prone for larger or multiple datasets. To the best of our knowledge, the only available computational tool designed for that purpose is Interactome3D (https://interactome3d.irbbarcelona.org/) [18]. Its main goals are to propose 3D models of the protein complexes of a given species, and to provide insights about the protein residues involved in the interactions. It consists of a database setup to provide a compendium of the structural information on the PPIs available for a given organism. It integrates experimental structures or models of protein complexes for that organism and information extracted from external resources such as IntAct (https://www.ebi.ac.uk/intact/) [19]. Models are either extracted from ModBase (https://modbase.compbio.ucsf.edu/) [20] or built by comparative modeling - at more than 40% sequence identity, combining blast, 3did (https://3did.irbbarcelona.org/) [21] and MODELLER (https://salilab.org/modeller/) [22]. Currently

it covers 18 organisms. For a query species, the server takes as input a list of proteins or their binary interactions, to return structural annotations of the PPIs network of the query. Another popular resource to annotate PPIs is the STRING database (https://string-db.org/) [23]. Unlike Interactome3D, STRING considers protein structural data as "accessory information" and does not use it to build PPIs networks. Rather, it uses other sources of information, such as genetic interactions, text mining, or PPIs from other databases. Finally, our comparative analysis also includes GeneMANIA [24] (http://genemania.org), as it can perform functional annotation and prediction from a flat list of protein genes, by integrating data from PPIs databases, such as BioGRID (https://thebiogrid.org/) [25] or PathwayCommons (https://www.pathwaycommons.org/) [26]. Interestingly, GeneMANIA is able to predict ortholog PPIs (interologs) through the integration of data from other species. However, it is currently limited to only nine organisms. The gap between the information content reachable by such resources and structural information remains however unclear.

Here, we present an integrative protocol aimed at validating and connecting candidate protein partners detected by interactomics experiments, starting from the "flat" list of the detected proteins, i.e. without *a priori* information on the actual interactions. The method is based on a large-scale search in the PDB for close and remote homologs of each member of a given list. It is followed by an analysis in terms of complexes to which these homologs belong to, in order to identify the complexes grouping the input proteins and provide a first level of data structuring. This protocol also extracts predicted weak and transient PPIs, through the search for Short Linear Motifs (SLiMs) located in intrinsically disordered regions (IDRs). Finally, the whole network analysis is further enriched by integrating experimental PPIs downloaded from the BioGRID database, which can be informative when no structures are available. Using two

datasets generated by MS-based proteomics experiments as case studies, we describe an original strategy for validating and expanding interaction data, and discuss its added-value and limitations. Source code and instructions describing how to set up the full protocol are freely available at https://gitlab.rpbs.univ-paris-diderot.fr/src/proteo3Dnet.

2. MATERIALS AND METHODS

2.1 Enlarged homology search

Because only a minor portion (~1%) of protein sequences have a 3D structure available in the PDB, we perform a search for known structures to be used as templates for comparative modeling. The main component of our protocol consists in mining the full set of PDB structures to find those templates, and possibly some structures that gathers several of the proteins of the list established by proteomics experiments, i.e. PDB entries corresponding to the structure of complexes involving several proteins of the list identified by the interactomics experiment. Fold-recognition was first performed automatically using the @TOME-2 server (http://atome.cbs.cnrs.fr/ATOME_V3/index.html) [27] to unravel the structural coverage for our queries. However, this does not directly highlight the multimeric organization of the template and hence the putative interactions among the query proteins. We resumed the search for homologs using only HHsearch (https://github.com/soedinglab/hh-suite version 3.0.0) [28] and considered further only templates with a probability >95.0%. Using such cut-off, the risk to miss some remote templates exists - although in our experience it is very limited. However, we prefer here to limit such risk rather than to possibly introduce false positive templates. Detecting templates with a large confidence, the underlying hypothesis is that if several proteins form a complex in a 3D structure, their homologs are likely to interact similarly, i.e. to correspond to

interologs. A specific difficulty comes from the fact that to perform efficiently, HHsearch searches for homologs in a subset of the PDB whose redundancy has been reduced to 70% sequence identity, through a clustering procedure - all proteins of a cluster are represented by only one cluster representative. Because any cluster representative identified by HHsearch as the closest homolog is not necessarily the closest homolog that could be found in the entire PDB, and because it could occur that non representatives of the clusters correspond to PDB entries containing the structures of complexes of interest, we enlarge the HHsearch results with all the members of a cluster. However, we still need to estimate the sequence identity of the query to each member of the cluster to identify the cluster member best corresponding to the query. Here, a new difficulty comes from the fact that HHsearch clusters are determined based on the complete protein sequence, and that in many cases, the resolved structure corresponds to only one part of it. In some cases, for two members of the same cluster, the regions for which a structure could be determined do not overlap at all. This difficulty is overcome by using MaxCluster (http://www.sbg.bio.ic.ac.uk/maxcluster/) (version 0.6.6), to subdivide, when necessary, each cluster in sub-clusters. Then, for each sub-cluster, we build a multiple sequence alignment (MSA) using the MUSTANG program (version 3.2.3) [29], which uses both protein sequence and structure information. Because it makes use of 3D information, this 3D based MSA is expected to be of better quality than an alignment based on the sequence alone. Finally, for any homolog identified by HHsearch, we use MAFFT (version 7.407) [30] to merge the pairwise alignment produced by HHsearch with the MSA and we calculate the sequence identity between the target and each sequence in the MSA. This way, each homolog initially identified by HHsearch may be replaced by a cluster member of higher sequence similarity to the query.

2.2 Data integration

After the search for 3D structures, each protein from the input list is associated with one, several, or no PDB entries. When several members of the input list are found in the same PDB entry (3D complex, biological unit), they are considered as interacting partners. Conversely, when a 3D complex found contains a protein chain not present in the proteomics dataset, this chain corresponds to a potentially missing partner of the complex; the subunit is therefore labeled as undetected. It is not necessarily a false negative, as it may be part of a non-obligate complex. An index of the number of times the chains are seen in the same complex is built to identify the proteins present in the complexes but not in the flat list, and to evaluate if they are part of the obligatory complex. Regarding the homo-oligomeric state, the number of copies of a protein chain within a complex is determined by extracting information about the stoichiometry in the PDB (author annotated biological unit). For example, a structure annotated by 'A3' ("Global Stoichiometry" field) is a homotrimer, whereas the 'A2B' annotation corresponds to a heterotrimer that contains two copies of the chain A and one of the chain B. For cases where there are several molecular assemblies assigned, the one that is labeled as "assigned by authors" (if any) is systematically preferred. Finally, it may happen that no 3D structure can be found for a candidate protein. In this case, the pipeline also integrates an additional source of interaction data: the Biological General Repository for Interaction Datasets (BioGRID) [25], from which only physical (not genetic) interactions between proteins are retrieved, except those associated with the 'Far Western', 'Co-fractionation', 'Co-localization', 'Biochemical Activity', and 'High Throughput' experimental systems.

2.3 Search for SLiMs

SLiMS are short motifs known to mediate PPIs and their identification can be useful to decipher weak interactions. They are located in disordered regions of proteins, and often in missing parts of resolved structures, We search into the protein sequences of the flat list, for occurrence of

such motifs—represented as regular expressions—as documented in the ELM database [31]. We only consider motifs found within disordered binding regions, as predicted by the ANCHOR2 score [32]. The latter is calculated with the IUPred2A software [32] and only motifs with an average ANCHOR2 score >0.95 are kept. ELM links motifs with interacting PFAM domains. Thus occurrence of such PFAM domains—as annotated in Uniprot—are also searched within the input proteins. Occurrences of pairs involving SLiMs-PFAM domains correspond to potential PPIs. We emphasize that, although ELM contains information on fully validated such interactions, the procedure relies for a large part on predictions and consequently, some erroneous pairs could be suggested, and some others missed. In our experience however, the added value of this detection largely counterbalances such risks.

2.4 Datasets

Proteasome 20S. The proteasome dataset was retrieved from [33]. Briefly, proteasome complexes were immuno-purified from in vivo formaldehyde crosslinked human U937 cells using the MCP21 antibody targeting the α-2 subunit of the 20S core particle. After trypsin digestion and extraction, peptides were identified by nano-liquid chromatography (LC) MS/MS. The database search was performed using the Mascot Daemon software (version 2.3.2, Matrix Science). Detailed information can be found in the original paper [33].

Pragmin. The Pragmin dataset was retrieved from [34]. Note that the 61 proteins of this dataset do not contain the Pragmin itself. Briefly, a SILAC-based quantitative proteomic analysis, which allowed the identification in a semi-quantitative manner of Pragmin interactors in FLAG-Pragmin transfected human HEK293T cells, was undertaken. It reproducibly yielded more than 52 specific interactors, with a mean ratio >4 and found two protein-kinases that were

prominently associated with Pragmin (mean ratio >12), including the Ser/Thr kinases AMPK and the TK CSK. Previous Pragmin interactors such as Src [35] and SgK269/PEAK1 [36] were, however, not recovered in our SILAC analysis. Detailed information can be found in the original paper [34].

3. RESULTS AND DISCUSSION

Figure 2 presents an overview of the full protocol that we introduce here. We discuss shortly its main lines, details of each component being presented in the Materials and Methods section. Starting from the list of proteins established by proteomics experiments—that could be, for instance, specified as a series of UniProt identifiers—it consists of 4 steps:

1) The search for homologs of each of the input proteins. The main task of our protocol consists in finding at least one PDB structure that gathers several of the UniProt identifiers from the list. However, only a minor portion (~1%) of UniProt entries have a 3D structure available in the PDB—due to the difficulty of determining protein structures experimentally. To overcome this difficulty, the protocol performs instead a search for homologs using HHsearch [28], which is presently one of the best-performing approaches to detect remote homologs. Being oriented towards the identification of templates for homology modeling, it returns only one representative per cluster of similar structures of the PDB, which is undesired in our case as it might miss associations between input proteins. For this reason, the standard HHsearch procedure is followed by a specific search in the clusters to identify the cluster member best matching a query. This step is, by far, the most time consuming part of the protocol. Although the number of input sequences is in theory not limited, it can require several hours for an input of over 700 sequences, depending on the running machine.

2) Given all homologs for all input proteins, a second step is the search for PDB entries combining homologs of several input proteins, so as to identify complexes of resolved structure likely to involve the proteins of the flat list. Complexes corresponding to strict interologs are identified, which makes possible the identification of the core obligate complexes versus the identification of context dependent binders. Information on homo-oligomery is also searched automatically.
3) A third step consists of the identification of segments likely to correspond to SLiMs (i.e. short segments belonging to regions predicted to have a weak binding affinity for some specific sequence motifs) following the rules developed for the database ELM (see methods).
4) The last step consists in linking the results to the information retrieved from the BioGRID database.

As a result, it becomes possible to build a graph in which the nodes correspond to the proteins of the list and the edges to the interactions identified between them. This graph can be interpreted in terms of direct or indirect binding to the bait. The interface of the binding partners identified in terms of 3D contacts (from PDB), or sequence motifs (SLiMS) provide the means for further exploration.

We present below how our protocol performs for two recently described interactomes of the proteasome and pragmin. Those interactomes vary in the number of proteins detected and in their complexity. The proteasome interactome was taken to test the ability of the protocol to get confronted with complex structural data involving orthologs and paralogs, whereas the pragmin interactome appears to contain information that requires more integrative analyses. As the originality of our protocol mostly lies in the search for homologs and SLiM data integration, we focus on these aspects rather than on the added value of integrating BioGRID data. Finally,

as such examples from the proteasome interactome show the importance of detecting remotely homologous complexes, we have included them in Figure 3 to illustrate the concept of interologs.

3.1 Proteasome interactome

The 20S proteasome is a huge proteolytic complex involved in the regulation of intracellular protein homeostasis. Its 20S catalytic core is composed of 14 different subunits that assemble as a 716 kDa barrel made of 4 heptameric stacked rings [37]. The 20S proteasome can then interact with different regulators, the most abundant of which is the 19S regulatory particle, that is made of ATPase (RPT) and non-ATPase (RPN) subunits and forms the ATP-dependant 26S ubiquitin proteasome system. On top of its many substrates, which only transiently contact the catalytic core, more than 300 Proteasome Interacting Proteins (PIPs) are known to interact with the 20S proteasome. Due to its central role in homeostasis, the proteasome has been thoroughly studied from a structural point of view by cryo-Electron Microscopy [38–40], X-ray crystallography [41] and Nuclear Magnetic Resonance [42]. MS-based techniques have also brought important insights on the primary to quaternary structures of the many different proteasome complexes. While top-down proteomics informs on the proteoforms present within the 20S core particle [43,44], bottom-up proteomics is commonly used to identify novel proteasome partners in complex mixtures [45], and physical interactions with putative binders of the proteasome have been confirmed by native MS [46,47]. Finally cross-linking MS and bioinformatics modeling have been extensively used to better characterize highly enriched proteasome complexes [48,49] but also, more recently, to identify and model the interactions with new partners found in complex mixtures or even *in vivo* [50,51]. While distance restraints are not compatible with our current pipeline, lists of interactors identified by cross-linking MS experiments can readily be used as an input.

Figure 4 depicts a graph summarizing the results of the protocol. Of the 192 partners identified by immuno-purification (IP) [25] of the 20S proteasome (via its constitutive α2 subunit, in red in Figure 5), only 9 (CFDP1, CG050, GPAM1, NDUF4, PIP30, PMGT2, POMP and RL29 STK19) were not assigned reliably to any structural template. For a small number (17) only very distant similarity but consistent (e.g.: F120A) could be detected with sequence identity ranging from 12 to 22%. Among the latter, three proteins (PSMG1, PSMG2 and PSME4) are involved in proteasome maturation. On the contrary, for 109 proteins identical structures were matched. For 145 and 164 proteins, sequence identity shared with the detected templates for structure modeling was above 65% and above 35% respectively. Hence, an almost complete structural coverage (95%) was reached (http://atome4.cbs.cnrs.fr/ms2model_proteasome).

Up to 70% of the detected partners were assigned to proteins found in homo-multimeric complexes (Table S1). We found for example the expected homo-heptameric activators PA28α (PDB 5MSJ[52]) and PA28β (PDB 5MSK[52]). Interestingly, these homo-heptamers are not physiological, as they naturally form within the cell an α4β3 hetero-heptamer, that is better at activating the 20S proteasome [52]. There is no structure available yet for the homo-heptameric PA28γ, and the best homology model proposed was the one of PA28α (PDB 5MSJ). As this part of the pipeline is strictly looking for homomers, the α-subunits α1 to α7 that form the hetero-heptameric α-ring of the human 20S proteasome, were assigned to the distantly related homo-heptameric α-ring from *Archaeoglobus fulgidus* (PDB: 1J2P[53]).

In the next step, 128 complexes were identified (Table S2), containing at least 2 identical and/or homologous proteins from our IP and 20 of these were 100% identical, including 8 dimers, 2 trimers, a 10 hetero-oligomers ranging from 7-mer to 33-mer. However 96 out of 192 proteins could not be assigned to any of these 128 complexes (50% of isolated proteins). The ELM search rescues 17 of these proteins and includes them in the interaction network (41% of

isolated proteins). The Biogrid search further added 25 more proteins to the interactome, which led to a total of 54 (28%) unexplained preys, that can either be false negatives, new interactors or transient substrates of the proteasome.

All the 33 subunits of the 26S proteasome (Figure 5, red, blue and green) bound to USP14-UbAI (Figure 5, purple, PDB: 5GJQ[54]) were found in the IP, meaning that this complex is most likely present . 21 Smaller sub-complexes of the 26S ranging from 20 (PDB: 5L4G) to 31 (PDB: 5GJR[54]) subunits, were also found. Structures of even smaller complexes or subcomplexes (7 to 17 subunits) of the 19S human proteasome regulatory particle itself (Figure 5, green) were also listed. As expected, the human standard 20S proteasome (PDB: 4R3O[55]) was identified by the pipeline (Figure 5, red and blue), together with its 228 homologous structures from mouse, bovin and yeast. A challenge for the yeast proteins was to automatically discriminate the correct pairs of orthologs among the large set of paralogous subunits in human proteasome (7 α- and 7 β-subunits). Interestingly, we could check that all 14 pairs of orthologs between yeast and human were correctly assigned despite important phylogenetic divergence. Furthermore, an alternative to this standard 20S, is called the immunoproteasome, and was also proposed (PDB: 6AVO[56]), as the three immuno-catalytic subunits (namely PSB9, PSB10, and PSB8) were also on the list of α2's partners.

Other complexes involved ubiquitin (UBB), known to target many substrates and enzymatic regulators to the 19S regulator when the latter is bound to the 20S catalytic core. More precisely, it was found in complexes containing ubiquitin binding motifs or 26S-bound deubiquitinating enzymes such as the ternary complexes UBB-RPN13-RPN2 and UBB-RPN13-UCHL5 or heterodimers such as UBB-RAD23A and UBB-RL40. This last dimer is particular because RL40 is the precursor of Ub made of one chain of Ub fused to the 60S ribosomal protein L40. RL40 was actually identified with the same peptide set as UBB, meaning that no peptide from

L40 were identified and that RL40 could well be a false positive. The heterodimers UBB-UCHL5 (PDB: 4UF6[57]) and RPN13-UCHL5 (PDB: 4UEM[57]) were also identified, as they are subcomplexes of the UBB-RPN13-UCHL5 trimer (PDB: 4UEL[57]). Other complexes containing RPN13 include RPN13-RPN2, RPN13-UBQLN2. We then found PDBs corresponding to the heterodimers of RPN10, another subunit of the 19S proteasome, and RAD23A & RAD23B which are multi-Ub chain receptors involved in the shuttling of polyubiquitin substrates to 26S proteasomal degradation.

The 108 remaining complexes were found with an identity that is <100% and correspond for example to close homologues, as is the case for the previously mentioned PA28α4β3 hetero-heptamer from mice (PDB: 5MX5[52]), the UBB-RPN10 heterodimer from yeast (PDB: 5LN1[58]) or the RPN8-RPN11 complex from yeast (PDB: 4OCN[59]). Other complexes were largely incomplete, such as the human 80S ribosome (10 subunits out of 82), the human respiratory complex (3 out of 68), the 48S preinitiation complex (8 out of 36), the human spliceosome (3 out of 51) and the human anaphase promoting complex (2 out of 19), to mention a few.

Finally, a number of putative homologues have caught our attention: it is the case of the E3 Ub-protein ligase that would form a heterodimer with the previously mentioned RL40, based on crystal structures of homologous E3 ligases bound to selective Ub variant probes [60] (PDBs: 5HPK, 5HPL, 5HPS, 5HPT[60]). Another complex between the RL40 precursor and UBP14, a proteasome-associated deubiquitinase is proposed, based on the crystal structure of a complex with the homologous deubiquitinase UBP21 (PDB: 3I3T[61]). Because RL40 was identified via its ubiquitin motif only, these two complexes have to be taken with great caution as they could be false-positives. Of high interest is the predicted interaction between the 20S subunit α5 (PSMA5) and the proteasome assembly chaperone 3 (PAC3, PSMG3_HUMAN) based on the structure of their yeast homologues (2Z5C[16]). A related complex involves the two other main

proteasome assembly chaperones 1 and 2 that are known to form a heterodimeric complex and is predicted to interact with the 20S subunit α6 (PSMA1) based on the yeast homologous proteasome crystallized in complex with Pba1-Pba2 (PDB 4G4S[62]). Remarkably, proteasome chaperones in human and yeast diverged drastically in evolution and share very little sequence identity (around 20%). This case underlines the interest of using HHsearch to reveal remote homology relationships between orthologous proteins. Other expected homology models include the interactions between RPN11 and RAD23A (Ub receptor) or between RPN11 and RPN8.

Other proposed complexes were unexpected and caught our attention, such as the proposed heterodimer between MDC1 and TE2IP that was based on the crystal structure of MDC1 bound to a homologue of TE2IP (TOPBP1). We did not expect this interaction between these two partners within our immuno-purification, because none of these proteins is apparently related to the 20S and its common proteasome interactors, except for the ELM interaction proposed between MDC1 and RPN10 (PSMD4) (Table S3). However, both proteins are involved in the DNA damage response and might either bind to the 20S proteasome as a dimeric subcomplex or constitute proteasomal substrates co-immuno-purified with the protease [63].

3.2 Pragmin interactome

Pragmin is a recently characterized pseudo-kinase from rat and homologous to the human SGK223 [64]. Pragmin was shown to strongly activate a tyrosine protein-kinase Csk and to regulate cell migration [34,65] by a mechanism yet to be fully unraveled.

Among the 61 candidate partners of the Pragmin interactome [34], it was possible to identify at least one solved structure for 60 cases (Table S4) and

http://atome4.cbs.cnrs.fr/ms2model_sgk223. For 48 cases, sequence identity shared with the detected template(s) for structure modeling was above 65% and above 35% for 7 more proteins. 25 structures were identified with a sequence identity of 100%. Remarkably, a near complete structural coverage was reached, thanks to remote homology detection.

Further analyses revealed that 50 proteins were predicted to form homomultimers (Table S4). Figure 6 depicts a graph summarizing the results of the protocol. Most of these (45 candidates) correspond to homodimers, though tri-, tetra-, hexa- and decamers were also found. 15 proteins (14 dimers and one decamer) matched the sequence of the template structure with at least 99% sequence identity which cast little doubt on the actual oligomeric organization for those proteins. For 11 other sequences, one observed a sequence identity of more than 65% with multimeric structures, again suggesting one could infer conservation of the same oligomeric state by homology. For another 12 proteins, sequence identity with multimeric templates was between 62% and 39% and the conservation of the same oligomeric states is more questionable and should be investigated for other evidence. Lower similarities were detected for 14 other sequences but further analyses are necessary to precise the potential assembly of those proteins. The remaining 9 sequences showed no obvious similarity with any multimeric assemblages. Obviously, this already provided quite significant coverage and valuable information compared to a "flat' list of interactants. However, this does not provide direct clues to validate any of the PPIs detected by AP-MS.

Thirty proteins could not be associated with any heteromultimeric complex. Most are expected to be isolated polypeptide chains (or pure homomultimers). 3D information allowed nevertheless the prediction of several complexes (Table S5), and in some cases, structure analysis brings complementary information to the raw interactomic knowledge.

The simultaneous detection of AAPK1, AAKB1, AAKG1- and AAKG2 suggested that the

AMP-activated protein kinase (AMPK) complex corresponds to a heterotetramer. However, this complex is heterotrimeric and made of only three subunits α1 (or α2), β1 (or β2), and γ1 (or γ2) (Figures 7A and 7B). Therefore, the closely related paralogues AAKG1 (γ1) and AAKG2 (γ2) are not expected to participate in the same macromolecular assemblies. Instead, this suggests that these two heterotrimers could co-exist, namely α1β1γ1 and α1β1γ2, which corresponds to an interesting case of false positive identification involving mutually exclusive interactions.

Another complex that was partially identified would be composed of CYFP1, NCKP1, and ABI1 proteins, which are found together within the native structure of the human WAVE regulatory complex (PDB entries 3P8C[66] and 4N78[67]). Again, the observed arrangement in the crystal structure suggested that the paralogous CYFP2 belongs to an alternative complex with NCKP1 and ABI1. The crystal structure also highlights that these proteins form extensive and intricate interactions with two undetected sequences (WASF1) and (BRK1) (Figure 7C) that could correspond to false negatives. Indeed, WASF1 was actually detected in the AP-MS results, but did not pass the standard statistical threshold and was, at first, discarded from the final list of candidates (Serge Urbach, personal communication). The rather short BRK1 protein (75 residues) was not detected at all, although it forms a 55-residue long helix buried in the complex. We wonder whether the small size of BRK1 prevented its straightforward identification by MS (despite the presence of numerous basic residues all along its sequence) or that it was effectively absent from the sample. Of note, BRK1 is detected in interaction with some WAVE components in distinct MS studies according to BioGRID but never systematically associated with all of them.

Similarly, the interactions between the four candidate proteins ARPC2, ARPC4, ARP2, and

ARP3 were confirmed by the identification of the structure of the bovin Arp2/3 complex (entries 1K8K[68] and 3DXK[69]). However, the crystal structure comes as an hetero-heptamer, while three polypeptide chains of this assembly (namely ARC1B, ARPC3, and ARPC5) are missing in this AP-MS flat list. Interestingly, these missing partners appear at the periphery of the complex (Figure 7D) and might not be obligatory or, the interactions with these proteins could be more labile than those between the ARP2-ARP3-ARPC2-ARPC4 core. Alternatively, one of the missing partners, ARPC1, binds to the ARP2/3 complex through its WD40 domain. It is interesting to note that several proteins (KEAP1, DcAF7, GBLP, KIF21B and potentially CB044) detected in the Pragmin interactome contain a WD-40 domain and at least one may replace ARPC1 here. Finally, it is important to note that the detection of this ARP2/3 complex together with that of the WAVE regulatory complex makes sense as they are known to be involved in regulating the actin nucleating activity, in agreement with the role of Pragmin in promoting cell migration [70].

Smaller heterocomplexes were also predicted based on structural conservation such as the two heterodimeric chaperones HS71A-HS71B and HSP7C-HS105 which correspond to the experimental complexes PDB 1XQS[71] and PDB 3C7N[72]. So far, we cannot extend this heterodimerization to the other HSP70-like chaperones found in this interactomic study (namely HSP72, HSP76, HS74L, GRP75 and GRP78).

Of note, the above structural analysis provided no clues to identify which preys actually bind to the Pragmin bait as no direct interaction with Pragmin could be detected, so far. In fact, Pragmin contains a pseudo-protein-kinase domain, a dimerization domain [64] as well as a very long (~950 aa) N-terminal tail. The latter is predicted to be natively disordered and to contain

short linear motifs targeting domains such as SH2, SH3, 14-3-3, PWWP or WD40. The first such motif located within the N-terminus of Pragmin and identified by the ELM analysis (Table S6) was the EPIYA sequence, a motif recognized and actively phosphorylated by CSK through its SH2 and protein-kinase domains, respectively. We also detected a proline-rich motif potentially recognized by either PWWP or SH3 domains, and this prediction matches the recently detected binding of CrkII to the equivalent motif present in PEAK3, a recently discovered paralogue of Pragmin [73]. These motifs can explain the detection of Csk and Crk (or its closed homologue CrkL) in the interactome of Pragmin.

In addition, ELM analysis suggested a potential connection of STBX4 with Pragmin through its WW domain while the C-terminus of Csk could be recognized by the PDZ domain of STBX4. This would tightly connect these three proteins in an interleaved manner. Other proline-rich segments exist in Pragmin N-terminus and in its preys while several of the latter (Csk, Crk, CrkL, GRB2 and ABI1) contain SH3 domain, an intricate network may be at play. As an example, ABI1, a subunit of the WAVE regulatory complex possesses both a SH3 domain and a PxxDY motif potentially recognized by a SH3 domain. This may connect the WAVE regulatory complex with the complex formed by Pragmin, CSK and CRK (see Fig S1).
Obviously, numerous experiments will be necessary to confirm (or not) these predicted interactions but we illustrate here how our protocol can provide clues to guide such experimental validations. As a whole, the data returned by our protocol helped to reveal an integrated network converging toward cell migration on one side and oxidative stress response on the other side.

4. DISCUSSION AND PERSPECTIVES

In the present study, we have introduced a full protocol that starts with a list of proteins detected by interatomics experiments and integrates knowledge from the available structures of complexes and predictions of weak interactions to propose a structured overview of the collection of input proteins in terms of their interactions. To overcome the limitation of the rather low number of homomeric and heteromeric non-redundant complex structures resolved to date (~30,000 and 7,700, respectively), our protocol elaborates over a remote homology detection approach.

From the two case studies presented here, the added value of using such remote homology search seems obvious. First, it is striking that we could detect either the proteins or homologs of them for all of the 61 proteins of the Pragmin dataset but one, and for 182 of the 192 proteins of the proteasome dataset, i.e. for close to 96% of the cases. Following, the results obtained from the challenging case of the proteasome test case illustrate well that (i) the quality of the scores of the remote detection distinguish well between orthologs and paralogs, and (ii) prove the ability of our procedure to identify core obligate complexes. Moreover it allows to "hierarchize" the flat list of proteins in terms of interactions, even for sequence identities as low as 20%. Regarding the proteasome dataset, we were indeed able to link together 138 of the 192 proteins, as deduced from 128 identified complex structures that contain at least 2 identical and/or homologous proteins from the input list. For the Pragmin dataset, our results well illustrate the ability of the approach to decipher more subtle cases, such as for instance the existence of mutually exclusive complexes involving different but overlapping subset of proteins. They also illustrate the potential added-value of predicting lower affinity interactions implying SLiMS to further decipher the interactions of the input. Finally, it is also interesting that the protocol was able to return information about homo-oligomers for both test sets.

Compared to other approaches such as Interactome3D, one advantage of our protocol is that it does not depend on any precalculation, and can infer information from different species. One can exemplify this for our two use-cases. For sake of clarity, we focus on the Pragmin case. For this use-case, our method connects 48 input proteins, into a single cluster, through 172 interactions: 21 structure-based, 60 SLiMs-based, and 71 from BioGRID. Interactome3D returned, starting from the 61 proteins, heterocomplex interactions for 27 proteins, through 32 connections: 18 structure-based and 14 others. These interactions define three clusters, which contain 2, 4, and 21 proteins. More in detail, our protocol identified as true positives the connections between all the partners of the ARPC2/3 complex, when Interactome3D, only did so for those between ARP2 and ARPC4, all other connections coming from theoretical models. This difference lies in the fact that the Interactome3D analysis was performed for *H. sapiens*, which is the organism from which the input dataset has been produced. Consequently, Interactome3D did not take into account the 19 structures obtained from *Bos taurus* proteins (which share 100% sequence identity with their human counterparts), which highlights the interest of considering homology detection across species. Interactome3D also identified a connection between AAKG1 and AAKG2, while these proteins are actually two alternative partners of the AMPK complex, as discussed above. With regard to the WAVE regulatory complex, Interactome3D failed to validate the true positive interaction between ABI1 and CYFP1, whereas our protocol identified a link between these two proteins, based on two PDB structures, 4N78 and 3P8C, found with 95.0% and 94.5% sequence identity, respectively.

Considering the STRING database, it returned for Pragmin 142 interactions, connecting 50 of the 61 entries, separated into four clusters of 2, 2, 3, and 43 proteins. STRING identifies the second largest number of connections—2 more than our protocol and 23 more than

Interactome3D—but we recall STRING integrates information sources such as genetic associations and text mining, which do not necessarily correspond to physical interactions. Also, the "protein homology" links established by STRING are of completely different nature than those produced by Interactome3D and our method. Such links actually represent an evolutionary relationship between the two proteins. For example, the PLOD2 and PLOD3 proteins do not physically interact and, yet, STRING connects them, based on their sequence similarity. Results returned by STRING face the same issue as Interactome3D for the AMPK complex, highlighting again the added value of considering structural information. For the WAVE complex, STRING identified the true positive interaction between ABI1 and CYFP1. Interestingly, both STRING and our method detected the genuine interaction between AP2A1 and PP2BA. However, STRING identified this PPI, because it is reported in the databases it integrates. Our method, on the other hand, found it through the search for interologs, which identified 13 PDB structures with 29% sequence identity. Interestingly, STRING does not connect certain partners that our method identified as even closer interologs. For example, the interaction between AP2A1 and 2A5G is supported by 5 PDB structures at 57% identity, and that between TNKS1 and DYR1A is supported by both the 3D-based (36% identity) and SLiMs-based analyses.

Considering the results obtained with GeneMANIA, the 61 input proteins are all connected as a single cluster, through 447 documented and 96 predicted PPIs. The integrative algorithm includes 20 additional nodes to the network. Compared to the other methods, GeneMANIA yields, by far, the highest number of connections. Remarkably, it does not generate any of the aforementioned false negatives. However, the results are nor exempt from false positives. For example, as observed with the Interactome3D and STRING methods, the γ1 and γ2 subunits of

the AMPK complex (AAKG1 and AAKG2, respectively) are wrongly connected. The server also adds the AAKB2 node to the input data and connects it with AAKB1, although these two proteins are mutually exclusive partners of the AMPK complex. A similar example regards the PLOD3 protein mentioned above, which is connected to the newly added PLOD1, despite the absence of any documented interaction between these two proteins. Finally, although the WAVE and Arp2/3 complexes do interact within the Pragmin network, the connections established by GeneMANIA, both within and between these two complexes, are not "hierarchized".

In summary, our protocol completely overlaps and extends the results of Interactome3D, in terms of interactions found in the two datasets. Conversely, it identified fewer associations than STRING or GeneMANIA, but these are more relevant regarding the definition of physical PPIs. The same trends were observed using the proteasome use-case, where STRING connects 163 of the 192 input proteins, GeneMANIA connects all the 192 input proteins as a single cluster (through 3,490 physical and 2,470 predicted interactions), while both our protocol and Interactome3D connect 138 and 90 candidate partners, respectively. Note that because our use-cases corresponded to human sequences, an organism precalculated in the Interactome3D database, similar results obtained from Interactome3D and our protocol were expected.

Altogether, the procedure we introduced here, applied to the two use-cases presented, performed as well or better than a reference approach such as Interactome3D. Furthermore, we expect it to have obvious added value in assisting users to analyse and to explore interactomics data. Notably, our protocol simplifies the identification of subsets of proteins participating to obligate interactions, it finds links between different such blocks, and it proposes putative interactions at a lower affinity involving SLiMS. Possible directions for methodological

improvement could encompass the explicit 3D modeling of the complexes, more detailed analysis of the protein-protein interfaces (e.g. in terms of co-evolution), or establishing connections with emerging technologies such as integrated top-down native proteomics [74,75]. In terms of applications, we believe that our pipeline will be of much use in the near future, due to the increasing number of interactomes performed using *in vivo* protein-proximity labeling (such as BioID[76]) or *in vivo* cross-linking approaches. Further developments of the pipeline could include (i) including quantitative information in the scoring of the (sub)-complexes and result outputs, as already proposed by other tools such as SAINT [77], QPROT [78], COMPASS [79] or TopS [80], and (ii) integrating distance-restraints from cross-linking experiments for homology modelling, as performed by dedicated scoring functions [81] or bioinformatic suites such as the Integrative Modeling Platform (IMP) [82] or HADDOCK2.2 Web server [83].

SUPPORTING INFORMATION

The following supporting information is available free of charge at ACS website http://pubs.acs.org

Figure S1. Sub-network including Pragmin (PRAG1) and the ARP2/3 and WRC complexes.

Pages S-4 to S-9. Markdown-formatted user manual of the stand-alone program.

Table S1. Results of the search for homologs and homo-oligomers, for the proteasome dataset.

Table S2. Results of the 3D-complex detection, for the proteasome dataset.

Table S3. Results of the SLiM-based analysis, for the Pragmin dataset.

Table S4. Results of the search for homologs and homo-oligomers, for the Pragmin dataset.

Table S5. Results of the 3D-complex detection, for the Pragmin dataset.

Table S6. Results of the SLiM-based analysis, for the Pragmin dataset.

ACKNOWLEDGEMENTS

This work was partly supported by the “Investissement d’Avenir Infrastructures Nationales en Biologie et Santé” grants ANR-10-INBS-05-01 (French Infrastructure for Integrated Structural Biology - FRISBI), ANR-10-INBS-08 (Proteomics French Infrastructure - ProFI) and ANR-11-INBS-0013 (French Institute of Bioinformatics - IFB).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

(1) Luck, K.; Sheynkman, G. M.; Zhang, I.; Vidal, M. Proteome-Scale Human Interactomics. *Trends Biochem. Sci.* **2017**, *42* (5), 342–354. https://doi.org/10.1016/j.tibs.2017.02.006.

(2) Rolland, T.; Taşan, M.; Charloteaux, B.; Pevzner, S. J.; Zhong, Q.; Sahni, N.; Yi, S.; Lemmens, I.; Fontanillo, C.; Mosca, R.; Kamburov, A.; Ghiassian, S. D.; Yang, X.; Ghamsari, L.; Balcha, D.; Begg, B. E.; Braun, P.; Brehme, M.; Broly, M. P.; Carvunis, A.-R.; Convery-Zupan, D.; Corominas, R.; Coulombe-Huntington, J.; Dann, E.; Dreze, M.; Dricot, A.; Fan, C.; Franzosa, E.; Gebreab, F.; Gutierrez, B. J.; Hardy, M. F.; Jin, M.; Kang, S.; Kiros, R.; Lin, G. N.; Luck, K.; MacWilliams, A.; Menche, J.; Murray, R. R.; Palagi, A.; Poulin, M. M.; Rambout, X.; Rasla, J.; Reichert, P.; Romero, V.; Ruyssinck, E.; Sahalie, J. M.; Scholz, A.; Shah, A. A.; Sharma, A.; Shen, Y.; Spirohn, K.; Tam, S.; Tejeda, A. O.; Trigg, S. A.; Twizere, J.-C.; Vega, K.; Walsh, J.; Cusick, M. E.; Xia, Y.; Barabási, A.-L.; Iakoucheva, L. M.; Aloy, P.; De Las Rivas, J.; Tavernier, J.; Calderwood, M. A.; Hill, D. E.; Hao, T.; Roth, F. P.; Vidal, M. A Proteome-Scale Map of the Human Interactome Network. *Cell* **2014**, *159* (5), 1212–1226. https://doi.org/10.1016/j.cell.2014.10.050.

(3) Chien, C. T.; Bartel, P. L.; Sternglanz, R.; Fields, S. The Two-Hybrid System: A Method to Identify and Clone Genes for Proteins That Interact with a Protein of Interest. *Proc. Natl. Acad. Sci.* **1991**, *88* (21), 9578–9582. https://doi.org/10.1073/pnas.88.21.9578.

(4) Brückner, A.; Polge, C.; Lentze, N.; Auerbach, D.; Schlattner, U. Yeast Two-Hybrid, a Powerful Tool for Systems Biology. *Int. J. Mol. Sci.* **2009**, *10* (6), 2763–2788. https://doi.org/10.3390/ijms10062763.

(5) Tavernier, J.; Eyckerman, S.; Lemmens, I.; Heyden, J. V. der; Vandekerckhove, J.; Ostade, X. V. MAPPIT: A Cytokine Receptor-Based Two-Hybrid Method in Mammalian Cells1. *Clin. Exp. Allergy* **2002**, *32* (10), 1397–1404. https://doi.org/10.1046/j.1365-2745.2002.01520.x.

(6) Nyfeler, B.; Michnick, S. W.; Hauri, H.-P. Capturing Protein Interactions in the Secretory Pathway of Living Cells. *Proc. Natl. Acad. Sci.* **2005**, *102* (18), 6350–6355. https://doi.org/10.1073/pnas.0501976102.

(7) Bauer, A.; Kuster, B. Affinity Purification-Mass Spectrometry. *Eur. J. Biochem.* **2003**, *270* (4), 570–578. https://doi.org/10.1046/j.1432-1033.2003.03428.x.

(8) Dunham, W. H.; Mullin, M.; Gingras, A.-C. Affinity-Purification Coupled to Mass Spectrometry: Basic Principles and Strategies. *Proteomics* **2012**, *12* (10), 1576–1590. https://doi.org/10.1002/pmic.201100523.

(9) Gingras, A.-C.; Abe, K. T.; Raught, B. Getting to Know the Neighborhood: Using Proximity-Dependent Biotinylation to Characterize Protein Complexes and Map Organelles. *Curr. Opin. Chem. Biol.* **2019**, *48*, 44–54. https://doi.org/10.1016/j.cbpa.2018.10.017.

(10) Fabre, B.; Lambour, T.; Garrigues, L.; Amalric, F.; Vigneron, N.; Menneteau, T.; Stella, A.; Monsarrat, B.; Van den Eynde, B.; Burlet-Schiltz, O.; Bousquet-Dubouch, M.-P. Deciphering Preferential Interactions within Supramolecular Protein Complexes: The Proteasome Case. *Mol. Syst. Biol.* **2015**, *11* (1), 771. https://doi.org/10.15252/msb.20145497.

(11) Chavez, J. D.; Mohr, J. P.; Mathay, M.; Zhong, X.; Keller, A.; Bruce, J. E. Systems Structural Biology Measurements by in Vivo Cross-Linking with Mass Spectrometry.

Nat. Protoc. **2019**, *14* (8), 2318–2343. https://doi.org/10.1038/s41596-019-0181-3.

(12) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

(13) Aloy, P.; Ceulemans, H.; Stark, A.; Russell, R. B. The Relationship Between Sequence and Interaction Divergence in Proteins. *J. Mol. Biol.* **2003**, *332* (5), 989–998. https://doi.org/10.1016/j.jmb.2003.07.006.

(14) Andreani, J.; Guerois, R. Evolution of Protein Interactions: From Interactomes to Interfaces. *Arch. Biochem. Biophys.* **2014**, *554*, 65–75. https://doi.org/10.1016/j.abb.2014.05.010.

(15) Faure, G.; Andreani, J.; Guerois, R. InterEvol Database: Exploring the Structure and Evolution of Protein Complex Interfaces. *Nucleic Acids Res.* **2012**, *40* (D1), D847–D856. https://doi.org/10.1093/nar/gkr845.

(16) Yashiroda, H.; Mizushima, T.; Okamoto, K.; Kameyama, T.; Hayashi, H.; Kishimoto, T.; Niwa, S.; Kasahara, M.; Kurimoto, E.; Sakata, E.; Takagi, K.; Suzuki, A.; Hirano, Y.; Murata, S.; Kato, K.; Yamane, T.; Tanaka, K. Crystal Structure of a Chaperone Complex That Contributes to the Assembly of Yeast 20S Proteasomes. *Nat. Struct. Mol. Biol.* **2008**, *15* (3), 228–236. https://doi.org/10.1038/nsmb.1386.

(17) Gouw, M.; Michael, S.; Sámano-Sánchez, H.; Kumar, M.; Zeke, A.; Lang, B.; Bely, B.; Chemes, L. B.; Davey, N. E.; Deng, Z.; Diella, F.; Gürth, C.-M.; Huber, A.-K.; Kleinsorg, S.; Schlegel, L. S.; Palopoli, N.; Roey, K. V.; Altenberg, B.; Reményi, A.; Dinkel, H.; Gibson, T. J. The Eukaryotic Linear Motif Resource – 2018 Update. *Nucleic Acids Res.* **2018**, *46* (D1), D428–D434. https://doi.org/10.1093/nar/gkx1077.

(18) Mosca, R.; Céol, A.; Aloy, P. Interactome3D: Adding Structural Details to Protein Networks. *Nat. Methods* **2013**, *10* (1), 47–53. https://doi.org/10.1038/nmeth.2289.

(19) Orchard, S.; Ammari, M.; Aranda, B.; Breuza, L.; Briganti, L.; Broackes-Carter, F.; Campbell, N. H.; Chavali, G.; Chen, C.; del-Toro, N.; Duesbury, M.; Dumousseau, M.; Galeota, E.; Hinz, U.; Iannuccelli, M.; Jagannathan, S.; Jimenez, R.; Khadake, J.; Lagreid, A.; Licata, L.; Lovering, R. C.; Meldal, B.; Melidoni, A. N.; Milagros, M.; Peluso, D.; Perfetto, L.; Porras, P.; Raghunath, A.; Ricard-Blum, S.; Roechert, B.; Stutz, A.; Tognolli, M.; van Roey, K.; Cesareni, G.; Hermjakob, H. The MIntAct Project—IntAct as a Common Curation Platform for 11 Molecular Interaction Databases. *Nucleic Acids Res.* **2014**, *42* (D1), D358–D363. https://doi.org/10.1093/nar/gkt1115.

(20) Pieper, U.; Webb, B. M.; Dong, G. Q.; Schneidman-Duhovny, D.; Fan, H.; Kim, S. J.; Khuri, N.; Spill, Y. G.; Weinkam, P.; Hammel, M.; Tainer, J. A.; Nilges, M.; Sali, A. ModBase, a Database of Annotated Comparative Protein Structure Models and Associated Resources. *Nucleic Acids Res.* **2014**, *42* (D1), D336–D346. https://doi.org/10.1093/nar/gkt1144.

(21) 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**, *42* (D1), D374–D379. https://doi.org/10.1093/nar/gkt887.

(22) Webb, B.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinforma.* **2016**, *54* (1), 5.6.1-5.6.37. https://doi.org/10.1002/cpbi.3.

(23) Szklarczyk, D.; Gable, A. L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N. T.; Morris, J. H.; Bork, P.; Jensen, L. J.; Mering, C. von. STRING V11:

Protein–Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets. *Nucleic Acids Res.* **2019**, *47* (D1), D607–D613. https://doi.org/10.1093/nar/gky1131.

(24) Franz, M.; Rodriguez, H.; Lopes, C.; Zuberi, K.; Montojo, J.; Bader, G. D.; Morris, Q. GeneMANIA Update 2018. *Nucleic Acids Res.* **2018**, *46* (W1), W60–W64. https://doi.org/10.1093/nar/gky311.

(25) Oughtred, R.; Stark, C.; Breitkreutz, B.-J.; Rust, J.; Boucher, L.; Chang, C.; Kolas, N.; O'Donnell, L.; Leung, G.; McAdam, R.; Zhang, F.; Dolma, S.; Willems, A.; Coulombe-Huntington, J.; Chatr-aryamontri, A.; Dolinski, K.; Tyers, M. The BioGRID Interaction Database: 2019 Update. *Nucleic Acids Res.* **2019**, *47* (D1), D529–D541. https://doi.org/10.1093/nar/gky1079.

(26) Rodchenkov, I.; Babur, O.; Luna, A.; Aksoy, B. A.; Wong, J. V.; Fong, D.; Franz, M.; Siper, M. C.; Cheung, M.; Wrana, M.; Mistry, H.; Mosier, L.; Dlin, J.; Wen, Q.; O'Callaghan, C.; Li, W.; Elder, G.; Smith, P. T.; Dallago, C.; Cerami, E.; Gross, B.; Dogrusoz, U.; Demir, E.; Bader, G. D.; Sander, C. Pathway Commons 2019 Update: Integration, Analysis and Exploration of Pathway Data. *Nucleic Acids Res.* **2020**, *48* (D1), D489–D497. https://doi.org/10.1093/nar/gkz946.

(27) Pons, J.-L.; Labesse, G. @TOME-2: A New Pipeline for Comparative Modeling of Protein–Ligand Complexes. *Nucleic Acids Res.* **2009**, *37* (suppl_2), W485–W491. https://doi.org/10.1093/nar/gkp368.

(28) Söding, J. Protein Homology Detection by HMM–HMM Comparison. *Bioinformatics* **2005**, *21* (7), 951–960. https://doi.org/10.1093/bioinformatics/bti125.

(29) Konagurthu, A. S.; Whisstock, J. C.; Stuckey, P. J.; Lesk, A. M. MUSTANG: A Multiple Structural Alignment Algorithm. *Proteins Struct. Funct. Bioinforma.* **2006**, *64* (3), 559–574. https://doi.org/10.1002/prot.20921.

(30) Katoh, K.; Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **2013**, *30* (4), 772–780. https://doi.org/10.1093/molbev/mst010.

(31) Kumar, M.; Gouw, M.; Michael, S.; Sámano-Sánchez, H.; Pancsa, R.; Glavina, J.; Diakogianni, A.; Valverde, J. A.; Bukirova, D.; Čalyševa, J.; Palopoli, N.; Davey, N. E.; Chemes, L. B.; Gibson, T. J. ELM—the Eukaryotic Linear Motif Resource in 2020. *Nucleic Acids Res.* **2020**, *48* (D1), D296–D306. https://doi.org/10.1093/nar/gkz1030.

(32) Mészáros, B.; Erdős, G.; Dosztányi, Z. IUPred2A: Context-Dependent Prediction of Protein Disorder as a Function of Redox State and Protein Binding. *Nucleic Acids Res.* **2018**, *46* (W1), W329–W337. https://doi.org/10.1093/nar/gky384.

(33) Fabre, B.; Lambour, T.; Delobel, J.; Amalric, F.; Monsarrat, B.; Burlet-Schiltz, O.; Bousquet-Dubouch, M.-P. Subcellular Distribution and Dynamics of Active Proteasome Complexes Unraveled by a Workflow Combining in Vivo Complex Cross-Linking and Quantitative Proteomics. *Mol. Cell. Proteomics* **2013**, *12* (3), 687–699. https://doi.org/10.1074/mcp.M112.023317.

(34) Lecointre, C.; Simon, V.; Kerneur, C.; Allemand, F.; Fournet, A.; Montarras, I.; Pons, J.-L.; Gelin, M.; Brignatz, C.; Urbach, S.; Labesse, G.; Roche, S. Dimerization of the Pragmin Pseudo-Kinase Regulates Protein Tyrosine Phosphorylation. *Structure* **2018**, *26* (4), 545-554.e4. https://doi.org/10.1016/j.str.2018.01.017.

(35) Leroy, C.; Fialin, C.; Sirvent, A.; Simon, V.; Urbach, S.; Poncet, J.; Robert, B.; Jouin, P.;

Roche, S. Quantitative Phosphoproteomics Reveals a Cluster of Tyrosine Kinases That Mediates Src Invasive Activity in Advanced Colon Carcinoma Cells. *Cancer Res.* **2009**, *69* (6), 2279–2286. https://doi.org/10.1158/0008-5472.CAN-08-2354.

(36) Liu, L.; Phua, Y. W.; Lee, R. S.; Ma, X.; Jenkins, Y.; Novy, K.; Humphrey, E. S.; Chan, H.; Shearer, R.; Ong, P. C.; Dai, W.; Saunders, D. N.; Lucet, I. S.; Daly, R. J. Homo- and Heterotypic Association Regulates Signaling by the SgK269/PEAK1 and SgK223 Pseudokinases. *J. Biol. Chem.* **2016**, *291* (41), 21571–21583. https://doi.org/10.1074/jbc.M116.748897.

(37) Fabre, B.; Lambour, T.; Garrigues, L.; Ducoux-Petit, M.; Amalric, F.; Monsarrat, B.; Burlet-Schiltz, O.; Bousquet-Dubouch, M.-P. Label-Free Quantitative Proteomics Reveals the Dynamics of Proteasome Complexes Composition and Stoichiometry in a Wide Range of Human Cell Lines. *J. Proteome Res.* **2014**, *13* (6), 3027–3037. https://doi.org/10.1021/pr500193k.

(38) Matyskiela, M. E.; Lander, G. C.; Martin, A. Conformational Switching of the 26S Proteasome Enables Substrate Degradation. *Nat. Struct. Mol. Biol.* **2013**, *20* (7), 781–788. https://doi.org/10.1038/nsmb.2616.

(39) Rêgo, A. T.; Fonseca, P. C. A. da. Characterization of Fully Recombinant Human 20S and 20S-PA200 Proteasome Complexes. *Mol. Cell* **2019**, *76* (1), 138-147.e5. https://doi.org/10.1016/j.molcel.2019.07.014.

(40) Wehmer, M.; Rudack, T.; Beck, F.; Aufderheide, A.; Pfeifer, G.; Plitzko, J. M.; Förster, F.; Schulten, K.; Baumeister, W.; Sakata, E. Structural Insights into the Functional Cycle of the ATPase Module of the 26S Proteasome. *Proc. Natl. Acad. Sci.* **2017**. https://doi.org/10.1073/pnas.1621129114.

(41) Huber, E. M.; Basler, M.; Schwab, R.; Heinemeyer, W.; Kirk, C. J.; Groettrup, M.; Groll, M. Immuno- and Constitutive Proteasome Crystal Structures Reveal Differences in Substrate and Inhibitor Specificity. *Cell* **2012**, *148* (4), 727–738. https://doi.org/10.1016/j.cell.2011.12.030.

(42) Ruschak, A. M.; Kay, L. E. Proteasome Allostery as a Population Shift between Interchanging Conformers. *Proc. Natl. Acad. Sci.* **2012**. https://doi.org/10.1073/pnas.1213640109.

(43) Gersch, M.; Hackl, M. W.; Dubiella, C.; Dobrinevski, A.; Groll, M.; Sieber, S. A. A Mass Spectrometry Platform for a Streamlined Investigation of Proteasome Integrity, Posttranslational Modifications, and Inhibitor Binding. *Chem. Biol.* **2015**, *22* (3), 404–411. https://doi.org/10.1016/j.chembiol.2015.01.004.

(44) Lesne, J.; Bousquet, M.-P.; Marcoux, J.; Locard-Paulet, M. Top-Down and Intact Protein Mass Spectrometry Data Visualization for Proteoform Analysis Using VisioProt-MS. *Bioinforma. Biol. Insights* **2019**, *13*, 1177932219868223. https://doi.org/10.1177/1177932219868223.

(45) Jonik-Nowak, B.; Menneteau, T.; Fesquet, D.; Baldin, V.; Bonne-Andrea, C.; Méchali, F.; Fabre, B.; Boisguerin, P.; Rossi, S. de; Henriquet, C.; Pugnière, M.; Ducoux-Petit, M.; Burlet-Schiltz, O.; Lamond, A. I.; Fort, P.; Boulon, S.; Bousquet, M.-P.; Coux, O. PIP30/FAM192A Is a Novel Regulator of the Nuclear Proteasome Activator PA28γ. *Proc. Natl. Acad. Sci.* **2018**, *115* (28), E6477–E6486. https://doi.org/10.1073/pnas.1722299115.

(46) Moscovitz, O.; Ben-Nissan, G.; Fainer, I.; Pollack, D.; Mizrachi, L.; Sharon, M. The

Parkinson's-Associated Protein DJ-1 Regulates the 20S Proteasome. *Nat. Commun.* **2015**, *6* (1), 1–13. https://doi.org/10.1038/ncomms7609.

(47) Olshina, M. A.; Arkind, G.; Kumar Deshmukh, F.; Fainer, I.; Taranavsky, M.; Hayat, D.; Ben-Dor, S.; Ben-Nissan, G.; Sharon, M. Regulation of the 20S Proteasome by a Novel Family of Inhibitory Proteins. *Antioxid. Redox Signal.* **2020**, *32* (9), 636–655. https://doi.org/10.1089/ars.2019.7816.

(48) Lasker, K.; Förster, F.; Bohn, S.; Walzthoeni, T.; Villa, E.; Unverdorben, P.; Beck, F.; Aebersold, R.; Sali, A.; Baumeister, W. Molecular Architecture of the 26S Proteasome Holocomplex Determined by an Integrative Approach. *Proc. Natl. Acad. Sci.* **2012**. https://doi.org/10.1073/pnas.1120559109.

(49) Wang, X.; Cimermancic, P.; Yu, C.; Schweitzer, A.; Chopra, N.; Engel, J. L.; Greenberg, C. H.; Huszagh, A. S.; Beck, F.; Sakata, E.; Yang, Y.; Novitsky, E. J.; Leitner, A.; Nanni, P.; Kahraman, A.; Guo, X.; Dixon, J. E.; Rychnovsky, S. D.; Aebersold, R.; Baumeister, W.; Sali, A.; Huang, L. Molecular Details Underlying Dynamic Structures and Regulation of the Human 26S Proteasome. *Mol. Cell. Proteomics* **2017**. https://doi.org/10.1074/mcp.M116.065326.

(50) Wang, X.; Chemmama, I. E.; Yu, C.; Huszagh, A.; Xu, Y.; Viner, R.; Block, S. A.; Cimermancic, P.; Rychnovsky, S. D.; Ye, Y.; Sali, A.; Huang, L. The Proteasome-Interacting Ecm29 Protein Disassembles the 26S Proteasome in Response to Oxidative Stress. *J. Biol. Chem.* **2017**, jbc.M117.803619. https://doi.org/10.1074/jbc.M117.803619.

(51) Wang, X.; Huang, L. Dissecting Dynamic and Heterogeneous Proteasome Complexes Using In Vivo Cross-Linking-Assisted Affinity Purification and Mass Spectrometry. In *The Ubiquitin Proteasome System: Methods and Protocols*; Mayor, T., Kleiger, G., Eds.; Methods in Molecular Biology; Springer: New York, NY, 2018; pp 401–410. https://doi.org/10.1007/978-1-4939-8706-1_25.

(52) Huber, E. M.; Groll, M. The Mammalian Proteasome Activator PA28 Forms an Asymmetric A4β3 Complex. *Struct. Lond. Engl. 1993* **2017**, *25* (10), 1473-1480.e3. https://doi.org/10.1016/j.str.2017.07.013.

(53) Groll, M.; Brandstetter, H.; Bartunik, H.; Bourenkow, G.; Huber, R. Investigations on the Maturation and Regulation of Archaebacterial Proteasomes††We Dedicate This Paper to the Memory of Eraldo Antonini, Eminent Biochemist, Prematurely Deceased 20 Years Ago, on 19 March 1983. *J. Mol. Biol.* **2003**, *327* (1), 75–83. https://doi.org/10.1016/S0022-2836(03)00080-9.

(54) Huang, X.; Luan, B.; Wu, J.; Shi, Y. An Atomic Structure of the Human 26S Proteasome. *Nat. Struct. Mol. Biol.* **2016**, *23* (9), 778–785. https://doi.org/10.1038/nsmb.3273.

(55) Harshbarger, W.; Miller, C.; Diedrich, C.; Sacchettini, J. Crystal Structure of the Human 20S Proteasome in Complex with Carfilzomib. *Structure* **2015**, *23* (2), 418–424. https://doi.org/10.1016/j.str.2014.11.017.

(56) Santos, R. de L. A.; Bai, L.; Singh, P. K.; Murakami, N.; Fan, H.; Zhan, W.; Zhu, Y.; Jiang, X.; Zhang, K.; Assker, J. P.; Nathan, C. F.; Li, H.; Azzi, J.; Lin, G. Structure of Human Immunoproteasome with a Reversible and Noncompetitive Inhibitor That Selectively Inhibits Activated Lymphocytes. *Nat. Commun.* **2017**, *8* (1), 1–11. https://doi.org/10.1038/s41467-017-01760-5.

(57) Sahtoe, D. D.; van Dijk, W. J.; El Oualid, F.; Ekkebus, R.; Ovaa, H.; Sixma, T. K.

Mechanism of UCH-L5 Activation and Inhibition by DEUBAD Domains in RPN13 and INO80G. *Mol. Cell* **2015**, *57* (5), 887–900. https://doi.org/10.1016/j.molcel.2014.12.039.

(58) Keren-Kaplan, T.; Peters, L. Z.; Levin-Kravets, O.; Attali, I.; Kleifeld, O.; Shohat, N.; Artzi, S.; Zucker, O.; Pilzer, I.; Reis, N.; Glickman, M. H.; Ben-Aroya, S.; Prag, G. Structure of Ubiquitylated-Rpn10 Provides Insight into Its Autoregulation Mechanism. *Nat. Commun.* **2016**, *7* (1), 1–12. https://doi.org/10.1038/ncomms12960.

(59) Pathare, G. R.; Nagy, I.; Śledź, P.; Anderson, D. J.; Zhou, H.-J.; Pardon, E.; Steyaert, J.; Förster, F.; Bracher, A.; Baumeister, W. Crystal Structure of the Proteasomal Deubiquitylation Module Rpn8-Rpn11. *Proc. Natl. Acad. Sci.* **2014**, *111* (8), 2984–2989. https://doi.org/10.1073/pnas.1400546111.

(60) Zhang, W.; Wu, K.-P.; Sartori, M. A.; Kamadurai, H. B.; Ordureau, A.; Jiang, C.; Mercredi, P. Y.; Murchie, R.; Hu, J.; Persaud, A.; Mukherjee, M.; Li, N.; Doye, A.; Walker, J. R.; Sheng, Y.; Hao, Z.; Li, Y.; Brown, K. R.; Lemichez, E.; Chen, J.; Tong, Y.; Harper, J. W.; Moffat, J.; Rotin, D.; Schulman, B. A.; Sidhu, S. S. System-Wide Modulation of HECT E3 Ligases with Selective Ubiquitin Variant Probes. *Mol. Cell* **2016**, *62* (1), 121–136. https://doi.org/10.1016/j.molcel.2016.02.005.

(61) Ernst, A.; Avvakumov, G.; Tong, J.; Fan, Y.; Zhao, Y.; Alberts, P.; Persaud, A.; Walker, J. R.; Neculai, A.-M.; Neculai, D.; Vorobyov, A.; Garg, P.; Beatty, L.; Chan, P.-K.; Juang, Y.-C.; Landry, M.-C.; Yeh, C.; Zeqiraj, E.; Karamboulas, K.; Allali-Hassani, A.; Vedadi, M.; Tyers, M.; Moffat, J.; Sicheri, F.; Pelletier, L.; Durocher, D.; Raught, B.; Rotin, D.; Yang, J.; Moran, M. F.; Dhe-Paganon, S.; Sidhu, S. S. A Strategy for Modulation of Enzymes in the Ubiquitin System. *Science* **2013**, *339* (6119), 590–595. https://doi.org/10.1126/science.1230161.

(62) Stadtmueller, B. M.; Kish-Trier, E.; Ferrell, K.; Petersen, C. N.; Robinson, H.; Myszka, D. G.; Eckert, D. M.; Formosa, T.; Hill, C. P. Structure of a Proteasome Pba1-Pba2 Complex IMPLICATIONS FOR PROTEASOME ASSEMBLY, ACTIVATION, AND BIOLOGICAL FUNCTION. *J. Biol. Chem.* **2012**, *287* (44), 37371–37382. https://doi.org/10.1074/jbc.M112.367003.

(63) Wolf-Levy, H.; Javitt, A.; Eisenberg-Lerner, A.; Kacen, A.; Ulman, A.; Sheban, D.; Dassa, B.; Fishbain-Yoskovitz, V.; Carmona-Rivera, C.; Kramer, M. P.; Nudel, N.; Regev, I.; Zahavi, L.; Elinger, D.; Kaplan, M. J.; Morgenstern, D.; Levin, Y.; Merbl, Y. Revealing the Cellular Degradome by Mass Spectrometry Analysis of Proteasome-Cleaved Peptides. *Nat. Biotechnol.* **2018**, *36* (11), 1110–1116. https://doi.org/10.1038/nbt.4279.

(64) O'Rourke, R. L.; Daly, R. J. The Pseudokinases SgK269 and SgK223: A Novel Oncogenic Alliance in Human Cancer. *Cell Adhes. Migr.* **2018**, *12* (6), 524–528. https://doi.org/10.1080/19336918.2017.1394570.

(65) Senda, Y.; Murata-Kamiya, N.; Hatakeyama, M. C-Terminal Src Kinase-Mediated EPIYA Phosphorylation of Pragmin Creates a Feed-Forward C-Terminal Src Kinase Activation Loop That Promotes Cell Motility. *Cancer Sci.* **2016**, *107* (7), 972–980. https://doi.org/10.1111/cas.12962.

(66) Chen, Z.; Borek, D.; Padrick, S. B.; Gomez, T. S.; Metlagel, Z.; Ismail, A. M.; Umetani, J.; Billadeau, D. D.; Otwinowski, Z.; Rosen, M. K. Structure and Control of the Actin Regulatory WAVE Complex. *Nature* **2010**, *468* (7323), 533–538. https://doi.org/10.1038/nature09623.

(67) Chen, B.; Brinkmann, K.; Chen, Z.; Pak, C. W.; Liao, Y.; Shi, S.; Henry, L.; Grishin, N. V.; Bogdan, S.; Rosen, M. K. The WAVE Regulatory Complex Links Diverse Receptors to the Actin Cytoskeleton. *Cell* **2014**, *156* (1), 195–207. https://doi.org/10.1016/j.cell.2013.11.048.

(68) Robinson, R. C.; Turbedsky, K.; Kaiser, D. A.; Marchand, J.-B.; Higgs, H. N.; Choe, S.; Pollard, T. D. Crystal Structure of Arp2/3 Complex. *Science* **2001**, *294* (5547), 1679–1684. https://doi.org/10.1126/science.1066333.

(69) Nolen, B. J.; Tomasevic, N.; Russell, A.; Pierce, D. W.; Jia, Z.; McCormick, C. D.; Hartman, J.; Sakowicz, R.; Pollard, T. D. Characterization of Two Classes of Small Molecule Inhibitors of Arp2/3 Complex. *Nature* **2009**, *460* (7258), 1031–1034. https://doi.org/10.1038/nature08231.

(70) Werfel, T. A.; Elion, D. L.; Rahman, B.; Hicks, D. J.; Sanchez, V.; Gonzalez-Ericsson, P. I.; Nixon, M. J.; James, J. L.; Balko, J. M.; Scherle, P.; Koblish, H. K.; Cook, R. S. Treatment-Induced Tumor Cell Apoptosis and Secondary Necrosis Drive Tumor Progression in the Residual Tumor Microenvironment through MerTK and IDO-1. *Cancer Res.* **2018**. https://doi.org/10.1158/0008-5472.CAN-18-1106.

(71) Shomura, Y.; Dragovic, Z.; Chang, H.-C.; Tzvetkov, N.; Young, J. C.; Brodsky, J. L.; Guerriero, V.; Hartl, F. U.; Bracher, A. Regulation of Hsp70 Function by HspBP1: Structural Analysis Reveals an Alternate Mechanism for Hsp70 Nucleotide Exchange. *Mol. Cell* **2005**, *17* (3), 367–379. https://doi.org/10.1016/j.molcel.2004.12.023.

(72) Schuermann, J. P.; Jiang, J.; Cuellar, J.; Llorca, O.; Wang, L.; Gimenez, L. E.; Jin, S.; Taylor, A. B.; Demeler, B.; Morano, K. A.; Hart, P. J.; Valpuesta, J. M.; Lafer, E. M.; Sousa, R. Structure of the Hsp110:Hsc70 Nucleotide Exchange Machine. *Mol. Cell* **2008**, *31* (2), 232–243. https://doi.org/10.1016/j.molcel.2008.05.006.

(73) Lopez, M. L.; Lo, M.; Kung, J. E.; Dudkiewicz, M.; Jang, G. M.; Dollen, J. V.; Johnson, J. R.; Krogan, N. J.; Pawłowski, K.; Jura, N. PEAK3/C19orf35 Pseudokinase, a New NFK3 Kinase Family Member, Inhibits CrkII through Dimerization. *Proc. Natl. Acad. Sci.* **2019**, *116* (31), 15495–15504. https://doi.org/10.1073/pnas.1906360116.

(74) Li, H.; Nguyen, H. H.; Ogorzalek Loo, R. R.; Campuzano, I. D. G.; Loo, J. A. An Integrated Native Mass Spectrometry and Top-down Proteomics Method That Connects Sequence to Structure and Function of Macromolecular Complexes. *Nat. Chem.* **2018**, *10* (2), 139–148. https://doi.org/10.1038/nchem.2908.

(75) Skinner, O. S.; Haverland, N. A.; Fornelli, L.; Melani, R. D.; Do Vale, L. H. F.; Seckler, H. S.; Doubleday, P. F.; Schachner, L. F.; Srzentić, K.; Kelleher, N. L.; Compton, P. D. Top-down Characterization of Endogenous Protein Complexes with Native Proteomics. *Nat. Chem. Biol.* **2018**, *14* (1), 36–41. https://doi.org/10.1038/nchembio.2515.

(76) Lambert, J.-P.; Tucholska, M.; Go, C.; Knight, J. D. R.; Gingras, A.-C. Proximity Biotinylation and Affinity Purification Are Complementary Approaches for the Interactome Mapping of Chromatin-Associated Protein Complexes. *J. Proteomics* **2015**, *118*, 81–94. https://doi.org/10.1016/j.jprot.2014.09.011.

(77) Choi, H.; Larsen, B.; Lin, Z.-Y.; Breitkreutz, A.; Mellacheruvu, D.; Fermin, D.; Qin, Z. S.; Tyers, M.; Gingras, A.-C.; Nesvizhskii, A. I. SAINT: Probabilistic Scoring of Affinity Purification–Mass Spectrometry Data. *Nat. Methods* **2011**, *8* (1), 70–73. https://doi.org/10.1038/nmeth.1541.

(78) Choi, H.; Kim, S.; Fermin, D.; Tsou, C.-C.; Nesvizhskii, A. I. QPROT: Statistical Method

for Testing Differential Expression Using Protein-Level Intensity Data in Label-Free Quantitative Proteomics. *J. Proteomics* **2015**, *129*, 121–126. https://doi.org/10.1016/j.jprot.2015.07.036.

(79) Wenger, C. D.; Phanstiel, D. H.; Lee, M. V.; Bailey, D. J.; Coon, J. J. COMPASS: A Suite of Pre- and Post-Search Proteomics Software Tools for OMSSA. *PROTEOMICS* **2011**, *11* (6), 1064–1074. https://doi.org/10.1002/pmic.201000616.

(80) Sardiu, M. E.; Gilmore, J. M.; Groppe, B. D.; Dutta, A.; Florens, L.; Washburn, M. P. Topological Scoring of Protein Interaction Networks. *Nat. Commun.* **2019**, *10* (1), 1–14. https://doi.org/10.1038/s41467-019-09123-y.

(81) Bullock, J. M. A.; Sen, N.; Thalassinos, K.; Topf, M. Modeling Protein Complexes Using Restraints from Crosslinking Mass Spectrometry. *Structure* **2018**, *26* (7), 1015-1024.e2. https://doi.org/10.1016/j.str.2018.04.016.

(82) Webb, B.; Lasker, K.; Velázquez-Muriel, J.; Schneidman-Duhovny, D.; Pellarin, R.; Bonomi, M.; Greenberg, C.; Raveh, B.; Tjioe, E.; Russel, D.; Sali, A. Modeling of Proteins and Their Assemblies with the Integrative Modeling Platform. In *Structural Genomics: General Applications*; Chen, Y. W., Ed.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2014; pp 277–295. https://doi.org/10.1007/978-1-62703-691-7_20.

(83) van Zundert, G. C. P.; Rodrigues, J. P. G. L. M.; Trellet, M.; Schmitz, C.; Kastritis, P. L.; Karaca, E.; Melquiond, A. S. J.; van Dijk, M.; de Vries, S. J.; Bonvin, A. M. J. J. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **2016**, *428* (4), 720–725. https://doi.org/10.1016/j.jmb.2015.09.014.

(84) Kazakou, K.; Holloway, D. E.; Prior, S. H.; Subramanian, V.; Acharya, K. R. Ribonuclease A Homologues of the Zebrafish: Polymorphism, Crystal Structures of Two Representatives and Their Evolutionary Implications. *J. Mol. Biol.* **2008**, *380* (1), 206–222. https://doi.org/10.1016/j.jmb.2008.04.070.

(85) Xiao, B.; Sanders, M. J.; Carmena, D.; Bright, N. J.; Haire, L. F.; Underwood, E.; Patel, B. R.; Heath, R. B.; Walker, P. A.; Hallen, S.; Giordanetto, F.; Martin, S. R.; Carling, D.; Gamblin, S. J. Structural Basis of AMPK Regulation by Small Molecule Activators. *Nat. Commun.* **2013**, *4* (1), 1–10. https://doi.org/10.1038/ncomms4017.

Figure captions:

Figure 1. Different cases introducing true, false positive, and false negative interactions from the flat interactomics data. Case 1: the obligate complex includes one partner that is missing in the flat list, resulting in false negative interactions (black links). Only the links in red (true positive interactions) involve proteins of the input list. The black links cannot be inferred from the input. Case 2: One protein of the input interacts with several partners, but in a mutually exclusive way. No interaction between those partners (blue and purple) exists in reality (false positive interaction).

Figure 2. Flowchart describing the integrative pipeline for the detection of protein complexes. Steps 1 to 4, described in the Results section, are represented by rectangles. 1: The search for homologs of each of the input proteins. 2: Assembling information of step 1 to identify hetero-multimers and homo-multimers. 3: Search for Short Linear Motifs (SLiMs) 4: links to BioGRID.

Figure 3. Illustration of the subunits detected in the proteasome sample and assigned to homologous complexes of known structures in the PDB. (**A**) Complex ID c035 corresponds to an assembly detected for 15 subunits of the sample matching the structure of Blm10 activator bound to the 20S proteasome (PDB:4V7O[84]) in yeast. The 14 subunits of the 20S proteasome and PSME4 in human (PA200) were assigned to a specific chain in the PDB structure. Noticeably, orthologous relationships between human and yeast subunits were automatically assigned in a correct manner despite the number of potentially misleading paralogous relationships. Moreover, ortholog of PA200 in yeast, named Blm10, was detected despite the high sequence divergence between both homologs sharing only 20 % sequence identity. (**B**)

Complex ID c074 identified a homologous complex structure for a set of subunits partly overlapping those in (A), namely the 20S subunits. Remote homology relationships could be detected at even higher sequence divergence between two proteasome chaperones (PSMG1 and PSMG2) for which a structure of the orthologous subunits was available in yeast (PDB:4G4S). (**C**) Complex ID c097 identified a homologous complex for two subunits involving homologs of PSA5 and PSMG3, the latter sharing high sequence divergence (19 % sequence identity) with its yeast ortholog, named POC3. Interestingly in the yeast complex (PDB:2Z5C), a third subunit, named POC4, is found tightly bound suggesting that a homologous subunit could have been detected in human. The remote homologs of yeast POC4, named PSMG4 in human, was not present in the list of input proteins suggesting potential issues in the detection threshold or in the capacity of spotting out this protein by mass spectrometry.

Figure 4. Graph representation of the PPI network identified for the proteasome dataset. Nodes colored in cyan, magenta and orange correspond to input proteins, nodes added by the PDB search and BioGRID, respectively. The bait protein (PSA2) is represented by a green node. Four types of edges represent interactions between proteins. (i) Thick, opaque: Edges identified by the structure-based analysis and connecting input proteins belonging to hetero-multimers. The colors (blue, green, yellow, red, black) depend on the evolutionary distance (sequence identity ≥95%, 80%, 50%, 30%, 0%, respectively) between the two input proteins and their homologs/paralogs. (ii) Thin, magenta: Edges produced by the structure-based analysis and connecting input proteins and undetected partners. (iii) Thin, blue: Edges produced by the ELM analysis and connecting input proteins. (iv) Thin and orange: Edges produced by the BioGRID analysis and connecting input proteins and additional partners from BioGRID.

Figure 5. Cryo-EM structure (PDB: 5GJQ) of the 20S catalytic core (blue) in complex with the 19S regulator (green) forming the ATP-ase dependent 26S proteasome bound to USP14 & UBB (purple). The two copies of the immuno-purification bait (α2) are represented in red. All these 33 proteins were immuno-purified. Structures of the 20S (red and blue), the 19S (green), the 26S (red, blue and green) and the 26S bound to USP14 & UBB (red, blue, green and purple) can be found in the PDB, making possible to distinguish three layers of interaction with the bait.

Figure 6. Graph representation of the PPI network identified for the Pragmin dataset. For this representation, the bait protein (PRAG1) has been added to the original dataset (green node).

Figure 7. Space-filling representation of protein complexes of the Pramgin use-case, as identified by the pipeline. (**A**) The pentameric structure of the WAVE regulatory complex (PDB entry 3p8c), with the subunits WASF1 and BRK1, missing in AP-MS data, colored in white and black, respectively. (**B**) The heptameric structure of the Arp2/3 complex (PDB entry 1k8k), with the subunits ARPC1B (white), ARPC3 (black), and ARPC5 (gray). (**C**) The α1β2γ1 and (**D**) α2β1γ1 isoforms of the AMPK complex, both found by our integrative approach (PDB entries 2vq8[84] and 4cfe[85], respectively); both structures are aligned on the γ1 subunit (colored in purple).

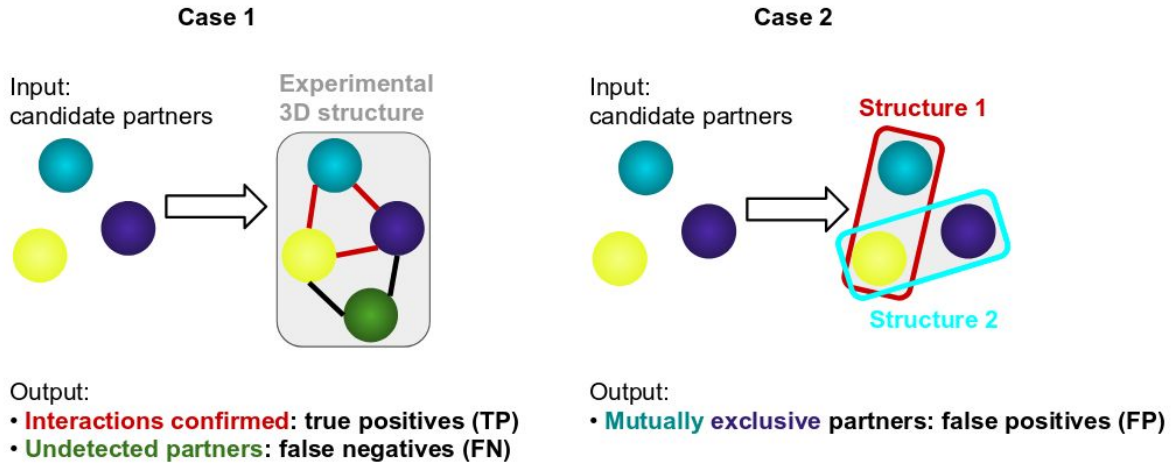

Case 1
Input:
candidate partners
Experimental
3D structure
Output:
• Interactions confirmed: true positives (TP)
• Undetected partners: false negatives (FN)
Case 2
Input:
candidate partners
Structure 1
Structure 2
Output:
• Mutually exclusive partners: false positives (FP)

Figure 1

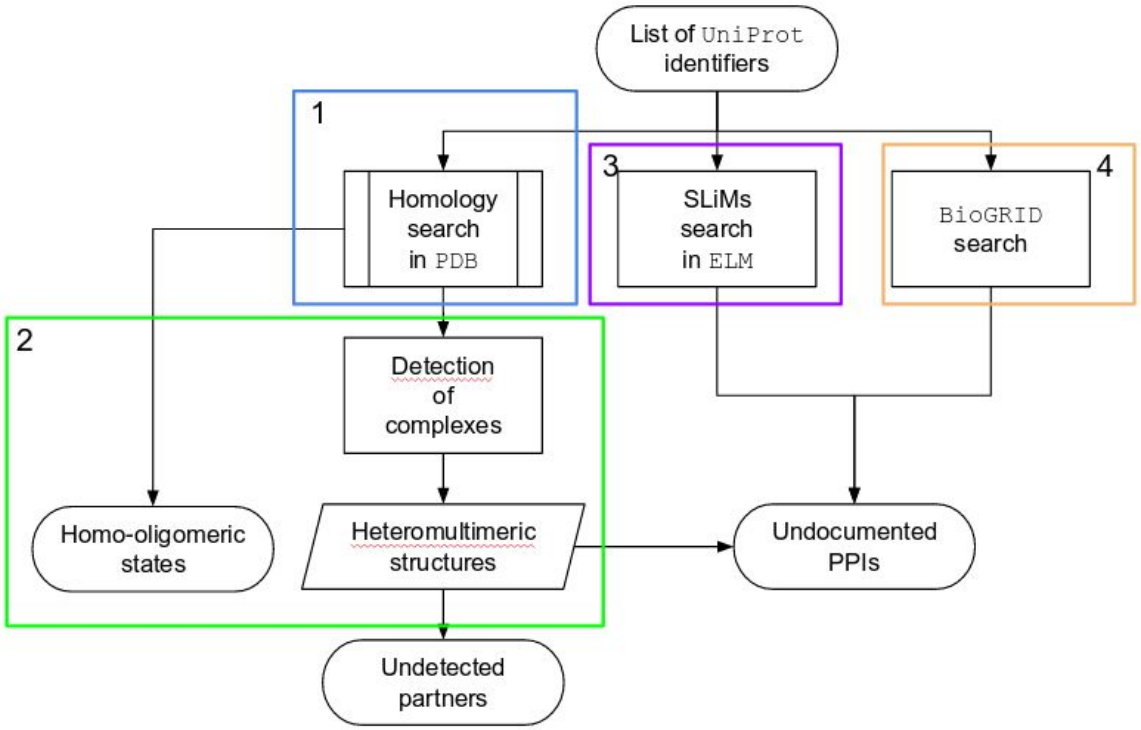
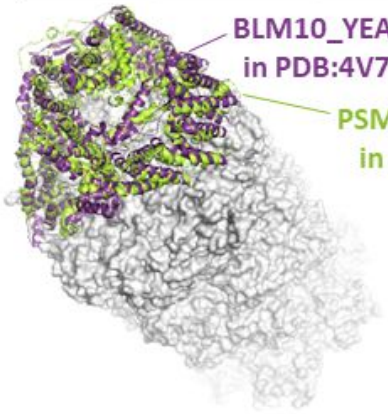

List of UniProt identifiers
1
Homology search in PDB
3
SLiMs search in ELM
4
BioGRID search
2
Detection of complexes
Homo-oligomeric states
Heteromultimeric structures
Undocumented PPIs
Undetected partners

Figure 2

A

Complex ID	Max avg id%	Inputs found	Structures found
c035	51.6	15	4V7O

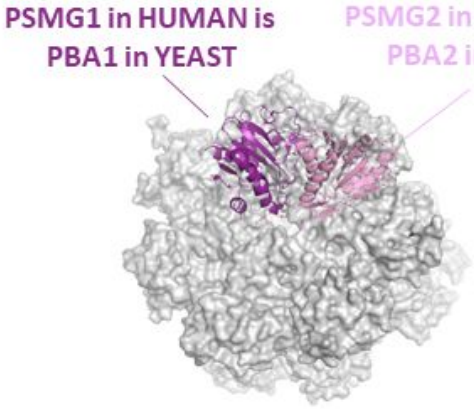


Validated in PDB (6KWY & 6REY) →

	Subunit	% Seq identity	Chain assigned
1	PSB5	65	A2
2	PSA5	61	AJ
3	PSA2	60	AG
4	PSB7	58	AM
5	PSA7	58	AI
6	PSA6	55	AA
7	PSB3	54	AN
8	PSB6	54	AB
9	PSA1	53	AK
10	PSA4	50	AH
11	PSA3	50	AL
12	PSB1	48	A3
13	PSB4	44	A4
14	PSB2	44	A1
15	PSME4	20	A6

B

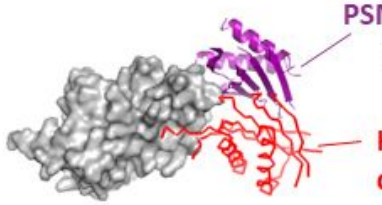
Complex ID	Max avg id%	Inputs found	Structures found
c074	50	16	4G4S



	Subunit	% Seq identity	Chain assigned
1	PSB5	65	L
2	PSA2	62	B
3	PSA5	61	E
4	PSA7	60	D
5	PSB7	58	I
6	PSA6	55	A
7	PSB6	55	H
8	PSB3	54	J
9	PSA1	53	F
10	PSA4	50	C
11	PSA3	50	G
12	PSB1	48	M
13	PSB2	45	K
14	PSB4	43	N
15	PSMG2	21	P
16	PSMG1	18	O

C

Complex ID	Max avg id%	Inputs found	Structures found
c097	37.5	2	2Z5C



	Subunit	% Seq identity	Chain assigned
1	PSA5	56	C
2	PSMG3	19	B
x	x	x	x

Figure 3

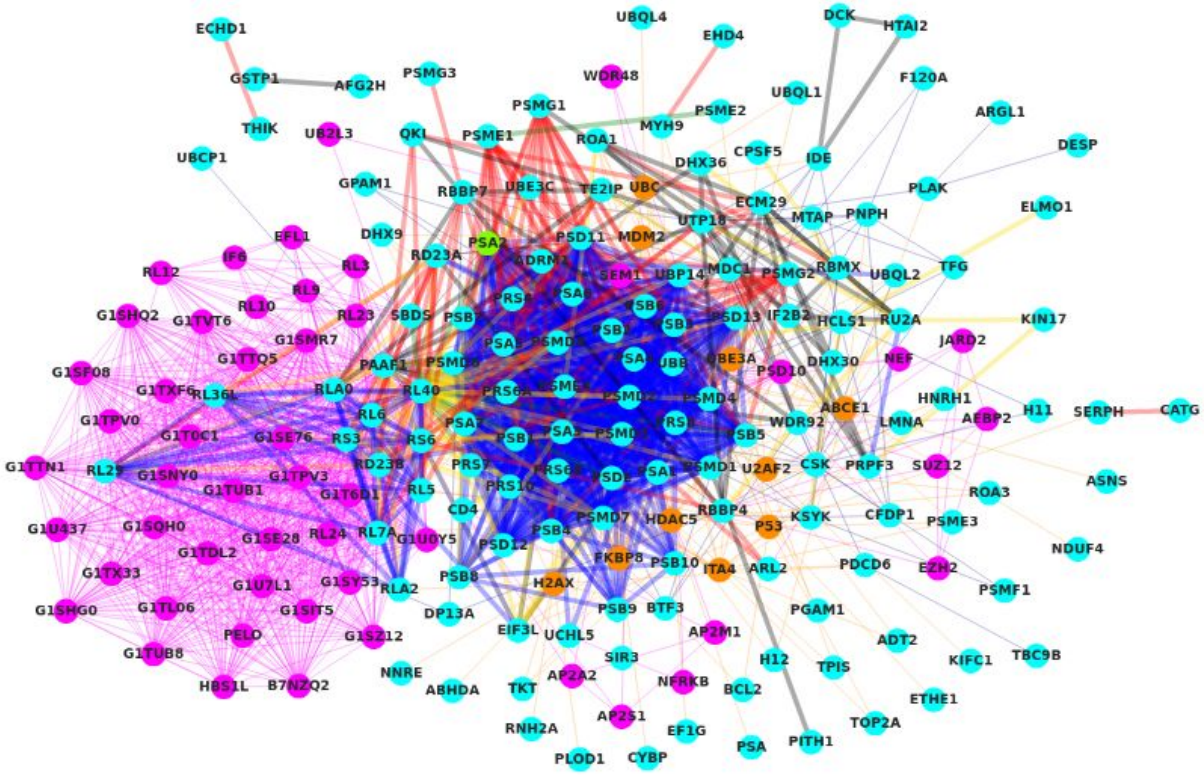

ECHD1
UBQL4
EHD4
DCK
HTAI2
GSTP1
AFG2H
PSMG3
WDR48
F120A
UBQL1
PSMG1
PSME2
ARGL1
THIK
UB2L3
QKI
PSME1
ROA1
MYH9
UBCP1
DHX36
CPSF5
IDE
DESP
GPAM1
RBBP7
UBE3C
TE2IP
UBC
PLAK
ECM29
UTP18
MTAP
PNPH
ELMO1
EFL1
DHX9
PSA2
PSD11
MDM2
IF6
RD23A
ADRM1
SEM1
UBP14
MDC1
PSMG2
RBMX
UBQL2
TFG
RL12
RL3
RL9
PRS8
PSA4
G1SHQ2
RL10
G1TVT6
RL23
SBDS
PSB1
PSB6
PSD13
IF2B2
HCLS1
RU2A
KIN17
G1SMR7
PSA1
PSB3
PSB7
JARD2
G1TTQ5
PAAF1
PSMD8
PSA6
UBE3A
DHX30
NEF
G1SF08
PSD10
G1TXF6
RL36L
RLA0
RL40
PRS6A
PSME4
PSMD4
HNRH1
RL6
ABCE1
WDR92
LMNA
AFBP2
H11
SERPH
CATG
G1TPV0
G1T0C1
G1SE76
RS3
RS6
PSB2
PSA3
PSM0
PRS4
PSB5
G1TTN1
RL29
G1SNY0
G1TPV3
RD23B
PRS7
PRS8
PSD1
PSA5
PSMD1
U2AF2
CSK
PRPF3
SUZ12
ASNS
G1TUB1
G1T6D1
RL5
PRS10
ROA3
CD4
PSMD7
HDAC5
RBBP4
G1U437
G15QH0
RL24
RL7A
G1U0Y5
PSB4
P53
KSYK
CFDP1
PSME3
G1SE28
PSD12
G1TDL2
FKBP8
PSB10
ITA4
ARL2
PDCD6
EZH2
NDUF4
G1TX33
G1U7L1
G1SY53
RLA2
PSB8
H2AX
PSMF1
G1SHG0
G1TL06
G1SIT5
DP13A
PSB9
BTF3
PGAM1
PELO
G1SZ12
EIF3L
UCHL5
AP2M1
ADT2
G1TUB8
SIR3
H12
TPIS
KIFC1
TBC9B
NNRE
HBS1L
B7NZQ2
ABHDA
AP2A2
TKT
NFRKB
BCL2
ETHE1
AP2S1
TOP2A
RNH2A
EF1G
PSA
PITH1
PLOD1
CYBP

Figure 4

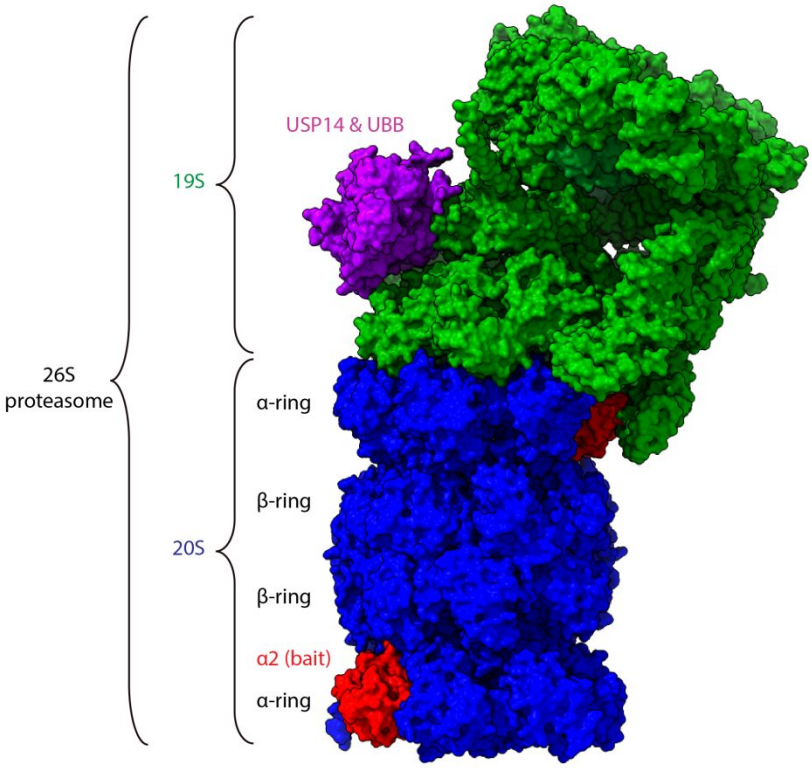

USP14 & UBB
19S
26S
proteasome
α-ring
β-ring
20S
β-ring
α2 (bait)
α-ring

Figure 5

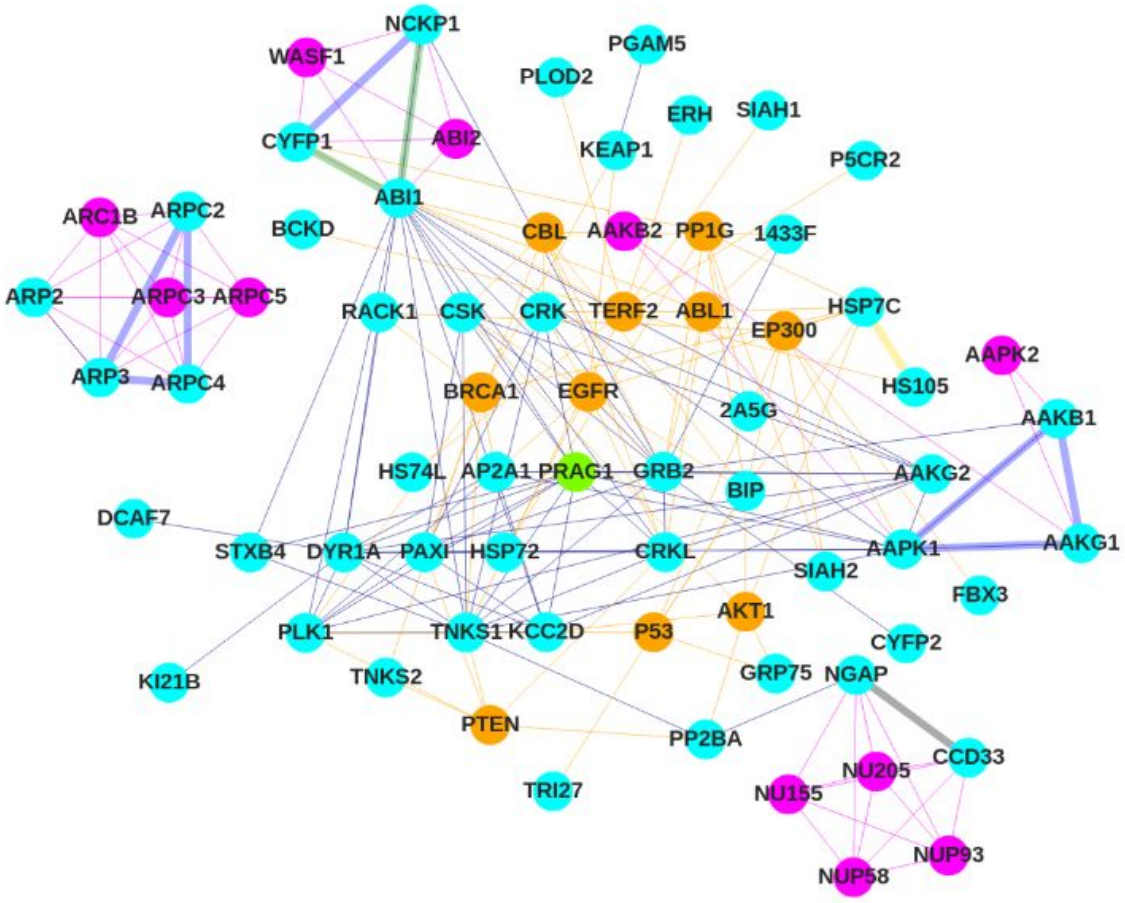

NCKP1
WASF1
PGAM5
PLOD2
ERH
SIAH1
CYFP1
ABI2
KEAP1
P5CR2
ABI1
ARC1B
ARPC2
BCKD
CBL
AAKB2
PP1G
1433F
ARP2
ARPC3
ARPC5
RACK1
CSK
CRK
TERF2
ABL1
HSP7C
EP300
AAPK2
ARP3
ARPC4
BRCA1
EGFR
HS105
2A5G
AAKB1
HS74L
AP2A1
PRAG1
GRB2
AAKG2
BIP
DCAF7
STXB4
DYR1A
PAXI
HSP72
CRKL
AAPK1
AAKG1
SIAH2
FBX3
AKT1
PLK1
TNKS1
KCC2D
P53
CYFP2
KI21B
TNKS2
GRP75
NGAP
PTEN
PP2BA
CCD33
NU205
TRI27
NU155
NUP93
NUP58

Figure 6

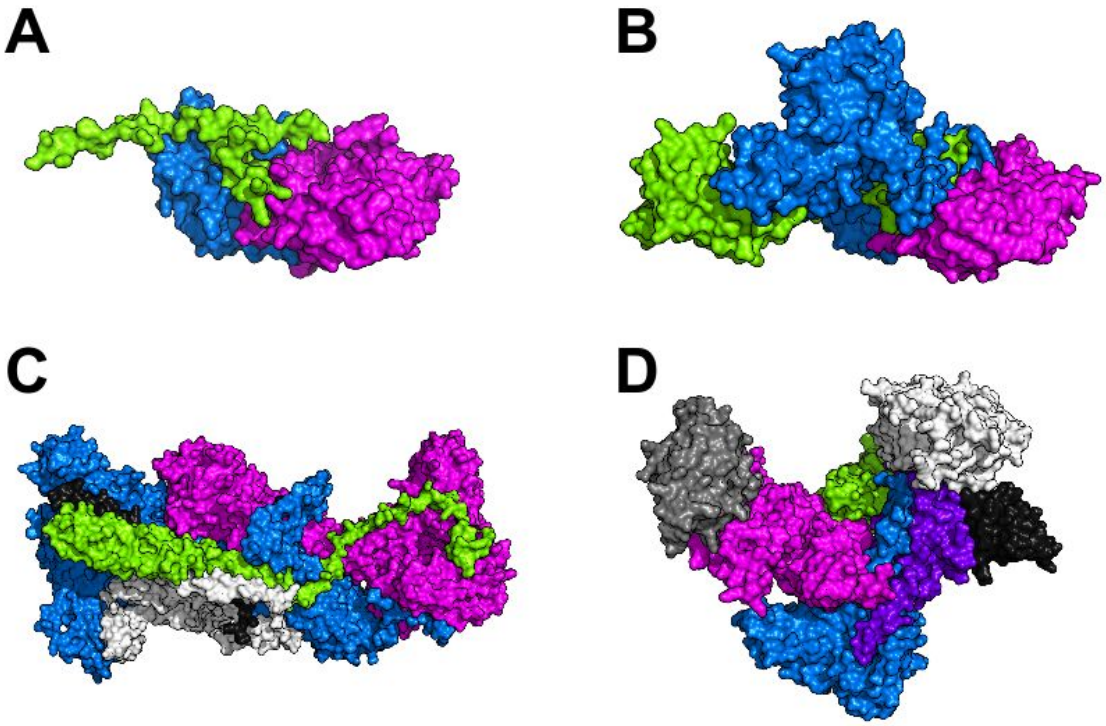
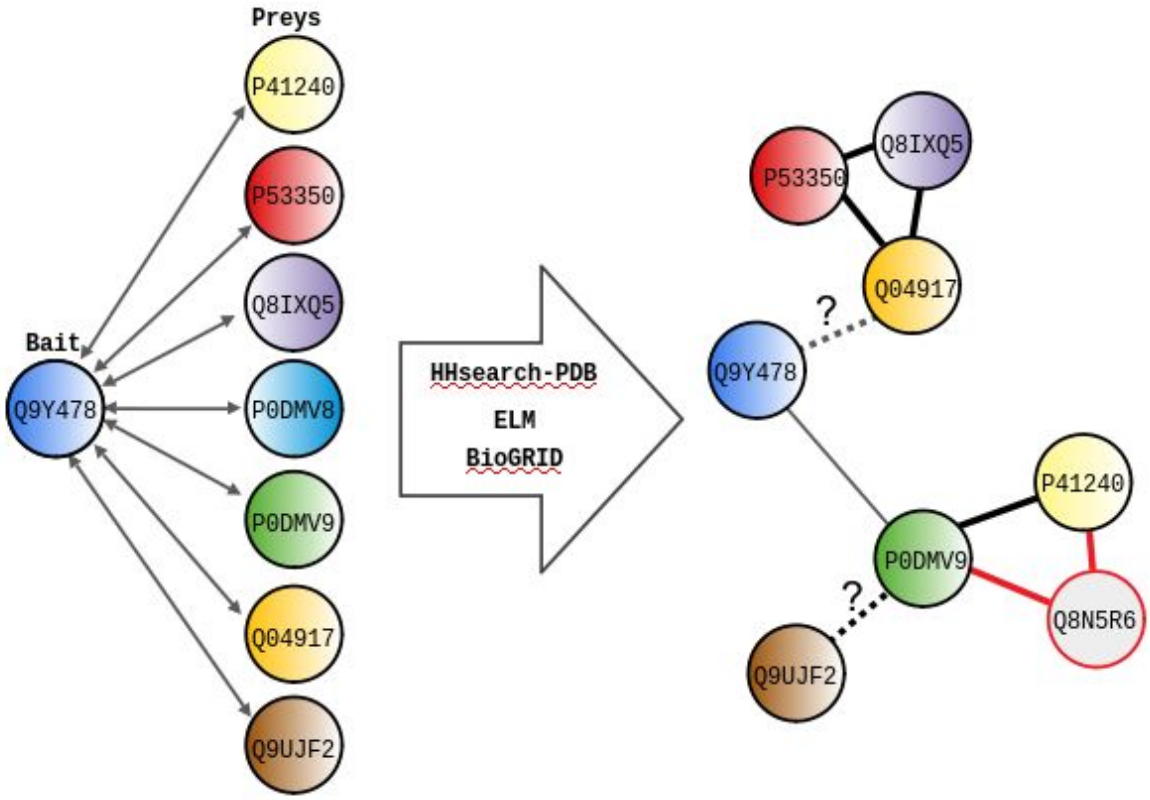


Figure 7

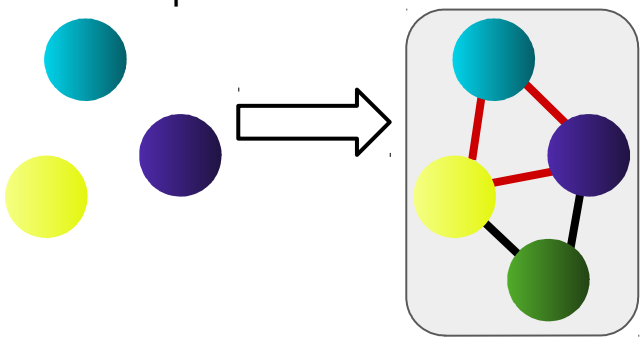
Table of Contents graphic (TOC)



Case 1

Input:
candidate partners

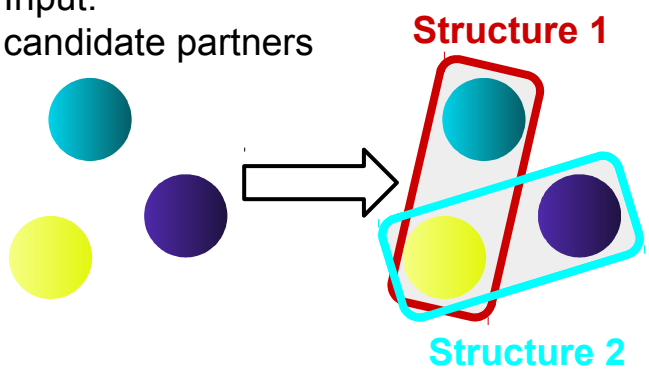
Experimental
3D structure



Output:
- **Interactions confirmed: true positives (TP)**
- **Undetected partners: false negatives (FN)**

Case 2

Input:
candidate partners



Output:
- **Mutually exclusive partners: false positives (FP)**

Figure 1

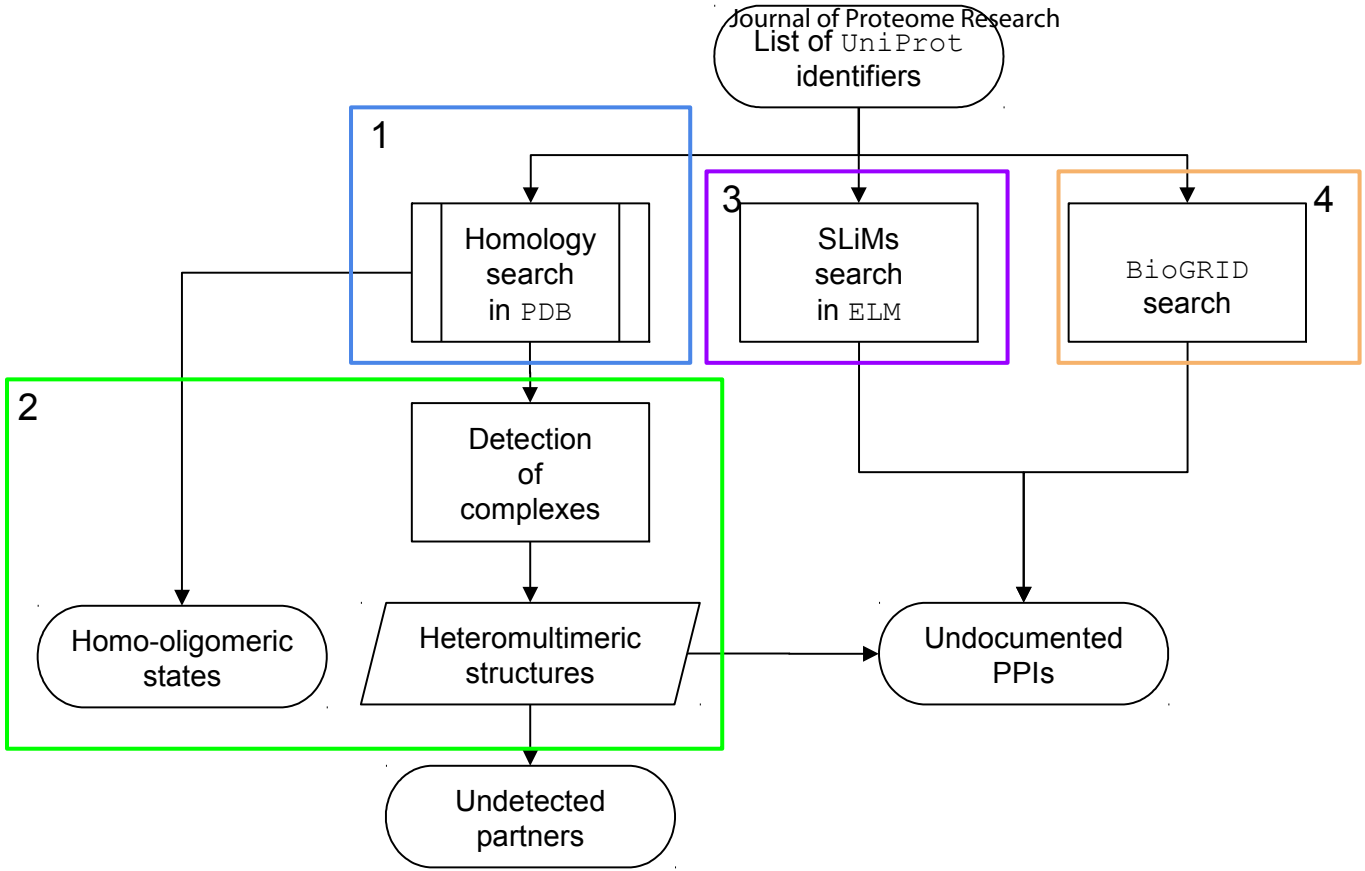
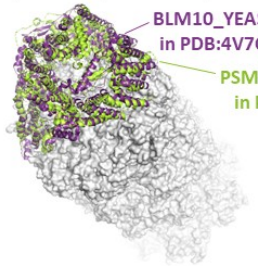

List of UniProt identifiers
1
Homology search in PDB
3
SLiMs search in ELM
4
BioGRID search
2
Detection of complexes
Homo-oligomeric states
Heteromultimeric structures
Undocumented PPIs
Undetected partners

Figure 2

A

Complex ID	Max avg id%	Inputs found	Structures found
c035	51.6	15	4V7O

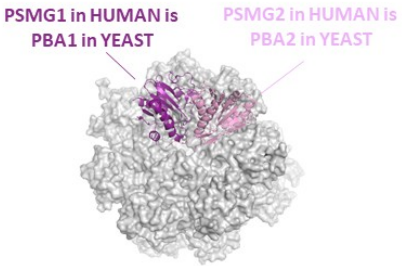


	Subunit	% Seq identity	Chain assigned
1	PSB5	65	A2
2	PSA5	61	AJ
3	PSA2	60	AG
4	PSB7	58	AM
5	PSA7	58	AI
6	PSA6	55	AA
7	PSB3	54	AN
8	PSB6	54	AB
9	PSA1	53	AK
10	PSA4	50	AH
11	PSA3	50	AL
12	PSB1	48	A3
13	PSB4	44	A4
14	PSB2	44	A1
15	PSME4	20	A6

Validated in PDB (6KWY & 6REY) →

B

Complex ID	Max avg id%	Inputs found	Structures found
c074	50	16	4G4S



	Subunit	% Seq identity	Chain assigned
1	PSB5	65	L
2	PSA2	62	B
3	PSA5	61	E
4	PSA7	60	D
5	PSB7	58	I
6	PSA6	55	A
7	PSB6	55	H
8	PSB3	54	J
9	PSA1	53	F
10	PSA4	50	C
11	PSA3	50	G
12	PSB1	48	M
13	PSB2	45	K
14	PSB4	43	N
15	PSMG2	21	P
16	PSMG1	18	O

C

Complex ID	Max avg id%	Inputs found	Structures found
c097	37.5	2	2Z5C

	Subunit	% Seq identity	Chain assigned
1	PSA5	56	C
2	PSMG3	19	B
x	x	x	x

Figure 3

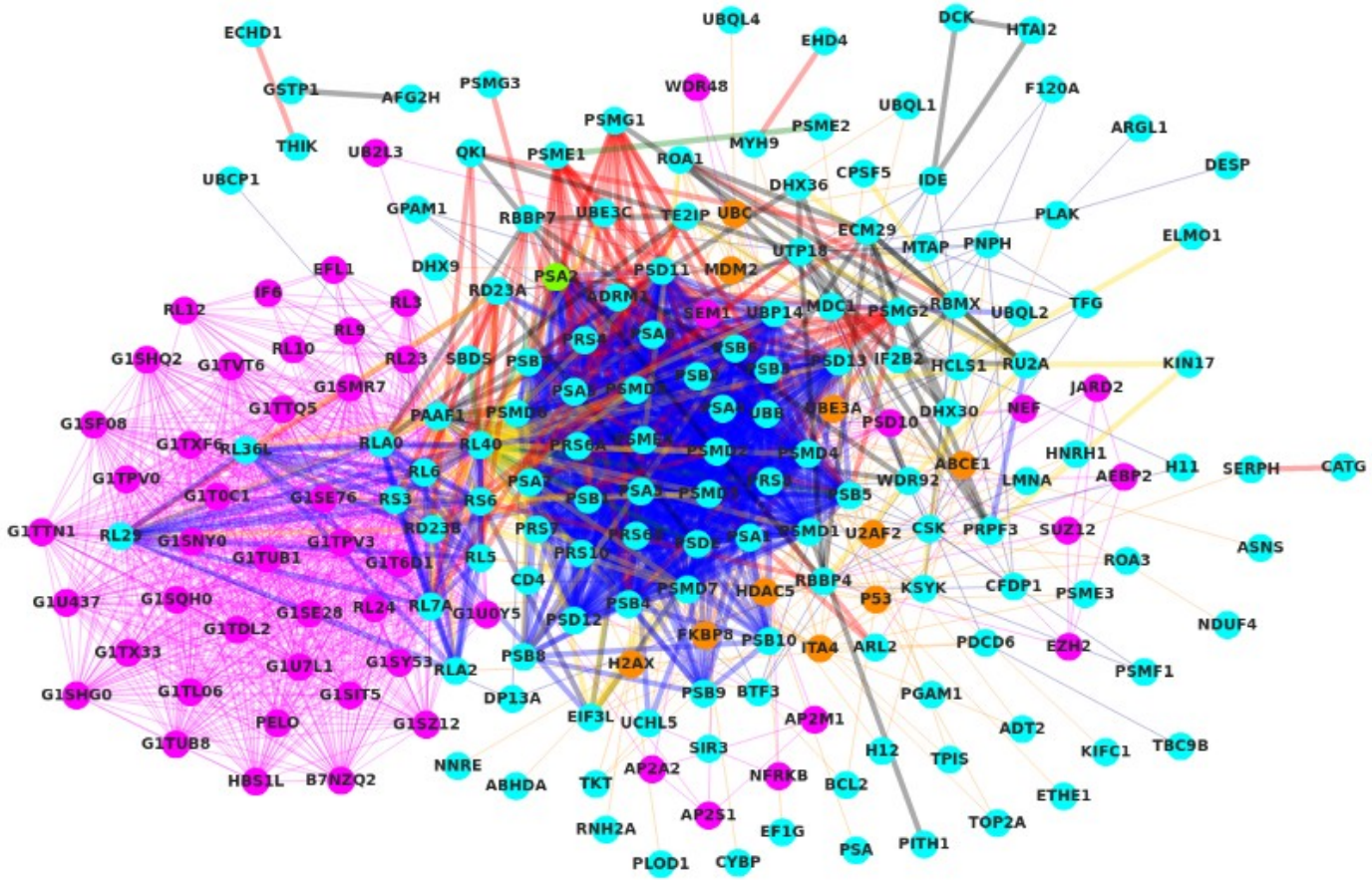


Figure 4

ACS Paragon Plus Environment

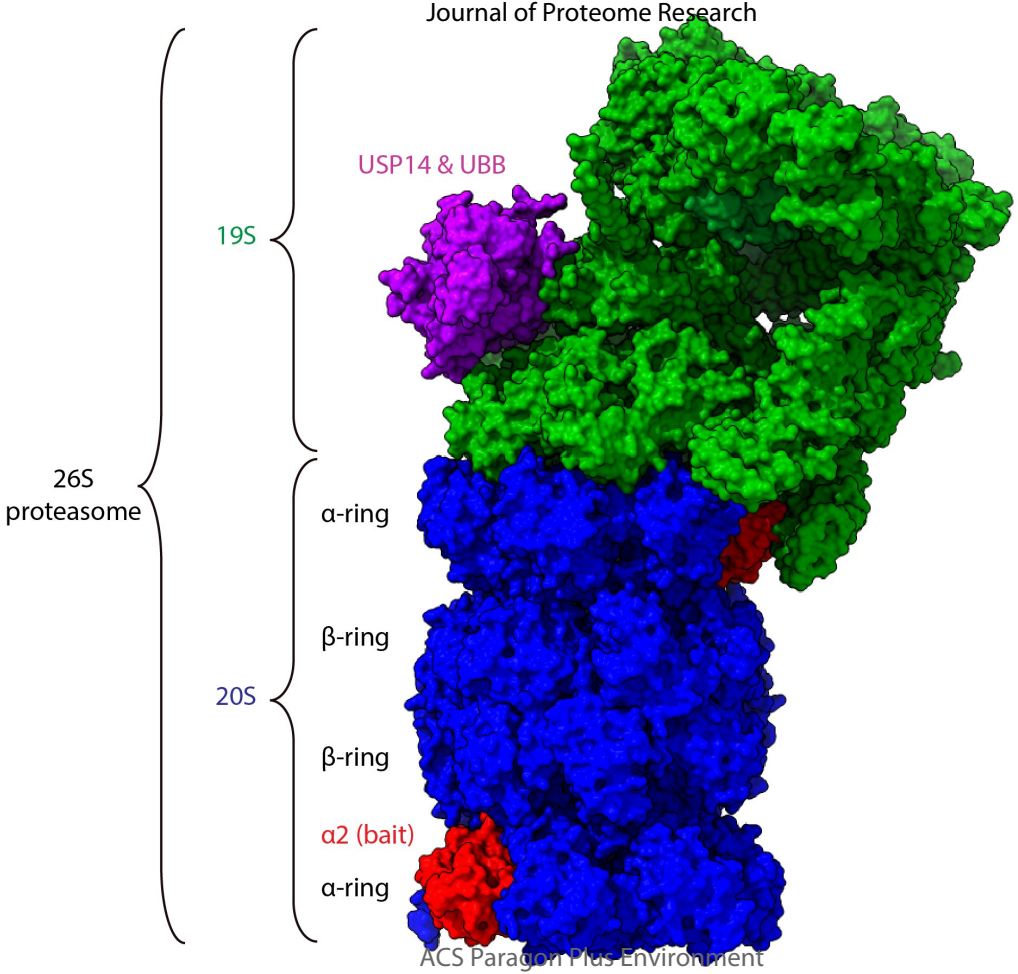
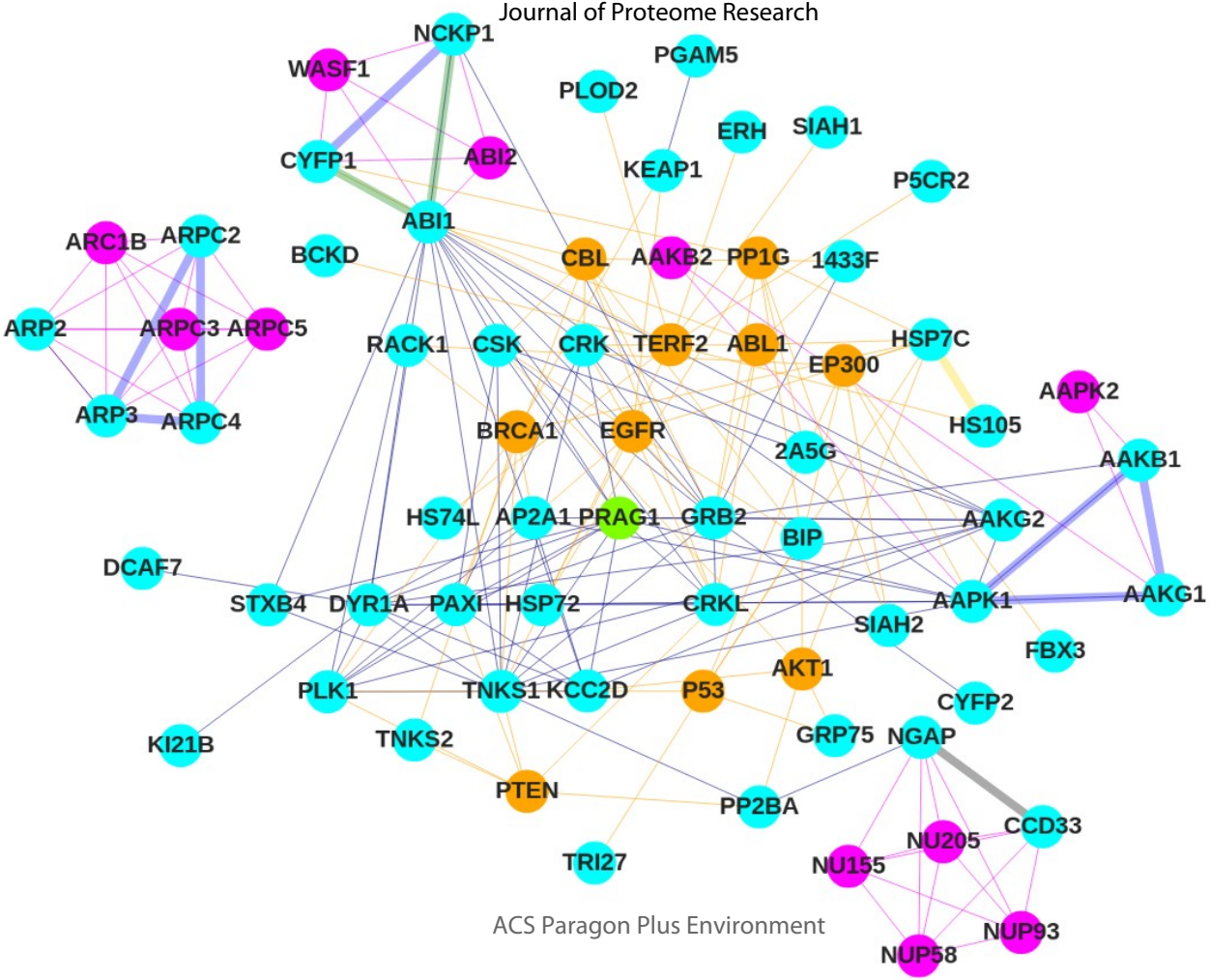


Figure 5



ACS Paragon Plus Environment

Figure 6

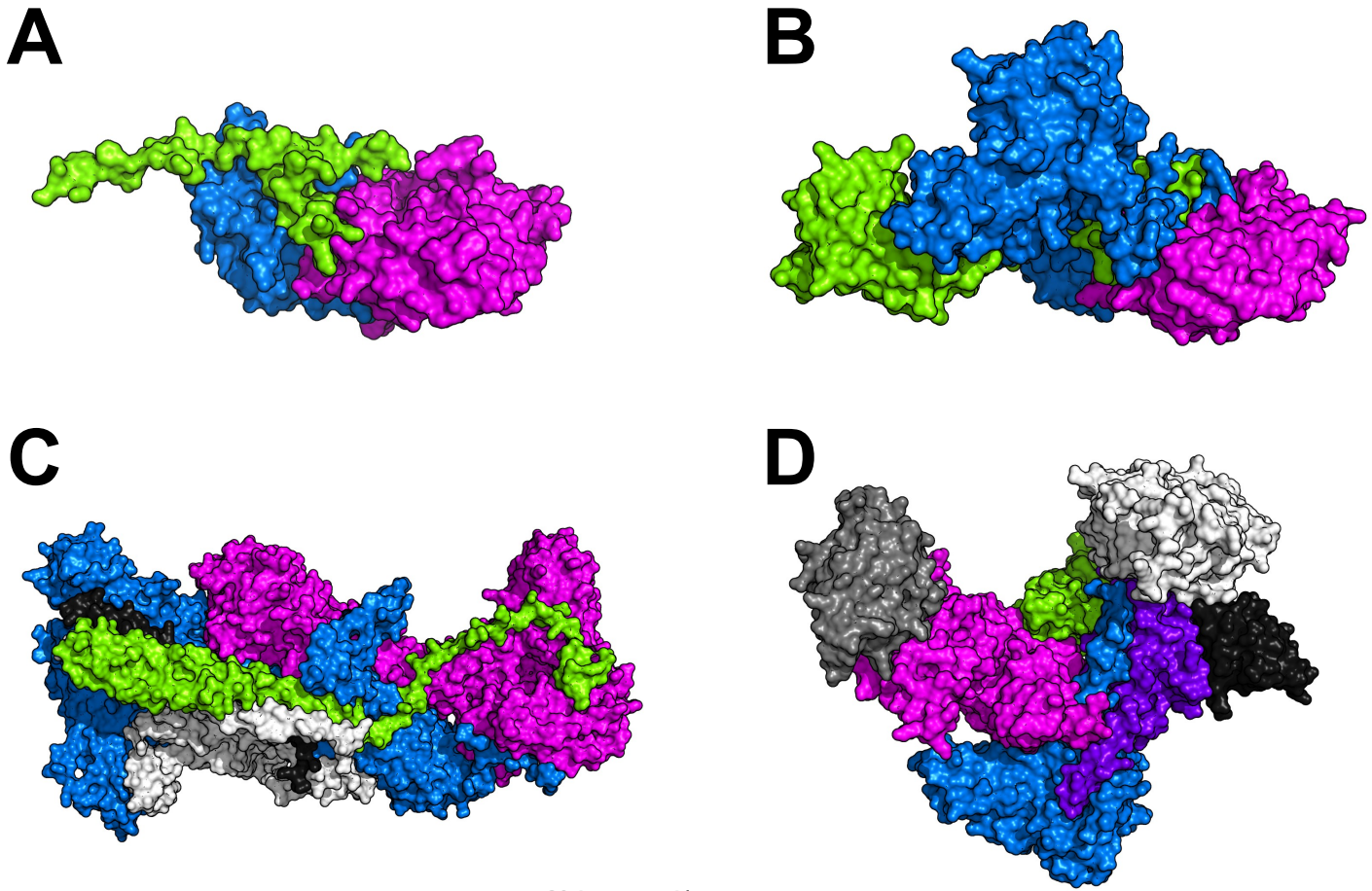


Figure 7