



**HAL**  
open science

## 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.

► **To cite this version:**

Guillaume Postic, Julien Marcoux, Victor Reys, Jessica Andreani, Yves Vandenbrouck, et al.. Probing Protein Interaction Networks by Combining MS-Based Proteomics and Structural Data Integration. *Journal of Proteome Research*, 2020, 19 (7), pp.2807-2820. 10.1021/acs.jproteome.0c00066 . hal-02571410

**HAL Id: hal-02571410**

**<https://hal.science/hal-02571410>**

Submitted on 9 Oct 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

This document is confidential and is proprietary to the American Chemical Society and its authors. Do not copy or disclose without written permission. If you have received this item in error, notify the sender and delete all copies.

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

Journal:	<i>Journal of Proteome Research</i>
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) Burllet-Schiltz, Odile; cnrs, ipbs Guerois, Raphael; Institute for Integrative Cell Biology Labesse, Gilles; CNRS, Centre de Biochimie Structurale Tuffery, Pierre; Université Paris Diderot, BFA, CNRS UMR8251, INSERM U1133

SCHOLARONE™  
Manuscripts

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

Guillaume Postic<sup>1,2</sup>, Julien Marcoux<sup>3</sup>, Victor Reys<sup>5</sup>, Jessica Andreani<sup>6</sup>, Yves Vandembrouck<sup>4</sup>, Marie-Pierre Bousquet<sup>3</sup>, Emmanuelle Mouton-Barbosa<sup>3</sup>, Sarah Cianféroni<sup>7</sup>, Odile Burette-Schiltz<sup>3</sup>, Raphael Guerois<sup>6</sup>, Gilles Labesse<sup>\*5</sup>, Pierre Tufféry<sup>\*1</sup>

<sup>1</sup>Université de Paris, BFA, UMR 8251, CNRS, ERL U1133, Inserm, RPBS, F-75013 Paris, France

<sup>2</sup>Institut Français de Bioinformatique (IFB), UMS 3601-CNRS, Université Paris-Saclay, Orsay, France

<sup>3</sup>Institut de Pharmacologie et de Biologie Structurale, IPBS, Université de Toulouse, CNRS, UPS, Toulouse, France

<sup>4</sup>Univ. Grenoble Alpes, INSERM, CEA, IRIG-BGE, U1038, 38000, Grenoble, France

<sup>5</sup>CBS, Univ. Montpellier, CNRS, INSERM, Montpellier, France

<sup>6</sup>Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France

<sup>7</sup>Laboratoire de Spectrométrie de Masse BioOrganique, Université de Strasbourg, CNRS, IPHC UMR 7178, 67000 Strasbourg, France

Corresponding authors: [pierre.tuffery@u-paris.fr](mailto:pierre.tuffery@u-paris.fr); [labesse@cbs.cnrs.fr](mailto: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

1  
2  
3 The vast majority of cellular functions are carried out by molecular machines made up of non-  
4 covalent protein–protein interactions (PPIs) and called protein complexes. However, PPI  
5 characterization is still far from being complete. Estimates are that between 120,000 and over  
6 one million such interactions occur in *Homo sapiens*<sup>1</sup>, but only on the order of 50,000 high  
7 quality PPIs are currently identified<sup>2</sup>. Turning to the structural level, estimates are that in the  
8 human proteome (restricted to proteins in the human interactome) two thirds of the proteins  
9 have either a structure experimentally resolved (~17,000 structures less than 95% identical in  
10 the PDB) or at least a part accessible as structural models. But again, moving to the human  
11 interactome, only 12% of interactions have a structure/model  
12 (<https://interactome3d.irbbarcelona.org/statistics.php>) and currently ~2,200 heteromeric  
13 complexes are structurally characterized. This has motivated a large effort of the community to  
14 identify ways to increase our knowledge on PPIs.  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32

33 Screening techniques to detect interacting partners fall into two categories: (i) genetic  
34 approaches, which measure mostly pairwise interactions and (ii) biochemical methods, which  
35 measure the co-complex interactions among multiple proteins. Both types of techniques have  
36 matured to the point where they now allow from focused to proteome-wide PPI network  
37 analyses. The genetic approaches consist of a test between two fusion proteins, whose  
38 interaction either activates the transcription of a reporter gene or restores the activity of a  
39 reporter enzyme. Examples of such binary techniques include the yeast-two hybrid system<sup>3</sup>,  
40 bacterial-two hybrid system<sup>4</sup>, MAMmalian Protein–Protein Interaction Trap<sup>5</sup> and protein  
41 complementation assays<sup>6</sup>. Large-scale biochemical methods mainly use enrichment strategies  
42 to capture a bait protein under conditions aiming at preserving native associations to its protein  
43 partners. Then, the bait is eluted with its potential interacting partners (preys), and subsequently  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 identified and quantified using liquid chromatography coupled to tandem mass spectrometry  
4 (LC-MS/MS). Typical co-complex techniques include the affinity purification-mass  
5 spectrometry (AP-MS) <sup>7,8</sup>, which involves tagging bait proteins with an epitope before  
6 purification of both bait and interacting partners using antibody affinity. An alternative method  
7 called proximity-dependant biotinylation consists in fusing an enzyme to the bait in order to  
8 modify its interacting partners <sup>9</sup>. Such techniques return a “flat” list of identified and quantified  
9 proteins that are all supposed to interact, directly or indirectly, with the bait protein. Although  
10 it can be associated with an observance number quantifying the number of occurrences of each  
11 peptide, at this stage, one issue is the absence of an organizational structure in the output list,  
12 to see which proteins do interact with which others. Proposing a hierarchical view of the  
13 interactions between the members of the flat list is extremely useful for biological interpretation,  
14 but tedious when done by hand. Moreover, one has also to consider that the list itself should be  
15 questioned: it may suffer from (i) the presence of identified PPIs that were wrongly assigned or  
16 do not occur physiologically and (ii) the non-detection of genuine associations—that we  
17 referred to hereafter as “false positives” and “false negatives”, respectively (see Figure 1).  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39

40 False positives in proteomics data may arise from various causes <sup>8</sup>. For example, when the bait  
41 protein is a member of different complexes in the cell, it will capture a set of prey proteins in  
42 which some never occur in the same complex (see Figure 1). Technically, for AP-MS  
43 experiments, non-specific binding of proteins to the antibody or the solid matrix may also  
44 contribute to the purification of proteins that do not specifically interact with the protein of  
45 interest. Experimental results are not either exempt from false negatives, as the integrity of the  
46 detected protein complexes may be compromised by different factors. For example, in  
47 biochemical co-complex methods, the purification step hinders transient and weak interactions  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 from being detected. Reversible <sup>10</sup> or irreversible <sup>11</sup> chemical cross-linking has been proposed  
4  
5 to overcome this problem, but at the price of introducing a bias in the pull-down data. Finally,  
6  
7 a low number of proteolytic peptides may lead to either false positives (identification of wrong  
8  
9 proteoforms, e.g. splice variants) or false negatives (signal below the detection threshold). With  
10  
11 regard to genetic binary approaches, transient complexes are considered, but all non-tagged  
12  
13 proteins are excluded from the analysis.  
14  
15  
16  
17  
18

19 One way to address these questions is to integrate the information of the multimeric protein 3D  
20  
21 structures available in the Protein Data Bank (PDB) (<http://www.rcsb.org/>) <sup>12</sup>. Indeed, the  
22  
23 experimental resolution of the structure of a complex (by methods such as X-ray, RMN, or  
24  
25 cryo-EM) supposes a high enough binding affinity between the partners and is, therefore, a  
26  
27 direct proof of the interaction. Moreover, the purification performed prior to the determination  
28  
29 of the protein structure avoids the presence of artifactual partners (*i.e.* false positives) in the  
30  
31 resulting 3D data. To date, this source of information encompasses the resolution of over 30,000  
32  
33 homomeric and 7,700 heteromeric non redundant complex structures. Furthermore, it is now  
34  
35 well established that interactions tend to be preserved throughout evolution and interface  
36  
37 binding modes are structurally conserved <sup>13–15</sup>, even for complexes involving partners with as  
38  
39 little as 20% sequence identity <sup>16</sup>. This provides a means for the large-scale identification of  
40  
41 complexes, even for species in which very few structures are available, through the search of  
42  
43 interologs (*i.e.* homologous complexes across species). Importantly, the oligomeric stability  
44  
45 required by protein structure determination techniques guarantees the integrity of the "obligate"  
46  
47 complex—which is the smallest set of subunits forming a stable complex, as observed among  
48  
49 various species—and, thus, the absence of false negatives. Therefore, this also provides a means  
50  
51 to control that all members of the obligate complex are observed in the interactome data, and  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 possibly, to identify false negatives. Finally, the knowledge of the structure directly provides  
4 information about the residues at the interface, providing the means to design further  
5 experiments to validate or modulate the interactions.  
6  
7  
8

9  
10 Turning to weaker or transient interactions, more and more evidence highlights the role of Short  
11 Linear Motifs (SLiMs), which correspond to short segments of the proteins mostly located  
12 within intrinsically disordered regions (IDRs). Many are documented in the PDB but are more  
13 difficult to detect by usual sequence comparison algorithms or fold recognition due to the small  
14 size of these SLiMs. In these cases, data accumulated in the Eukaryotic Linear Motif resource  
15 (ELM) (<http://elm.eu.org/>)<sup>17</sup> could prove helpful to decipher the potential interactions  
16 occurring between proteins of the “flat” list.  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27

28 Although the whole process of integrating structural and interactomics data can be done  
29 manually for small datasets, it requires advanced skills in searching structure information.  
30 Moreover, the task becomes highly tedious and error prone for larger or multiple datasets. To  
31 the best of our knowledge, the only available computational tool designed for that purpose is  
32 Interactome3D (<https://interactome3d.irbbarcelona.org/>)<sup>18</sup>. Its main goals are to propose 3D  
33 models of the protein complexes of a given species, and to provide insights about the protein  
34 residues involved in the interactions. It consists of a database setup to provide a compendium  
35 of the structural information on the PPIs available for a given organism. It integrates  
36 experimental structures or models of protein complexes for that organism and information  
37 extracted from external resources such as IntAct (<https://www.ebi.ac.uk/intact/>)<sup>19</sup>. Models are  
38 either extracted from ModBase (<https://modbase.compbio.ucsf.edu/>)<sup>20</sup> or built by comparative  
39 modeling - at more than 40% sequence identity, combining blast, 3did  
40 (<https://3did.irbbarcelona.org/>)<sup>21</sup> and MODELLER (<https://salilab.org/modeller/>)<sup>22</sup>. Currently  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 it covers 18 organisms. For a query species, the server takes as input a list of proteins or their  
4 binary interactions, to return structural annotations of the PPIs network of the query. Another  
5 popular resource to annotate PPIs is the STRING database (<https://string-db.org/>)<sup>23</sup>. Unlike  
6 Interactome3D, STRING considers protein structural data as “accessory information” and does  
7 not use it to build PPIs networks. Rather, it uses other sources of information, such as genetic  
8 interactions, text mining, or PPIs from other databases. Finally, our comparative analysis also  
9 includes GeneMANIA<sup>24</sup> (<http://genemania.org/>), as it can perform functional annotation and  
10 prediction from a flat list of protein genes, by integrating data from PPIs databases, such as  
11 BioGRID (<https://thebiogrid.org/>)<sup>25</sup> or PathwayCommons  
12 (<https://www.pathwaycommons.org/>)<sup>26</sup>. Interestingly, GeneMANIA is able to predict ortholog  
13 PPIs (interologs) through the integration of data from other species. However, it is currently  
14 limited to only nine organisms. The gap between the information content reachable by such  
15 resources and structural information remains however unclear.

16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35 Here, we present an integrative protocol aimed at validating and connecting candidate protein  
36 partners detected by interactomics experiments, starting from the “flat” list of the detected  
37 proteins, i.e. without *a priori* information on the actual interactions. The method is based on a  
38 large-scale search in the PDB for close and remote homologs of each member of a given list. It  
39 is followed by an analysis in terms of complexes to which these homologs belong to, in order  
40 to identify the complexes grouping the input proteins and provide a first level of data structuring.  
41 This protocol also extracts predicted weak and transient PPIs, through the search for Short  
42 Linear Motifs (SLiMs) located in intrinsically disordered regions (IDRs). Finally, the whole  
43 network analysis is further enriched by integrating experimental PPIs downloaded from the  
44 BioGRID database, which can be informative when no structures are available. Using two  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 datasets generated by MS-based proteomics experiments as case studies, we describe an  
4 original strategy for validating and expanding interaction data, and discuss its added-value and  
5  
6 limitations. Source code and instructions describing how to set up the full protocol are freely  
7  
8 available at <https://gitlab.rpbs.univ-paris-diderot.fr/src/proteo3Dnet>.  
9  
10  
11  
12

## 13 14 15 **2. MATERIALS AND METHODS**

### 16 17 18 19 *2.1 Enlarged homology search*

20  
21 Because only a minor portion (~1%) of protein sequences have a 3D structure available in the  
22  
23 PDB, we perform a search for known structures to be used as templates for comparative  
24  
25 modeling. The main component of our protocol consists in mining the full set of PDB structures  
26  
27 to find those templates, and possibly some structures that gathers several of the proteins of the  
28  
29 list established by proteomics experiments, i.e. PDB entries corresponding to the structure of  
30  
31 complexes involving several proteins of the list identified by the interactomics experiment.  
32  
33 Fold-recognition was first performed automatically using the @TOME-2 server  
34  
35 ([http://atome.cbs.cnrs.fr/ATOME\\_V3/index.html](http://atome.cbs.cnrs.fr/ATOME_V3/index.html))<sup>27</sup> to unravel the structural coverage for our  
36  
37 queries. However, this does not directly highlight the multimeric organization of the template  
38  
39 and hence the putative interactions among the query proteins. We resumed the search for  
40  
41 homologs using only HHsearch (<https://github.com/soedinglab/hh-suite> version 3.0.0)<sup>28</sup> and  
42  
43 considered further only templates with a probability >95.0%. Using such cut-off, the risk to  
44  
45 miss some remote templates exists - although in our experience it is very limited. However, we  
46  
47 prefer here to limit such risk rather than to possibly introduce false positive templates. Detecting  
48  
49 templates with a large confidence, the underlying hypothesis is that if several proteins form a  
50  
51 complex in a 3D structure, their homologs are likely to interact similarly, i.e. to correspond to  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 interologs. A specific difficulty comes from the fact that to perform efficiently, HHsearch  
4 searches for homologs in a subset of the PDB whose redundancy has been reduced to 70%  
5  
6 sequence identity, through a clustering procedure - all proteins of a cluster are represented by  
7  
8 only one cluster representative. Because any cluster representative identified by HHsearch as  
9  
10 the closest homolog is not necessarily the closest homolog that could be found in the entire  
11  
12 PDB, and because it could occur that non representatives of the clusters correspond to PDB  
13  
14 entries containing the structures of complexes of interest, we enlarge the HHsearch results with  
15  
16 all the members of a cluster. However, we still need to estimate the sequence identity of the  
17  
18 query to each member of the cluster to identify the cluster member best corresponding to the  
19  
20 query. Here, a new difficulty comes from the fact that HHsearch clusters are determined based  
21  
22 on the complete protein sequence, and that in many cases, the resolved structure corresponds to  
23  
24 only one part of it. In some cases, for two members of the same cluster, the regions for which  
25  
26 a structure could be determined do not overlap at all. This difficulty is overcome by using  
27  
28 MaxCluster (<http://www.sbg.bio.ic.ac.uk/maxcluster/>) (version 0.6.6), to subdivide, when  
29  
30 necessary, each cluster in sub-clusters. Then, for each sub-cluster, we build a multiple sequence  
31  
32 alignment (MSA) using the MUSTANG program (version 3.2.3)<sup>29</sup>, which uses both protein  
33  
34 sequence and structure information. Because it makes use of 3D information, this 3D based  
35  
36 MSA is expected to be of better quality than an alignment based on the sequence alone. Finally,  
37  
38 for any homolog identified by HHsearch, we use MAFFT (version 7.407)<sup>30</sup> to merge the  
39  
40 pairwise alignment produced by HHsearch with the MSA and we calculate the sequence  
41  
42 identity between the target and each sequence in the MSA. This way, each homolog initially  
43  
44 identified by HHsearch may be replaced by a cluster member of higher sequence similarity to  
45  
46 the query.  
47  
48  
49  
50  
51  
52  
53  
54

## 55 *2.2 Data integration*

56  
57  
58  
59  
60

1  
2  
3 After the search for 3D structures, each protein from the input list is associated with one, several,  
4 or no PDB entries. When several members of the input list are found in the same PDB entry  
5 (3D complex, biological unit), they are considered as interacting partners. Conversely, when a  
6 3D complex found contains a protein chain not present in the proteomics dataset, this chain  
7 corresponds to a potentially missing partner of the complex; the subunit is therefore labeled as  
8 undetected. It is not necessarily a false negative, as it may be part of a non-obligate complex.  
9  
10 An index of the number of times the chains are seen in the same complex is built to identify the  
11 proteins present in the complexes but not in the flat list, and to evaluate if they are part of the  
12 obligatory complex. Regarding the homo-oligomeric state, the number of copies of a protein  
13 chain within a complex is determined by extracting information about the stoichiometry in the  
14 PDB (author annotated biological unit). For example, a structure annotated by 'A3' ("Global  
15 Stoichiometry" field) is a homotrimer, whereas the 'A2B' annotation corresponds to a  
16 heterotrimer that contains two copies of the chain A and one of the chain B. For cases where  
17 there are several molecular assemblies assigned, the one that is labeled as "assigned by authors"  
18 (if any) is systematically preferred. Finally, it may happen that no 3D structure can be found  
19 for a candidate protein. In this case, the pipeline also integrates an additional source of  
20 interaction data: the Biological General Repository for Interaction Datasets (BioGRID)<sup>25</sup>, from  
21 which only physical (not genetic) interactions between proteins are retrieved, except those  
22 associated with the 'Far Western', 'Co-fractionation', 'Co-localization', 'Biochemical Activity',  
23 and 'High Throughput' experimental systems.

### 24 *2.3 Search for SLiMs*

25  
26 SLiMS are short motifs known to mediate PPIs and their identification can be useful to decipher  
27 weak interactions. They are located in disordered regions of proteins, and often in missing parts  
28 of resolved structures, We search into the protein sequences of the flat list, for occurrence of  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 such motifs—represented as regular expressions—as documented in the ELM database <sup>31</sup>. We  
4  
5 only consider motifs found within disordered binding regions, as predicted by the ANCHOR2  
6  
7 score <sup>32</sup>. The latter is calculated with the IUPred2A software <sup>32</sup> and only motifs with an average  
8  
9 ANCHOR2 score >0.95 are kept. ELM links motifs with interacting PFAM domains. Thus  
10  
11 occurrence of such PFAM domains—as annotated in Uniprot—are also searched within the  
12  
13 input proteins. Occurrences of pairs involving SLiMs-PFAM domains correspond to potential  
14  
15 PPIs. We emphasize that, although ELM contains information on fully validated such  
16  
17 interactions, the procedure relies for a large part on predictions and consequently, some  
18  
19 erroneous pairs could be suggested, and some others missed. In our experience however, the  
20  
21 added value of this detection largely counterbalances such risks.  
22  
23  
24  
25  
26  
27

#### 28 *2.4 Datasets*

29  
30 **Proteasome 20S.** The proteasome dataset was retrieved from <sup>33</sup>. Briefly, proteasome complexes  
31  
32 were immuno-purified from in vivo formaldehyde crosslinked human U937 cells using the  
33  
34 MCP21 antibody targeting the  $\alpha$ -2 subunit of the 20S core particle. After trypsin digestion and  
35  
36 extraction, peptides were identified by nano-liquid chromatography (LC) MS/MS. The database  
37  
38 search was performed using the Mascot Daemon software (version 2.3.2, Matrix Science).  
39  
40 Detailed information can be found in the original paper <sup>33</sup>.  
41  
42  
43  
44  
45

46  
47 **Pragmin.** The Pragmin dataset was retrieved from <sup>34</sup>. Note that the 61 proteins of this dataset  
48  
49 do not contain the Pragmin itself. Briefly, a SILAC-based quantitative proteomic analysis,  
50  
51 which allowed the identification in a semi-quantitative manner of Pragmin interactors in FLAG-  
52  
53 Pragmin transfected human HEK293T cells, was undertaken. It reproducibly yielded more than  
54  
55 52 specific interactors, with a mean ratio >4 and found two protein-kinases that were  
56  
57  
58  
59  
60

1  
2  
3 prominently associated with Pragmin (mean ratio >12), including the Ser/Thr kinases AMPK  
4 and the TK CSK. Previous Pragmin interactors such as Src<sup>35</sup> and Sgk269/PEAK1<sup>36</sup> were,  
5  
6 however, not recovered in our SILAC analysis. Detailed information can be found in the  
7  
8 original paper<sup>34</sup>.  
9  
10  
11  
12  
13

### 14 **3. RESULTS AND DISCUSSION**

15  
16 Figure 2 presents an overview of the full protocol that we introduce here. We discuss shortly  
17 its main lines, details of each component being presented in the Materials and Methods section.  
18  
19 Starting from the list of proteins established by proteomics experiments—that could be, for  
20  
21 instance, specified as a series of UniProt identifiers—it consists of 4 steps:  
22  
23  
24

- 25  
26 1) The search for homologs of each of the input proteins. The main task of our protocol  
27 consists in finding at least one PDB structure that gathers several of the UniProt  
28 identifiers from the list. However, only a minor portion (~1%) of UniProt entries have  
29 a 3D structure available in the PDB—due to the difficulty of determining protein  
30 structures experimentally. To overcome this difficulty, the protocol performs instead a  
31 search for homologs using HHsearch<sup>28</sup>, which is presently one of the best-performing  
32 approaches to detect remote homologs. Being oriented towards the identification of  
33 templates for homology modeling, it returns only one representative per cluster of  
34 similar structures of the PDB, which is undesired in our case as it might miss  
35 associations between input proteins. For this reason, the standard HHsearch procedure  
36 is followed by a specific search in the clusters to identify the cluster member best  
37 matching a query. This step is, by far, the most time consuming part of the protocol.  
38  
39 Although the number of input sequences is in theory not limited, it can require several  
40  
41 hours for an input of over 700 sequences, depending on the running machine.  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 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 homology 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,

1  
2  
3 as such examples from the proteasome interactome show the importance of detecting remotely  
4 homologous complexes, we have included them in Figure 3 to illustrate the concept of  
5  
6  
7  
8 interologs.  
9

### 10 11 12 *3.1 Proteasome interactome*

13  
14 The 20S proteasome is a huge proteolytic complex involved in the regulation of intracellular  
15 protein homeostasis. Its 20S catalytic core is composed of 14 different subunits that assemble  
16 as a 716 kDa barrel made of 4 heptameric stacked rings<sup>37</sup>. The 20S proteasome can then interact  
17 with different regulators, the most abundant of which is the 19S regulatory particle, that is made  
18 of ATPase (RPT) and non-ATPase (RPN) subunits and forms the ATP-dependant 26S ubiquitin  
19 proteasome system. On top of its many substrates, which only transiently contact the catalytic  
20 core, more than 300 Proteasome Interacting Proteins (PIPs) are known to interact with the 20S  
21 proteasome. Due to its central role in homeostasis, the proteasome has been thoroughly studied  
22 from a structural point of view by cryo-Electron Microscopy<sup>38-40</sup>, X-ray crystallography<sup>41</sup> and  
23 Nuclear Magnetic Resonance<sup>42</sup>. MS-based techniques have also brought important insights on  
24 the primary to quaternary structures of the many different proteasome complexes. While top-  
25 down proteomics informs on the proteoforms present within the 20S core particle<sup>43,44</sup>, bottom-  
26 up proteomics is commonly used to identify novel proteasome partners in complex mixtures<sup>45</sup>,  
27 and physical interactions with putative binders of the proteasome have been confirmed by  
28 native MS<sup>46,47</sup>. Finally cross-linking MS and bioinformatics modeling have been extensively  
29 used to better characterize highly enriched proteasome complexes<sup>48,49</sup> but also, more recently,  
30 to identify and model the interactions with new partners found in complex mixtures or even *in*  
31 *vivo*<sup>50,51</sup>. While distance restraints are not compatible with our current pipeline, lists of  
32 interactors identified by cross-linking MS experiments can readily be used as an input.  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 Figure 4 depicts a graph summarizing the results of the protocol. Of the 192 partners identified  
4 by immuno-purification (IP)<sup>25</sup> of the 20S proteasome (via its constitutive  $\alpha 2$  subunit, in red in  
5 Figure 5), only 9 (CFDP1, CG050, GPAM1, NDUF4, PIP30, PMGT2, POMP and RL29  
6 STK19) were not assigned reliably to any structural template. For a small number (17) only  
7 very distant similarity but consistent (e.g.: F120A) could be detected with sequence identity  
8 ranging from 12 to 22%. Among the latter, three proteins (PSMG1, PSMG2 and PSME4) are  
9 involved in proteasome maturation. On the contrary, for 109 proteins identical structures were  
10 matched. For 145 and 164 proteins, sequence identity shared with the detected templates for  
11 structure modeling was above 65% and above 35% respectively. Hence, an almost complete  
12 structural coverage (95%) was reached ([http://atome4.cbs.cnrs.fr/ms2model\\_proteasome](http://atome4.cbs.cnrs.fr/ms2model_proteasome)).  
13  
14

15  
16  
17 Up to 70% of the detected partners were assigned to proteins found in homo-multimeric  
18 complexes (Table S1). We found for example the expected homo-heptameric activators PA28 $\alpha$   
19 (PDB 5MSJ<sup>52</sup>) and PA28 $\beta$  (PDB 5MSK<sup>52</sup>). Interestingly, these homo-heptamers are not  
20 physiological, as they naturally form within the cell an  $\alpha 4\beta 3$  hetero-heptamer, that is better at  
21 activating the 20S proteasome<sup>52</sup>. There is no structure available yet for the homo-heptameric  
22 PA28 $\gamma$ , and the best homology model proposed was the one of PA28 $\alpha$  (PDB 5MSJ). As this  
23 part of the pipeline is strictly looking for homomers, the  $\alpha$ -subunits  $\alpha 1$  to  $\alpha 7$  that form the  
24 hetero-heptameric  $\alpha$ -ring of the human 20S proteasome, were assigned to the distantly related  
25 homo-heptameric  $\alpha$ -ring from *Archaeoglobus fulgidus* (PDB: 1J2P<sup>53</sup>).  
26  
27

28  
29 In the next step, 128 complexes were identified (Table S2), containing at least 2 identical and/or  
30 homologous proteins from our IP and 20 of these were 100% identical, including 8 dimers, 2  
31 trimers, a 10 hetero-oligomers ranging from 7-mer to 33-mer. However 96 out of 192 proteins  
32 could not be assigned to any of these 128 complexes (50% of isolated proteins). The ELM  
33 search rescues 17 of these proteins and includes them in the interaction network (41% of  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 isolated proteins). The Biogrid search further added 25 more proteins to the interactome, which  
4 led to a total of 54 (28%) unexplained preys, that can either be false negatives, new interactors  
5 or transient substrates of the proteasome.  
6  
7  
8

9  
10 All the 33 subunits of the 26S proteasome (Figure 5, red, blue and green) bound to USP14-  
11 UbAI (Figure 5, purple, PDB: 5GJQ<sup>54</sup>) were found in the IP, meaning that this complex is most  
12 likely present. 21 Smaller sub-complexes of the 26S ranging from 20 (PDB: 5L4G) to 31 (PDB:  
13 5GJR<sup>54</sup>) subunits, were also found. Structures of even smaller complexes or subcomplexes (7  
14 to 17 subunits) of the 19S human proteasome regulatory particle itself (Figure 5, green) were  
15 also listed. As expected, the human standard 20S proteasome (PDB: 4R3O<sup>55</sup>) was identified by  
16 the pipeline (Figure 5, red and blue), together with its 228 homologous structures from mouse,  
17 bovin and yeast. A challenge for the yeast proteins was to automatically discriminate the correct  
18 pairs of orthologs among the large set of paralogous subunits in human proteasome (7  $\alpha$ - and 7  
19  $\beta$ -subunits). Interestingly, we could check that all 14 pairs of orthologs between yeast and  
20 human were correctly assigned despite important phylogenetic divergence. Furthermore, an  
21 alternative to this standard 20S, is called the immunoproteasome, and was also proposed (PDB:  
22 6AVO<sup>56</sup>), as the three immuno-catalytic subunits (namely PSB9, PSB10, and PSB8) were also  
23 on the list of  $\alpha 2$ 's partners.  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41

42 Other complexes involved ubiquitin (UBB), known to target many substrates and enzymatic  
43 regulators to the 19S regulator when the latter is bound to the 20S catalytic core. More precisely,  
44 it was found in complexes containing ubiquitin binding motifs or 26S-bound deubiquitinating  
45 enzymes such as the ternary complexes UBB-RPN13-RPN2 and UBB-RPN13-UCHL5 or  
46 heterodimers such as UBB-RAD23A and UBB-RL40. This last dimer is particular because  
47 RL40 is the precursor of Ub made of one chain of Ub fused to the 60S ribosomal protein L40.  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 L40 were identified and that RL40 could well be a false positive. The heterodimers UBB-  
4 UCHL5 (PDB: 4UF6<sup>57</sup>) and RPN13-UCHL5 (PDB: 4UEM<sup>57</sup>) were also identified, as they are  
5  
6 subcomplexes of the UBB-RPN13-UCHL5 trimer (PDB: 4UEL<sup>57</sup>). Other complexes containing  
7  
8 RPN13 include RPN13-RPN2, RPN13-UBQLN2. We then found PDBs corresponding to the  
9  
10 heterodimers of RPN10, another subunit of the 19S proteasome, and RAD23A & RAD23B  
11  
12 which are multi-Ub chain receptors involved in the shuttling of polyubiquitin substrates to 26S  
13  
14 proteasomal degradation.  
15  
16

17  
18  
19 The 108 remaining complexes were found with an identity that is <100% and correspond for  
20  
21 example to close homologues, as is the case for the previously mentioned PA28 $\alpha$ 4 $\beta$ 3 hetero-  
22  
23 heptamer from mice (PDB: 5MX5<sup>52</sup>), the UBB-RPN10 heterodimer from yeast (PDB: 5LN1<sup>58</sup>)  
24  
25 or the RPN8-RPN11 complex from yeast (PDB: 4OCN<sup>59</sup>). Other complexes were largely  
26  
27 incomplete, such as the human 80S ribosome (10 subunits out of 82), the human respiratory  
28  
29 complex (3 out of 68), the 48S preinitiation complex (8 out of 36), the human spliceosome (3  
30  
31 out of 51) and the human anaphase promoting complex (2 out of 19), to mention a few.  
32  
33

34  
35 Finally, a number of putative homologues have caught our attention: it is the case of the E3 Ub-  
36  
37 protein ligase that would form a heterodimer with the previously mentioned RL40, based on  
38  
39 crystal structures of homologous E3 ligases bound to selective Ub variant probes <sup>60</sup> (PDBs:  
40  
41 5HPK, 5HPL, 5HPS, 5HPT<sup>60</sup>). Another complex between the RL40 precursor and UBP14, a  
42  
43 proteasome-associated deubiquitinase is proposed, based on the crystal structure of a complex  
44  
45 with the homologous deubiquitinase UBP21 (PDB: 3I3T<sup>61</sup>). Because RL40 was identified via  
46  
47 its ubiquitin motif only, these two complexes have to be taken with great caution as they could  
48  
49 be false-positives. Of high interest is the predicted interaction between the 20S subunit  $\alpha$ 5  
50  
51 (PSMA5) and the proteasome assembly chaperone 3 (PAC3, PSMG3\_HUMAN) based on the  
52  
53 structure of their yeast homologues (2Z5C<sup>16</sup>). A related complex involves the two other main  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 proteasome assembly chaperones 1 and 2 that are known to form a heterodimeric complex and  
4  
5 is predicted to interact with the 20S subunit  $\alpha 6$  (PSMA1) based on the yeast homologous  
6  
7 proteasome crystallized in complex with Pba1-Pba2 (PDB 4G4S<sup>62</sup>). Remarkably, proteasome  
8  
9 chaperones in human and yeast diverged drastically in evolution and share very little sequence  
10  
11 identity (around 20%). This case underlines the interest of using HHsearch to reveal remote  
12  
13 homology relationships between orthologous proteins. Other expected homology models  
14  
15 include the interactions between RPN11 and RAD23A (Ub receptor) or between RPN11 and  
16  
17 RPN8.  
18  
19  
20  
21  
22

23  
24 Other proposed complexes were unexpected and caught our attention, such as the proposed  
25  
26 heterodimer between MDC1 and TE2IP that was based on the crystal structure of MDC1 bound  
27  
28 to a homologue of TE2IP (TOPBP1). We did not expect this interaction between these two  
29  
30 partners within our immuno-purification, because none of these proteins is apparently related  
31  
32 to the 20S and its common proteasome interactors, except for the ELM interaction proposed  
33  
34 between MDC1 and RPN10 (PSMD4) (Table S3). However, both proteins are involved in the  
35  
36 DNA damage response and might either bind to the 20S proteasome as a dimeric subcomplex  
37  
38 or constitute proteasomal substrates co-immuno-purified with the protease <sup>63</sup>.  
39  
40  
41  
42  
43

### 44 *3.2 Pragmin interactome*

45  
46 Pragmin is a recently characterized pseudo-kinase from rat and homologous to the human  
47  
48 SGK223 <sup>64</sup>. Pragmin was shown to strongly activate a tyrosine protein-kinase Csk and to  
49  
50 regulate cell migration <sup>34,65</sup> by a mechanism yet to be fully unraveled.  
51  
52

53  
54 Among the 61 candidate partners of the Pragmin interactome <sup>34</sup>, it was possible to identify at  
55  
56 least one solved structure for 60 cases (Table S4) and  
57  
58  
59  
60

1  
2  
3 [http://atome4.cbs.cnrs.fr/ms2model\\_sgk223](http://atome4.cbs.cnrs.fr/ms2model_sgk223). For 48 cases, sequence identity shared with the  
4 detected template(s) for structure modeling was above 65% and above 35% for 7 more proteins.  
5  
6  
7  
8 25 structures were identified with a sequence identity of 100%. Remarkably, a near complete  
9 structural coverage was reached, thanks to remote homology detection.

10  
11  
12 Further analyses revealed that 50 proteins were predicted to form homomultimers (Table S4).  
13  
14 Figure 6 depicts a graph summarizing the results of the protocol. Most of these (45 candidates)  
15 correspond to homodimers, though tri-, tetra-, hexa- and decamers were also found. 15 proteins  
16 (14 dimers and one decamer) matched the sequence of the template structure with at least 99%  
17 sequence identity which cast little doubt on the actual oligomeric organization for those proteins.  
18  
19 For 11 other sequences, one observed a sequence identity of more than 65% with multimeric  
20 structures, again suggesting one could infer conservation of the same oligomeric state by  
21 homology. For another 12 proteins, sequence identity with multimeric templates was between  
22 62% and 39% and the conservation of the same oligomeric states is more questionable and  
23 should be investigated for other evidence. Lower similarities were detected for 14 other  
24 sequences but further analyses are necessary to precise the potential assembly of those proteins.  
25  
26 The remaining 9 sequences showed no obvious similarity with any multimeric assemblages.  
27  
28 Obviously, this already provided quite significant coverage and valuable information compared  
29 to a “flat” list of interactants. However, this does not provide direct clues to validate any of the  
30 PPIs detected by AP-MS.  
31  
32

33  
34  
35 Thirty proteins could not be associated with any heteromultimeric complex. Most are expected  
36 to be isolated polypeptide chains (or pure homomultimers). 3D information allowed  
37 nevertheless the prediction of several complexes (Table S5), and in some cases, structure  
38 analysis brings complementary information to the raw interactomic knowledge.

39  
40  
41 The simultaneous detection of AAPK1, AAKB1, AAKG1- and AAKG2 suggested that the  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 AMP-activated protein kinase (AMPK) complex corresponds to a heterotetramer. However,  
4 this complex is heterotrimeric and made of only three subunits  $\alpha 1$  (or  $\alpha 2$ ),  $\beta 1$  (or  $\beta 2$ ), and  $\gamma 1$   
5 (or  $\gamma 2$ ) (Figures 7A and 7B). Therefore, the closely related paralogues AAKG1 ( $\gamma 1$ ) and  
6 AAKG2 ( $\gamma 2$ ) are not expected to participate in the same macromolecular assemblies. Instead,  
7 this suggests that these two heterotrimers could co-exist, namely  $\alpha 1\beta 1\gamma 1$  and  $\alpha 1\beta 1\gamma 2$ , which  
8 corresponds to an interesting case of false positive identification involving mutually exclusive  
9 interactions.

10  
11  
12 Another complex that was partially identified would be composed of CYFP1, NCKP1, and  
13 ABI1 proteins, which are found together within the native structure of the human WAVE  
14 regulatory complex (PDB entries 3P8C<sup>66</sup> and 4N78<sup>67</sup>). Again, the observed arrangement in the  
15 crystal structure suggested that the paralogous CYFP2 belongs to an alternative complex with  
16 NCKP1 and ABI1. The crystal structure also highlights that these proteins form extensive and  
17 intricate interactions with two undetected sequences (WASF1) and (BRK1) (Figure 7C) that  
18 could correspond to false negatives. Indeed, WASF1 was actually detected in the AP-MS results,  
19 but did not pass the standard statistical threshold and was, at first, discarded from the final list  
20 of candidates (Serge Urbach, personal communication). The rather short BRK1 protein (75  
21 residues) was not detected at all, although it forms a 55-residue long helix buried in the complex.  
22 We wonder whether the small size of BRK1 prevented its straightforward identification by MS  
23 (despite the presence of numerous basic residues all along its sequence) or that it was effectively  
24 absent from the sample. Of note, BRK1 is detected in interaction with some WAVE components  
25 in distinct MS studies according to BioGRID but never systematically associated with all of  
26 them.

27  
28  
29 Similarly, the interactions between the four candidate proteins ARPC2, ARPC4, ARP2, and  
30

1  
2  
3 ARP3 were confirmed by the identification of the structure of the bovin Arp2/3 complex  
4  
5 (entries 1K8K<sup>68</sup> and 3DXK<sup>69</sup>). However, the crystal structure comes as an hetero-heptamer,  
6  
7 while three polypeptide chains of this assembly (namely ARC1B, ARPC3, and ARPC5) are  
8  
9 missing in this AP-MS flat list. Interestingly, these missing partners appear at the periphery of  
10  
11 the complex (Figure 7D) and might not be obligatory or, the interactions with these proteins  
12  
13 could be more labile than those between the ARP2-ARP3-ARPC2-ARPC4 core. Alternatively,  
14  
15 one of the missing partners, ARPC1, binds to the ARP2/3 complex through its WD40 domain.  
16  
17 It is interesting to note that several proteins (KEAP1, DcAF7, GBLP, KIF21B and potentially  
18  
19 CB044) detected in the Pragmin interactome contain a WD-40 domain and at least one may  
20  
21 replace ARPC1 here. Finally, it is important to note that the detection of this ARP2/3 complex  
22  
23 together with that of the WAVE regulatory complex makes sense as they are known to be  
24  
25 involved in regulating the actin nucleating activity, in agreement with the role of Pragmin in  
26  
27 promoting cell migration <sup>70</sup>.  
28  
29  
30  
31  
32  
33

34  
35 Smaller heterocomplexes were also predicted based on structural conservation such as the two  
36  
37 heterodimeric chaperones HS71A-HS71B and HSP7C-HS105 which correspond to the  
38  
39 experimental complexes PDB 1XQS<sup>71</sup> and PDB 3C7N<sup>72</sup>. So far, we cannot extend this  
40  
41 heterodimerization to the other HSP70-like chaperones found in this interactomic study  
42  
43 (namely HSP72, HSP76, HS74L, GRP75 and GRP78).  
44  
45  
46  
47  
48

49 Of note, the above structural analysis provided no clues to identify which preys actually bind  
50  
51 to the Pragmin bait as no direct interaction with Pragmin could be detected, so far. In fact,  
52  
53 Pragmin contains a pseudo-protein-kinase domain, a dimerization domain <sup>64</sup> as well as a very  
54  
55 long (~950 aa) N-terminal tail. The latter is predicted to be natively disordered and to contain  
56  
57  
58  
59  
60

1  
2  
3 short linear motifs targeting domains such as SH2, SH3, 14-3-3, PWWP or WD40. The first  
4 such motif located within the N-terminus of Pragmin and identified by the ELM analysis (Table  
5 S6) was the EPIYA sequence, a motif recognized and actively phosphorylated by CSK through  
6 its SH2 and protein-kinase domains, respectively. We also detected a proline-rich motif  
7 potentially recognized by either PWWP or SH3 domains, and this prediction matches the  
8 recently detected binding of CrkII to the equivalent motif present in PEAK3, a recently  
9 discovered paralogue of Pragmin <sup>73</sup>. These motifs can explain the detection of Csk and Crk (or  
10 its closed homologue CrkL) in the interactome of Pragmin.  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22

23  
24 In addition, ELM analysis suggested a potential connection of STBX4 with Pragmin through  
25 its WW domain while the C-terminus of Csk could be recognized by the PDZ domain of STBX4.  
26 This would tightly connect these three proteins in an interleaved manner. Other proline-rich  
27 segments exist in Pragmin N-terminus and in its preys while several of the latter (Csk, Crk,  
28 CrkL, GRB2 and ABI1) contain SH3 domain, an intricate network may be at play. As an  
29 example, ABI1, a subunit of the WAVE regulatory complex possesses both a SH3 domain and  
30 a PxxDY motif potentially recognized by a SH3 domain. This may connect the WAVE  
31 regulatory complex with the complex formed by Pragmin, CSK and CRK (see Fig S1).  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41

42 Obviously, numerous experiments will be necessary to confirm (or not) these predicted  
43 interactions but we illustrate here how our protocol can provide clues to guide such  
44 experimental validations. As a whole, the data returned by our protocol helped to reveal an  
45 integrated network converging toward cell migration on one side and oxidative stress response  
46 on the other side.  
47  
48  
49  
50  
51  
52  
53  
54  
55

#### 56 **4. DISCUSSION AND PERSPECTIVES**

57  
58  
59  
60

1  
2  
3 In the present study, we have introduced a full protocol that starts with a list of proteins  
4 detected by interatomics experiments and integrates knowledge from the available structures of  
5 complexes and predictions of weak interactions to propose a structured overview of the  
6 collection of input proteins in terms of their interactions. To overcome the limitation of the  
7 rather low number of homomeric and heteromeric non-redundant complex structures resolved  
8 to date (~30,000 and 7,700, respectively), our protocol elaborates over a remote homology  
9 detection approach.  
10  
11  
12  
13  
14  
15  
16  
17  
18

19 From the two case studies presented here, the added value of using such remote  
20 homology search seems obvious. First, it is striking that we could detect either the proteins or  
21 homologs of them for all of the 61 proteins of the Pragmin dataset but one, and for 182 of the  
22 192 proteins of the proteasome dataset, i.e. for close to 96% of the cases. Following, the results  
23 obtained from the challenging case of the proteasome test case illustrate well that (i) the quality  
24 of the scores of the remote detection distinguish well between orthologs and paralogs, and (ii)  
25 prove the ability of our procedure to identify core obligate complexes. Moreover it allows to  
26 “hierarchize” the flat list of proteins in terms of interactions, even for sequence identities as low  
27 as 20%. Regarding the proteasome dataset, we were indeed able to link together 138 of the 192  
28 proteins, as deduced from 128 identified complex structures that contain at least 2 identical  
29 and/or homologous proteins from the input list. For the Pragmin dataset, our results well  
30 illustrate the ability of the approach to decipher more subtle cases, such as for instance the  
31 existence of mutually exclusive complexes involving different but overlapping subset of  
32 proteins. They also illustrate the potential added-value of predicting lower affinity interactions  
33 implying SLiMS to further decipher the interactions of the input. Finally, it is also interesting  
34 that the protocol was able to return information about homo-oligomers for both test sets.  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 Compared to other approaches such as Interactome3D, one advantage of our protocol is that it  
4 does not depend on any precalculation, and can infer information from different species. One  
5 can exemplify this for our two use-cases. For sake of clarity, we focus on the Pragmin case. For  
6 this use-case, our method connects 48 input proteins, into a single cluster, through 172  
7 interactions: 21 structure-based, 60 SLiMs-based, and 71 from BioGRID. Interactome3D  
8 returned, starting from the 61 proteins, heterocomplex interactions for 27 proteins, through 32  
9 connections: 18 structure-based and 14 others. These interactions define three clusters, which  
10 contain 2, 4, and 21 proteins. More in detail, our protocol identified as true positives the  
11 connections between all the partners of the ARPC2/3 complex, when Interactome3D, only did  
12 so for those between ARP2 and ARPC4, all other connections coming from theoretical models.  
13 This difference lies in the fact that the Interactome3D analysis was performed for *H. sapiens*,  
14 which is the organism from which the input dataset has been produced. Consequently,  
15 Interactome3D did not take into account the 19 structures obtained from *Bos taurus* proteins  
16 (which share 100% sequence identity with their human counterparts), which highlights the  
17 interest of considering homology detection across species. Interactome3D also identified a  
18 connection between AAKG1 and AAKG2, while these proteins are actually two alternative  
19 partners of the AMPK complex, as discussed above. With regard to the WAVE regulatory  
20 complex, Interactome3D failed to validate the true positive interaction between ABI1 and  
21 CYFP1, whereas our protocol identified a link between these two proteins, based on two PDB  
22 structures, 4N78 and 3P8C, found with 95.0% and 94.5% sequence identity, respectively.  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50

51 Considering the STRING database, it returned for Pragmin 142 interactions, connecting 50 of  
52 the 61 entries, separated into four clusters of 2, 2, 3, and 43 proteins. STRING identifies the  
53 second largest number of connections—2 more than our protocol and 23 more than  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 Interactome3D—but we recall STRING integrates information sources such as genetic  
4 associations and text mining, which do not necessarily correspond to physical interactions. Also,  
5  
6 the "protein homology" links established by STRING are of completely different nature than  
7  
8 those produced by Interactome3D and our method. Such links actually represent an  
9  
10 evolutionary relationship between the two proteins. For example, the PLOD2 and PLOD3  
11  
12 proteins do not physically interact and, yet, STRING connects them, based on their sequence  
13  
14 similarity. Results returned by STRING face the same issue as Interactome3D for the AMPK  
15  
16 complex, highlighting again the added value of considering structural information. For the  
17  
18 WAVE complex, STRING identified the true positive interaction between ABI1 and CYFP1.  
19  
20 Interestingly, both STRING and our method detected the genuine interaction between AP2A1  
21  
22 and PP2BA. However, STRING identified this PPI, because it is reported in the databases it  
23  
24 integrates. Our method, on the other hand, found it through the search for interologs, which  
25  
26 identified 13 PDB structures with 29% sequence identity. Interestingly, STRING does not  
27  
28 connect certain partners that our method identified as even closer interologs. For example, the  
29  
30 interaction between AP2A1 and 2A5G is supported by 5 PDB structures at 57% identity, and  
31  
32 that between TNKS1 and DYR1A is supported by both the 3D-based (36% identity) and  
33  
34 SLiMs-based analyses.  
35  
36  
37  
38  
39  
40  
41  
42  
43

44 Considering the results obtained with GeneMANIA, the 61 input proteins are all connected as  
45  
46 a single cluster, through 447 documented and 96 predicted PPIs. The integrative algorithm  
47  
48 includes 20 additional nodes to the network. Compared to the other methods, GeneMANIA  
49  
50 yields, by far, the highest number of connections. Remarkably, it does not generate any of the  
51  
52 aforementioned false negatives. However, the results are not exempt from false positives. For  
53  
54 example, as observed with the Interactome3D and STRING methods, the  $\gamma$ 1 and  $\gamma$ 2 subunits of  
55  
56  
57  
58  
59  
60

1  
2  
3 the AMPK complex (AAKG1 and AAKG2, respectively) are wrongly connected. The server  
4 also adds the AAKB2 node to the input data and connects it with AAKB1, although these two  
5 proteins are mutually exclusive partners of the AMPK complex. A similar example regards the  
6 PLOD3 protein mentioned above, which is connected to the newly added PLOD1, despite the  
7 absence of any documented interaction between these two proteins. Finally, although the  
8 WAVE and Arp2/3 complexes do interact within the Pragmin network, the connections  
9 established by GeneMANIA, both within and between these two complexes, are not  
10 “hierarchized”.  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22

23  
24 In summary, our protocol completely overlaps and extends the results of Interactome3D, in  
25 terms of interactions found in the two datasets. Conversely, it identified fewer associations than  
26 STRING or GeneMANIA, but these are more relevant regarding the definition of physical PPIs.  
27 The same trends were observed using the proteasome use-case, where STRING connects 163  
28 of the 192 input proteins, GeneMANIA connects all the 192 input proteins as a single cluster  
29 (through 3,490 physical and 2,470 predicted interactions), while both our protocol and  
30 Interactome3D connect 138 and 90 candidate partners, respectively. Note that because our use-  
31 cases corresponded to human sequences, an organism precalculated in the Interactome3D  
32 database, similar results obtained from Interactome3D and our protocol were expected.  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43

44 Altogether, the procedure we introduced here, applied to the two use-cases presented,  
45 performed as well or better than a reference approach such as Interactome3D. Furthermore, we  
46 expect it to have obvious added value in assisting users to analyse and to explore interactomics  
47 data. Notably, our protocol simplifies the identification of subsets of proteins participating to  
48 obligate interactions, it finds links between different such blocks, and it proposes putative  
49 interactions at a lower affinity involving SLiMS. Possible directions for methodological  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 improvement could encompass the explicit 3D modeling of the complexes, more detailed  
4 analysis of the protein-protein interfaces (e.g. in terms of co-evolution), or establishing  
5 connections with emerging technologies such as integrated top-down native proteomics<sup>74,75</sup>. In  
6 terms of applications, we believe that our pipeline will be of much use in the near future, due  
7 to the increasing number of interactomes performed using *in vivo* protein-proximity labeling  
8 (such as BioID<sup>76</sup>) or *in vivo* cross-linking approaches. Further developments of the pipeline  
9 could include (i) including quantitative information in the scoring of the (sub)-complexes and  
10 result outputs, as already proposed by other tools such as SAINT<sup>77</sup>, QPROT<sup>78</sup>, COMPASS<sup>79</sup>  
11 or TopS<sup>80</sup>, and (ii) integrating distance-restraints from cross-linking experiments for homology  
12 modelling, as performed by dedicated scoring functions<sup>81</sup> or bioinformatic suites such as the  
13 Integrative Modeling Platform (IMP)<sup>82</sup> or HADDOCK2.2 Web server<sup>83</sup>.

## 30 SUPPORTING INFORMATION

31 The following supporting information is available free of charge at ACS website

32 <http://pubs.acs.org>

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

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

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

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

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

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

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

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

## 57 ACKNOWLEDGEMENTS

1  
2  
3 This work was partly supported by the “Investissement d’Avenir Infrastructures Nationales en  
4 Biologie et Santé” grants ANR-10-INBS-05-01 (French Infrastructure for Integrated Structural  
5 Biology - FRISBI), ANR-10-INBS-08 (Proteomics French Infrastructure - ProFI) and ANR-  
6 11-INBS-0013 (French Institute of Bioinformatics - IFB).  
7  
8  
9  
10  
11  
12  
13

#### 14 **CONFLICT OF INTEREST**

15  
16  
17 The authors declare no conflict of interest.  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## REFERENCES

- 1  
2  
3  
4  
5  
6 (1) Luck, K.; Sheynkman, G. M.; Zhang, I.; Vidal, M. Proteome-Scale Human Interactomics. *Trends Biochem. Sci.* **2017**, *42* (5), 342–354.  
7 <https://doi.org/10.1016/j.tibs.2017.02.006>.
- 8  
9 (2) Rolland, T.; Taşan, M.; Charlotheaux, B.; Pevzner, S. J.; Zhong, Q.; Sahni, N.; Yi, S.;  
10 Lemmens, I.; Fontanillo, C.; Mosca, R.; Kamburov, A.; Ghiassian, S. D.; Yang, X.;  
11 Ghamsari, L.; Balcha, D.; Begg, B. E.; Braun, P.; Brehme, M.; Broly, M. P.; Carvunis, A.-  
12 R.; Convery-Zupan, D.; Corominas, R.; Coulombe-Huntington, J.; Dann, E.; Dreze, M.;  
13 Dricot, A.; Fan, C.; Franzosa, E.; Gebreab, F.; Gutierrez, B. J.; Hardy, M. F.; Jin, M.;  
14 Kang, S.; Kiros, R.; Lin, G. N.; Luck, K.; MacWilliams, A.; Menche, J.; Murray, R. R.;  
15 Palagi, A.; Poulin, M. M.; Rambout, X.; Rasla, J.; Reichert, P.; Romero, V.; Ruysinck, E.;  
16 Sahalie, J. M.; Scholz, A.; Shah, A. A.; Sharma, A.; Shen, Y.; Spirohn, K.; Tam, S.; Tejada,  
17 A. O.; Trigg, S. A.; Twizere, J.-C.; Vega, K.; Walsh, J.; Cusick, M. E.; Xia, Y.; Barabási, A.-  
18 L.; Iakoucheva, L. M.; Aloy, P.; De Las Rivas, J.; Tavernier, J.; Calderwood, M. A.; Hill, D.  
19 E.; Hao, T.; Roth, F. P.; Vidal, M. A Proteome-Scale Map of the Human Interactome  
20 Network. *Cell* **2014**, *159* (5), 1212–1226. <https://doi.org/10.1016/j.cell.2014.10.050>.
- 21  
22 (3) Chien, C. T.; Bartel, P. L.; Sternglanz, R.; Fields, S. The Two-Hybrid System: A Method to  
23 Identify and Clone Genes for Proteins That Interact with a Protein of Interest. *Proc.*  
24 *Natl. Acad. Sci.* **1991**, *88* (21), 9578–9582. <https://doi.org/10.1073/pnas.88.21.9578>.
- 25  
26 (4) Brückner, A.; Polge, C.; Lentze, N.; Auerbach, D.; Schlattner, U. Yeast Two-Hybrid, a  
27 Powerful Tool for Systems Biology. *Int. J. Mol. Sci.* **2009**, *10* (6), 2763–2788.  
28 <https://doi.org/10.3390/ijms10062763>.
- 29  
30 (5) Tavernier, J.; Eyckerman, S.; Lemmens, I.; Heyden, J. V. der; Vandekerckhove, J.;  
31 Ostade, X. V. MAPPIT: A Cytokine Receptor-Based Two-Hybrid Method in Mammalian  
32 Cells1. *Clin. Exp. Allergy* **2002**, *32* (10), 1397–1404. <https://doi.org/10.1046/j.1365-2745.2002.01520.x>.
- 33  
34 (6) Nyfeler, B.; Michnick, S. W.; Hauri, H.-P. Capturing Protein Interactions in the  
35 Secretory Pathway of Living Cells. *Proc. Natl. Acad. Sci.* **2005**, *102* (18), 6350–6355.  
36 <https://doi.org/10.1073/pnas.0501976102>.
- 37  
38 (7) Bauer, A.; Kuster, B. Affinity Purification-Mass Spectrometry. *Eur. J. Biochem.* **2003**,  
39 *270* (4), 570–578. <https://doi.org/10.1046/j.1432-1033.2003.03428.x>.
- 40  
41 (8) Dunham, W. H.; Mullin, M.; Gingras, A.-C. Affinity-Purification Coupled to Mass  
42 Spectrometry: Basic Principles and Strategies. *Proteomics* **2012**, *12* (10), 1576–1590.  
43 <https://doi.org/10.1002/pmic.201100523>.
- 44  
45 (9) Gingras, A.-C.; Abe, K. T.; Raught, B. Getting to Know the Neighborhood: Using  
46 Proximity-Dependent Biotinylation to Characterize Protein Complexes and Map  
47 Organelles. *Curr. Opin. Chem. Biol.* **2019**, *48*, 44–54.  
48 <https://doi.org/10.1016/j.cbpa.2018.10.017>.
- 49  
50 (10) Fabre, B.; Lambour, T.; Garrigues, L.; Amalric, F.; Vigneron, N.; Menneteau, T.; Stella,  
51 A.; Monsarrat, B.; Van den Eynde, B.; Burlet-Schiltz, O.; Bousquet-Dubouch, M.-P.  
52 Deciphering Preferential Interactions within Supramolecular Protein Complexes: The  
53 Proteasome Case. *Mol. Syst. Biol.* **2015**, *11* (1), 771.  
54 <https://doi.org/10.15252/msb.20145497>.
- 55  
56 (11) Chavez, J. D.; Mohr, J. P.; Mathay, M.; Zhong, X.; Keller, A.; Bruce, J. E. Systems  
57 Structural Biology Measurements by in Vivo Cross-Linking with Mass Spectrometry.  
58  
59  
60

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60
- 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:

- 1  
2  
3 Protein–Protein Association Networks with Increased Coverage, Supporting Functional  
4 Discovery in Genome-Wide Experimental Datasets. *Nucleic Acids Res.* **2019**, *47* (D1),  
5 D607–D613. <https://doi.org/10.1093/nar/gky1131>.  
6  
7 (24) Franz, M.; Rodriguez, H.; Lopes, C.; Zuberi, K.; Montojo, J.; Bader, G. D.; Morris, Q.  
8 GeneMANIA Update 2018. *Nucleic Acids Res.* **2018**, *46* (W1), W60–W64.  
9 <https://doi.org/10.1093/nar/gky311>.  
10  
11 (25) Oughtred, R.; Stark, C.; Breitkreutz, B.-J.; Rust, J.; Boucher, L.; Chang, C.; Kolas, N.;  
12 O'Donnell, L.; Leung, G.; McAdam, R.; Zhang, F.; Dolma, S.; Willems, A.; Coulombe-  
13 Huntington, J.; Chatr-aryamontri, A.; Dolinski, K.; Tyers, M. The BioGRID Interaction  
14 Database: 2019 Update. *Nucleic Acids Res.* **2019**, *47* (D1), D529–D541.  
15 <https://doi.org/10.1093/nar/gky1079>.  
16  
17 (26) Rodchenkov, I.; Babur, O.; Luna, A.; Aksoy, B. A.; Wong, J. V.; Fong, D.; Franz, M.; Siper,  
18 M. C.; Cheung, M.; Wrana, M.; Mistry, H.; Mosier, L.; Dlin, J.; Wen, Q.; O'Callaghan, C.;  
19 Li, W.; Elder, G.; Smith, P. T.; Dallago, C.; Cerami, E.; Gross, B.; Dogrusoz, U.; Demir, E.;  
20 Bader, G. D.; Sander, C. Pathway Commons 2019 Update: Integration, Analysis and  
21 Exploration of Pathway Data. *Nucleic Acids Res.* **2020**, *48* (D1), D489–D497.  
22 <https://doi.org/10.1093/nar/gkz946>.  
23  
24 (27) Pons, J.-L.; Labesse, G. @TOME-2: A New Pipeline for Comparative Modeling of  
25 Protein–Ligand Complexes. *Nucleic Acids Res.* **2009**, *37* (suppl\_2), W485–W491.  
26 <https://doi.org/10.1093/nar/gkp368>.  
27  
28 (28) Söding, J. Protein Homology Detection by HMM–HMM Comparison. *Bioinformatics*  
29 **2005**, *21* (7), 951–960. <https://doi.org/10.1093/bioinformatics/bti125>.  
30  
31 (29) Konagurthu, A. S.; Whisstock, J. C.; Stuckey, P. J.; Lesk, A. M. MUSTANG: A Multiple  
32 Structural Alignment Algorithm. *Proteins Struct. Funct. Bioinforma.* **2006**, *64* (3), 559–  
33 574. <https://doi.org/10.1002/prot.20921>.  
34  
35 (30) Katoh, K.; Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7:  
36 Improvements in Performance and Usability. *Mol. Biol. Evol.* **2013**, *30* (4), 772–780.  
37 <https://doi.org/10.1093/molbev/mst010>.  
38  
39 (31) Kumar, M.; Gouw, M.; Michael, S.; Sámano-Sánchez, H.; Panca, R.; Glavina, J.;  
40 Diakogianni, A.; Valverde, J. A.; Bukirova, D.; Čalyševa, J.; Palopoli, N.; Davey, N. E.;  
41 Chemes, L. B.; Gibson, T. J. ELM—the Eukaryotic Linear Motif Resource in 2020.  
42 *Nucleic Acids Res.* **2020**, *48* (D1), D296–D306. <https://doi.org/10.1093/nar/gkz1030>.  
43  
44 (32) Mészáros, B.; Erdős, G.; Dosztányi, Z. IUPred2A: Context-Dependent Prediction of  
45 Protein Disorder as a Function of Redox State and Protein Binding. *Nucleic Acids Res.*  
46 **2018**, *46* (W1), W329–W337. <https://doi.org/10.1093/nar/gky384>.  
47  
48 (33) Fabre, B.; Lambour, T.; Delobel, J.; Amalric, F.; Monsarrat, B.; Burlet-Schiltz, O.;  
49 Bousquet-Dubouch, M.-P. Subcellular Distribution and Dynamics of Active Proteasome  
50 Complexes Unraveled by a Workflow Combining in Vivo Complex Cross-Linking and  
51 Quantitative Proteomics. *Mol. Cell. Proteomics* **2013**, *12* (3), 687–699.  
52 <https://doi.org/10.1074/mcp.M112.023317>.  
53  
54 (34) Lecointre, C.; Simon, V.; Kerneur, C.; Allemand, F.; Fournet, A.; Montarras, I.; Pons, J.-  
55 L.; Gelin, M.; Brignatz, C.; Urbach, S.; Labesse, G.; Roche, S. Dimerization of the  
56 Pragmin Pseudo-Kinase Regulates Protein Tyrosine Phosphorylation. *Structure* **2018**,  
57 *26* (4), 545–554.e4. <https://doi.org/10.1016/j.str.2018.01.017>.  
58  
59 (35) Leroy, C.; Fialin, C.; Sirvent, A.; Simon, V.; Urbach, S.; Poncet, J.; Robert, B.; Jouin, P.;  
60

- 1  
2  
3 Roche, S. Quantitative Phosphoproteomics Reveals a Cluster of Tyrosine Kinases That  
4 Mediates Src Invasive Activity in Advanced Colon Carcinoma Cells. *Cancer Res.* **2009**,  
5 *69* (6), 2279–2286. <https://doi.org/10.1158/0008-5472.CAN-08-2354>.  
6  
7 (36) Liu, L.; Phua, Y. W.; Lee, R. S.; Ma, X.; Jenkins, Y.; Novy, K.; Humphrey, E. S.; Chan, H.;  
8 Shearer, R.; Ong, P. C.; Dai, W.; Saunders, D. N.; Lucet, I. S.; Daly, R. J. Homo- and  
9 Heterotypic Association Regulates Signaling by the SgK269/PEAK1 and SgK223  
10 Pseudokinases. *J. Biol. Chem.* **2016**, *291* (41), 21571–21583.  
11 <https://doi.org/10.1074/jbc.M116.748897>.  
12  
13 (37) Fabre, B.; Lambour, T.; Garrigues, L.; Ducoux-Petit, M.; Amalric, F.; Monsarrat, B.;  
14 Burlet-Schiltz, O.; Bousquet-Dubouch, M.-P. Label-Free Quantitative Proteomics  
15 Reveals the Dynamics of Proteasome Complexes Composition and Stoichiometry in a  
16 Wide Range of Human Cell Lines. *J. Proteome Res.* **2014**, *13* (6), 3027–3037.  
17 <https://doi.org/10.1021/pr500193k>.  
18  
19 (38) Matyskiela, M. E.; Lander, G. C.; Martin, A. Conformational Switching of the 26S  
20 Proteasome Enables Substrate Degradation. *Nat. Struct. Mol. Biol.* **2013**, *20* (7), 781–  
21 788. <https://doi.org/10.1038/nsmb.2616>.  
22  
23 (39) Rêgo, A. T.; Fonseca, P. C. A. da. Characterization of Fully Recombinant Human 20S  
24 and 20S-PA200 Proteasome Complexes. *Mol. Cell* **2019**, *76* (1), 138-147.e5.  
25 <https://doi.org/10.1016/j.molcel.2019.07.014>.  
26  
27 (40) Wehmer, M.; Rudack, T.; Beck, F.; Aufderheide, A.; Pfeifer, G.; Plitzko, J. M.; Förster,  
28 F.; Schulten, K.; Baumeister, W.; Sakata, E. Structural Insights into the Functional Cycle  
29 of the ATPase Module of the 26S Proteasome. *Proc. Natl. Acad. Sci.* **2017**.  
30 <https://doi.org/10.1073/pnas.1621129114>.  
31  
32 (41) Huber, E. M.; Basler, M.; Schwab, R.; Heinemeyer, W.; Kirk, C. J.; Groettrup, M.; Groll,  
33 M. Immuno- and Constitutive Proteasome Crystal Structures Reveal Differences in  
34 Substrate and Inhibitor Specificity. *Cell* **2012**, *148* (4), 727–738.  
35 <https://doi.org/10.1016/j.cell.2011.12.030>.  
36  
37 (42) Ruschak, A. M.; Kay, L. E. Proteasome Allostery as a Population Shift between  
38 Interchanging Conformers. *Proc. Natl. Acad. Sci.* **2012**.  
39 <https://doi.org/10.1073/pnas.1213640109>.  
40  
41 (43) Gersch, M.; Hackl, M. W.; Dubiella, C.; Dobrinevski, A.; Groll, M.; Sieber, S. A. A Mass  
42 Spectrometry Platform for a Streamlined Investigation of Proteasome Integrity,  
43 Posttranslational Modifications, and Inhibitor Binding. *Chem. Biol.* **2015**, *22* (3), 404–  
44 411. <https://doi.org/10.1016/j.chembiol.2015.01.004>.  
45  
46 (44) Lesne, J.; Bousquet, M.-P.; Marcoux, J.; Locard-Paulet, M. Top-Down and Intact  
47 Protein Mass Spectrometry Data Visualization for Proteoform Analysis Using VisioProt-  
48 MS. *Bioinforma. Biol. Insights* **2019**, *13*, 1177932219868223.  
49 <https://doi.org/10.1177/1177932219868223>.  
50  
51 (45) Jonik-Nowak, B.; Menneteau, T.; Fesquet, D.; Baldin, V.; Bonne-Andrea, C.; Méchali, F.;  
52 Fabre, B.; Boisguerin, P.; Rossi, S. de; Henriquet, C.; Pugnière, M.; Ducoux-Petit, M.;  
53 Burlet-Schiltz, O.; Lamond, A. I.; Fort, P.; Boulon, S.; Bousquet, M.-P.; Coux, O.  
54 PIP30/FAM192A Is a Novel Regulator of the Nuclear Proteasome Activator PA28γ.  
55 *Proc. Natl. Acad. Sci.* **2018**, *115* (28), E6477–E6486.  
56 <https://doi.org/10.1073/pnas.1722299115>.  
57  
58 (46) Moscovitz, O.; Ben-Nissan, G.; Fainer, I.; Pollack, D.; Mizrachi, L.; Sharon, M. The  
59  
60

- 1  
2  
3 Parkinson's-Associated Protein DJ-1 Regulates the 20S Proteasome. *Nat. Commun.*  
4 **2015**, *6* (1), 1–13. <https://doi.org/10.1038/ncomms7609>.  
5  
6 (47) Olshina, M. A.; Arkind, G.; Kumar Deshmukh, F.; Fainer, I.; Taranavsky, M.; Hayat, D.;  
7 Ben-Dor, S.; Ben-Nissan, G.; Sharon, M. Regulation of the 20S Proteasome by a Novel  
8 Family of Inhibitory Proteins. *Antioxid. Redox Signal.* **2020**, *32* (9), 636–655.  
9 <https://doi.org/10.1089/ars.2019.7816>.  
10  
11 (48) Lasker, K.; Förster, F.; Bohn, S.; Walzthoeni, T.; Villa, E.; Unverdorben, P.; Beck, F.;  
12 Aebersold, R.; Sali, A.; Baumeister, W. Molecular Architecture of the 26S Proteasome  
13 Holocomplex Determined by an Integrative Approach. *Proc. Natl. Acad. Sci.* **2012**.  
14 <https://doi.org/10.1073/pnas.1120559109>.  
15  
16 (49) Wang, X.; Cimermancic, P.; Yu, C.; Schweitzer, A.; Chopra, N.; Engel, J. L.; Greenberg, C.  
17 H.; Huszagh, A. S.; Beck, F.; Sakata, E.; Yang, Y.; Novitsky, E. J.; Leitner, A.; Nanni, P.;  
18 Kahraman, A.; Guo, X.; Dixon, J. E.; Rychnovsky, S. D.; Aebersold, R.; Baumeister, W.;  
19 Sali, A.; Huang, L. Molecular Details Underlying Dynamic Structures and Regulation of  
20 the Human 26S Proteasome. *Mol. Cell. Proteomics* **2017**.  
21 <https://doi.org/10.1074/mcp.M116.065326>.  
22  
23 (50) Wang, X.; Chemmama, I. E.; Yu, C.; Huszagh, A.; Xu, Y.; Viner, R.; Block, S. A.;  
24 Cimermancic, P.; Rychnovsky, S. D.; Ye, Y.; Sali, A.; Huang, L. The Proteasome-  
25 Interacting Ecm29 Protein Disassembles the 26S Proteasome in Response to Oxidative  
26 Stress. *J. Biol. Chem.* **2017**, jbc.M117.803619.  
27 <https://doi.org/10.1074/jbc.M117.803619>.  
28  
29 (51) Wang, X.; Huang, L. Dissecting Dynamic and Heterogeneous Proteasome Complexes  
30 Using In Vivo Cross-Linking-Assisted Affinity Purification and Mass Spectrometry. In  
31 *The Ubiquitin Proteasome System: Methods and Protocols*; Mayor, T., Kleiger, G., Eds.;  
32 Methods in Molecular Biology; Springer: New York, NY, 2018; pp 401–410.  
33 [https://doi.org/10.1007/978-1-4939-8706-1\\_25](https://doi.org/10.1007/978-1-4939-8706-1_25).  
34  
35 (52) Huber, E. M.; Groll, M. The Mammalian Proteasome Activator PA28 Forms an  
36 Asymmetric A4 $\beta$ 3 Complex. *Struct. Lond. Engl.* **1993** **2017**, *25* (10), 1473–1480.e3.  
37 <https://doi.org/10.1016/j.str.2017.07.013>.  
38  
39 (53) Groll, M.; Brandstetter, H.; Bartunik, H.; Bourenkow, G.; Huber, R. Investigations on  
40 the Maturation and Regulation of Archaeobacterial Proteasomes<sup>††</sup>We Dedicate This  
41 Paper to the Memory of Eraldo Antonini, Eminent Biochemist, Prematurely Deceased  
42 20 Years Ago, on 19 March 1983. *J. Mol. Biol.* **2003**, *327* (1), 75–83.  
43 [https://doi.org/10.1016/S0022-2836\(03\)00080-9](https://doi.org/10.1016/S0022-2836(03)00080-9).  
44  
45 (54) Huang, X.; Luan, B.; Wu, J.; Shi, Y. An Atomic Structure of the Human 26S Proteasome.  
46 *Nat. Struct. Mol. Biol.* **2016**, *23* (9), 778–785. <https://doi.org/10.1038/nsmb.3273>.  
47  
48 (55) Harshbarger, W.; Miller, C.; Diedrich, C.; Sacchettini, J. Crystal Structure of the Human  
49 20S Proteasome in Complex with Carfilzomib. *Structure* **2015**, *23* (2), 418–424.  
50 <https://doi.org/10.1016/j.str.2014.11.017>.  
51  
52 (56) Santos, R. de L. A.; Bai, L.; Singh, P. K.; Murakami, N.; Fan, H.; Zhan, W.; Zhu, Y.; Jiang,  
53 X.; Zhang, K.; Assker, J. P.; Nathan, C. F.; Li, H.; Azzi, J.; Lin, G. Structure of Human  
54 Immunoproteasome with a Reversible and Noncompetitive Inhibitor That Selectively  
55 Inhibits Activated Lymphocytes. *Nat. Commun.* **2017**, *8* (1), 1–11.  
56 <https://doi.org/10.1038/s41467-017-01760-5>.  
57  
58 (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.; Mercedi, 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>.

- 1  
2  
3 (67) Chen, B.; Brinkmann, K.; Chen, Z.; Pak, C. W.; Liao, Y.; Shi, S.; Henry, L.; Grishin, N. V.;  
4 Bogdan, S.; Rosen, M. K. The WAVE Regulatory Complex Links Diverse Receptors to  
5 the Actin Cytoskeleton. *Cell* **2014**, *156* (1), 195–207.  
6 <https://doi.org/10.1016/j.cell.2013.11.048>.  
7  
8 (68) Robinson, R. C.; Turbedsky, K.; Kaiser, D. A.; Marchand, J.-B.; Higgs, H. N.; Choe, S.;  
9 Pollard, T. D. Crystal Structure of Arp2/3 Complex. *Science* **2001**, *294* (5547), 1679–  
10 1684. <https://doi.org/10.1126/science.1066333>.  
11  
12 (69) Nolen, B. J.; Tomasevic, N.; Russell, A.; Pierce, D. W.; Jia, Z.; McCormick, C. D.;  
13 Hartman, J.; Sakowicz, R.; Pollard, T. D. Characterization of Two Classes of Small  
14 Molecule Inhibitors of Arp2/3 Complex. *Nature* **2009**, *460* (7258), 1031–1034.  
15 <https://doi.org/10.1038/nature08231>.  
16  
17 (70) Werfel, T. A.; Elion, D. L.; Rahman, B.; Hicks, D. J.; Sanchez, V.; Gonzalez-Ericsson, P. I.;  
18 Nixon, M. J.; James, J. L.; Balko, J. M.; Scherle, P.; Koblisch, H. K.; Cook, R. S. Treatment-  
19 Induced Tumor Cell Apoptosis and Secondary Necrosis Drive Tumor Progression in the  
20 Residual Tumor Microenvironment through MerTK and IDO-1. *Cancer Res.* **2018**.  
21 <https://doi.org/10.1158/0008-5472.CAN-18-1106>.  
22  
23 (71) Shomura, Y.; Dragovic, Z.; Chang, H.-C.; Tzvetkov, N.; Young, J. C.; Brodsky, J. L.;  
24 Guerriero, V.; Hartl, F. U.; Bracher, A. Regulation of Hsp70 Function by HspBP1:  
25 Structural Analysis Reveals an Alternate Mechanism for Hsp70 Nucleotide Exchange.  
26 *Mol. Cell* **2005**, *17* (3), 367–379. <https://doi.org/10.1016/j.molcel.2004.12.023>.  
27  
28 (72) Schuermann, J. P.; Jiang, J.; Cuellar, J.; Llorca, O.; Wang, L.; Gimenez, L. E.; Jin, S.;  
29 Taylor, A. B.; Demeler, B.; Morano, K. A.; Hart, P. J.; Valpuesta, J. M.; Lafer, E. M.;  
30 Sousa, R. Structure of the Hsp110:Hsc70 Nucleotide Exchange Machine. *Mol. Cell*  
31 **2008**, *31* (2), 232–243. <https://doi.org/10.1016/j.molcel.2008.05.006>.  
32  
33 (73) Lopez, M. L.; Lo, M.; Kung, J. E.; Dudkiewicz, M.; Jang, G. M.; Dollen, J. V.; Johnson, J.  
34 R.; Krogan, N. J.; Pawłowski, K.; Jura, N. PEAK3/C19orf35 Pseudokinase, a New NFK3  
35 Kinase Family Member, Inhibits Crkl through Dimerization. *Proc. Natl. Acad. Sci.* **2019**,  
36 *116* (31), 15495–15504. <https://doi.org/10.1073/pnas.1906360116>.  
37  
38 (74) Li, H.; Nguyen, H. H.; Ogorzalek Loo, R. R.; Campuzano, I. D. G.; Loo, J. A. An Integrated  
39 Native Mass Spectrometry and Top-down Proteomics Method That Connects  
40 Sequence to Structure and Function of Macromolecular Complexes. *Nat. Chem.* **2018**,  
41 *10* (2), 139–148. <https://doi.org/10.1038/nchem.2908>.  
42  
43 (75) Skinner, O. S.; Haverland, N. A.; Fornelli, L.; Melani, R. D.; Do Vale, L. H. F.; Seckler, H.  
44 S.; Doubleday, P. F.; Schachner, L. F.; Srzentić, K.; Kelleher, N. L.; Compton, P. D. Top-  
45 down Characterization of Endogenous Protein Complexes with Native Proteomics.  
46 *Nat. Chem. Biol.* **2018**, *14* (1), 36–41. <https://doi.org/10.1038/nchembio.2515>.  
47  
48 (76) Lambert, J.-P.; Tucholska, M.; Go, C.; Knight, J. D. R.; Gingras, A.-C. Proximity  
49 Biotinylation and Affinity Purification Are Complementary Approaches for the  
50 Interactome Mapping of Chromatin-Associated Protein Complexes. *J. Proteomics*  
51 **2015**, *118*, 81–94. <https://doi.org/10.1016/j.jprot.2014.09.011>.  
52  
53 (77) Choi, H.; Larsen, B.; Lin, Z.-Y.; Breikreutz, A.; Mellacheruvu, D.; Fermin, D.; Qin, Z. S.;  
54 Tyers, M.; Gingras, A.-C.; Nesvizhskii, A. I. SAINT: Probabilistic Scoring of Affinity  
55 Purification–Mass Spectrometry Data. *Nat. Methods* **2011**, *8* (1), 70–73.  
56 <https://doi.org/10.1038/nmeth.1541>.  
57  
58 (78) Choi, H.; Kim, S.; Fermin, D.; Tsou, C.-C.; Nesvizhskii, A. I. QPROT: Statistical Method  
59  
60

- 1  
2  
3 for Testing Differential Expression Using Protein-Level Intensity Data in Label-Free  
4 Quantitative Proteomics. *J. Proteomics* **2015**, *129*, 121–126.  
5 <https://doi.org/10.1016/j.jprot.2015.07.036>.  
6  
7 (79) Wenger, C. D.; Phanstiel, D. H.; Lee, M. V.; Bailey, D. J.; Coon, J. J. COMPASS: A Suite of  
8 Pre- and Post-Search Proteomics Software Tools for OMSSA. *PROTEOMICS* **2011**, *11*  
9 (6), 1064–1074. <https://doi.org/10.1002/pmic.201000616>.  
10  
11 (80) Sardu, M. E.; Gilmore, J. M.; Groppe, B. D.; Dutta, A.; Florens, L.; Washburn, M. P.  
12 Topological Scoring of Protein Interaction Networks. *Nat. Commun.* **2019**, *10* (1), 1–14.  
13 <https://doi.org/10.1038/s41467-019-09123-y>.  
14  
15 (81) Bullock, J. M. A.; Sen, N.; Thalassinou, K.; Topf, M. Modeling Protein Complexes Using  
16 Restraints from Crosslinking Mass Spectrometry. *Structure* **2018**, *26* (7), 1015–1024.e2.  
17 <https://doi.org/10.1016/j.str.2018.04.016>.  
18  
19 (82) Webb, B.; Lasker, K.; Velázquez-Muriel, J.; Schneidman-Duhovny, D.; Pellarin, R.;  
20 Bonomi, M.; Greenberg, C.; Raveh, B.; Tjioe, E.; Russel, D.; Sali, A. Modeling of Proteins  
21 and Their Assemblies with the Integrative Modeling Platform. In *Structural Genomics:*  
22 *General Applications*; Chen, Y. W., Ed.; Methods in Molecular Biology; Humana Press:  
23 Totowa, NJ, 2014; pp 277–295. [https://doi.org/10.1007/978-1-62703-691-7\\_20](https://doi.org/10.1007/978-1-62703-691-7_20).  
24  
25 (83) van Zundert, G. C. P.; Rodrigues, J. P. G. L. M.; Trellet, M.; Schmitz, C.; Kastiris, P. L.;  
26 Karaca, E.; Melquiond, A. S. J.; van Dijk, M.; de Vries, S. J.; Bonvin, A. M. J. J. The  
27 HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular  
28 Complexes. *J. Mol. Biol.* **2016**, *428* (4), 720–725.  
29 <https://doi.org/10.1016/j.jmb.2015.09.014>.  
30  
31 (84) Kazakou, K.; Holloway, D. E.; Prior, S. H.; Subramanian, V.; Acharya, K. R. Ribonuclease  
32 A Homologues of the Zebrafish: Polymorphism, Crystal Structures of Two  
33 Representatives and Their Evolutionary Implications. *J. Mol. Biol.* **2008**, *380* (1), 206–  
34 222. <https://doi.org/10.1016/j.jmb.2008.04.070>.  
35  
36 (85) Xiao, B.; Sanders, M. J.; Carmena, D.; Bright, N. J.; Haire, L. F.; Underwood, E.; Patel, B.  
37 R.; Heath, R. B.; Walker, P. A.; Hallen, S.; Giordanetto, F.; Martin, S. R.; Carling, D.;  
38 Gamblin, S. J. Structural Basis of AMPK Regulation by Small Molecule Activators. *Nat.*  
39 *Commun.* **2013**, *4* (1), 1–10. <https://doi.org/10.1038/ncomms4017>.  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**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<sup>84</sup>) 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)**

1  
2  
3 Complex ID c074 identified a homologous complex structure for a set of subunits partly  
4 overlapping those in (A), namely the 20S subunits. Remote homology relationships could be  
5 detected at even higher sequence divergence between two proteasome chaperones (PSMG1 and  
6 PSMG2) for which a structure of the orthologous subunits was available in yeast (PDB:4G4S).  
7  
8  
9  
10  
11  
12 (C) Complex ID c097 identified a homologous complex for two subunits involving homologs  
13 of PSA5 and PSMG3, the latter sharing high sequence divergence (19 % sequence identity)  
14 with its yeast ortholog, named POC3. Interestingly in the yeast complex (PDB:2Z5C), a third  
15 subunit, named POC4, is found tightly bound suggesting that a homologous subunit could have  
16 been detected in human. The remote homologs of yeast POC4, named PSMG4 in human, was  
17 not present in the list of input proteins suggesting potential issues in the detection threshold or  
18 in the capacity of spotting out this protein by mass spectrometry.  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29

30  
31 **Figure 4.** Graph representation of the PPI network identified for the proteasome dataset. Nodes  
32 colored in cyan, magenta and orange correspond to input proteins, nodes added by the PDB  
33 search and BioGRID, respectively. The bait protein (PSA2) is represented by a green node.  
34  
35 Four types of edges represent interactions between proteins. (i) Thick, opaque: Edges identified  
36 by the structure-based analysis and connecting input proteins belonging to hetero-multimers.  
37  
38 The colors (blue, green, yellow, red, black) depend on the evolutionary distance (sequence  
39 identity  $\geq$ 95%, 80%, 50%, 30%, 0%, respectively) between the two input proteins and their  
40 homologs/paralogs. (ii) Thin, magenta: Edges produced by the structure-based analysis and  
41 connecting input proteins and undetected partners. (iii) Thin, blue: Edges produced by the ELM  
42 analysis and connecting input proteins. (iv) Thin and orange: Edges produced by the BioGRID  
43 analysis and connecting input proteins and additional partners from BioGRID.  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 **Figure 5.** Cryo-EM structure (PDB: 5GJQ) of the 20S catalytic core (blue) in complex with the  
4 19S regulator (green) forming the ATP-ase dependent 26S proteasome bound to USP14 & UBB  
5 (purple). The two copies of the immuno-purification bait ( $\alpha 2$ ) are represented in red. All these  
6 33 proteins were immuno-purified. Structures of the 20S (red and blue), the 19S (green), the  
7 26S (red, blue and green) and the 26S bound to USP14 & UBB (red, blue, green and purple)  
8 can be found in the PDB, making possible to distinguish three layers of interaction with the  
9 bait.  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22

23  
24 **Figure 6.** Graph representation of the PPI network identified for the Pragmin dataset. For this  
25 representation, the bait protein (PRAG1) has been added to the original dataset (green node).  
26  
27  
28  
29

30  
31 **Figure 7.** Space-filling representation of protein complexes of the Pramgin use-case, as  
32 identified by the pipeline. (A) The pentameric structure of the WAVE regulatory complex (PDB  
33 entry 3p8c), with the subunits WASF1 and BRK1, missing in AP-MS data, colored in white  
34 and black, respectively. (B) The heptameric structure of the Arp2/3 complex (PDB entry 1k8k),  
35 with the subunits ARPC1B (white), ARPC3 (black), and ARPC5 (gray). (C) The  $\alpha 1\beta 2\gamma 1$  and  
36 (D)  $\alpha 2\beta 1\gamma 1$  isoforms of the AMPK complex, both found by our integrative approach (PDB  
37 entries 2vq8<sup>84</sup> and 4cfe<sup>85</sup>, respectively); both structures are aligned on the  $\gamma 1$  subunit (colored  
38 in purple).  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

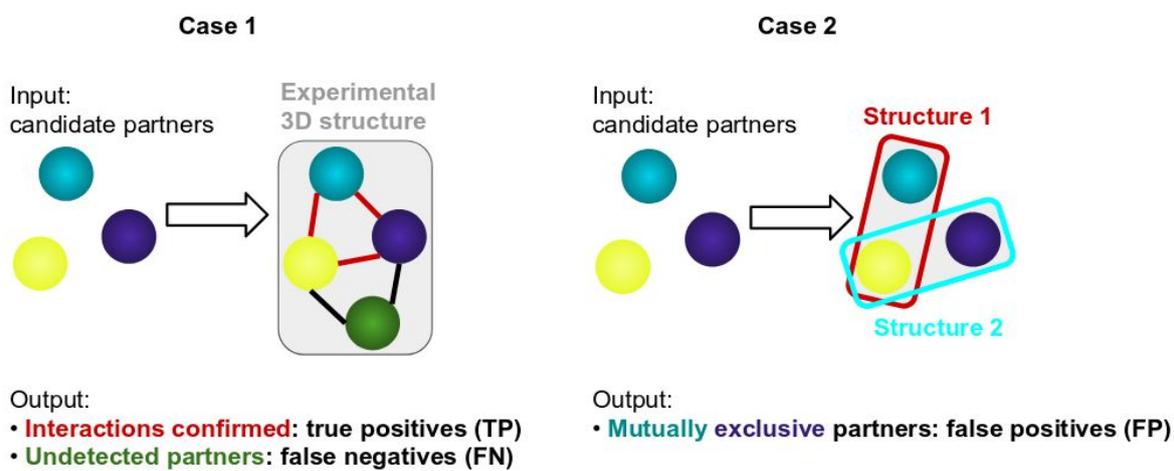


Figure 1

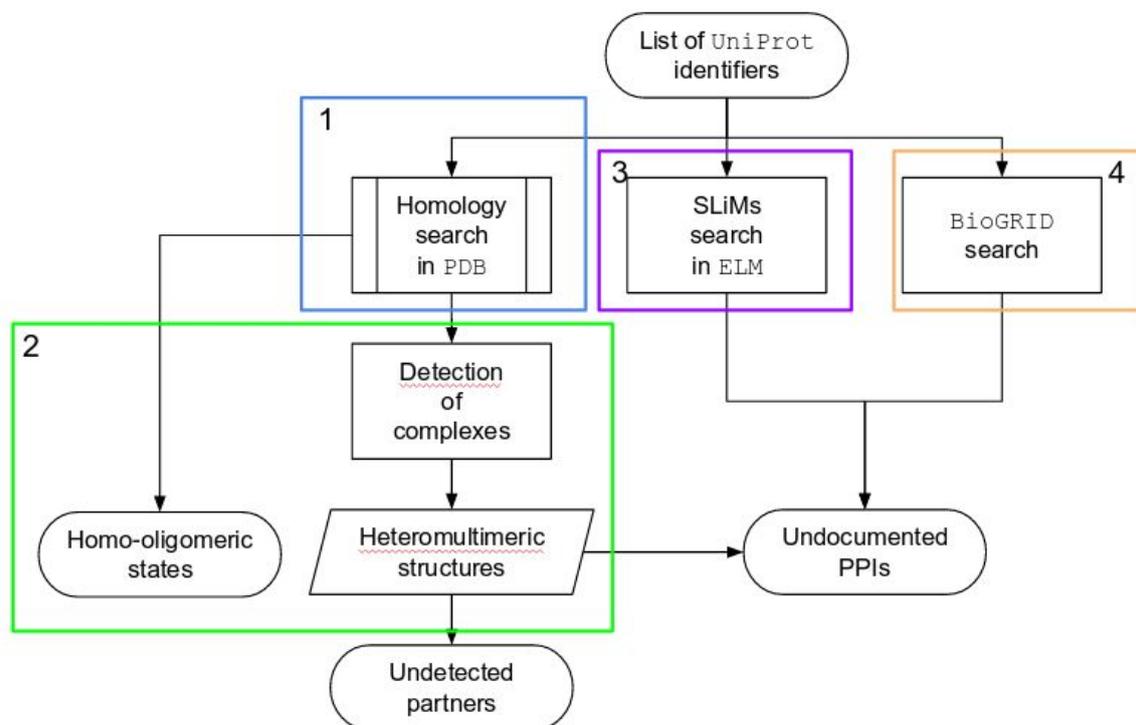


Figure 2

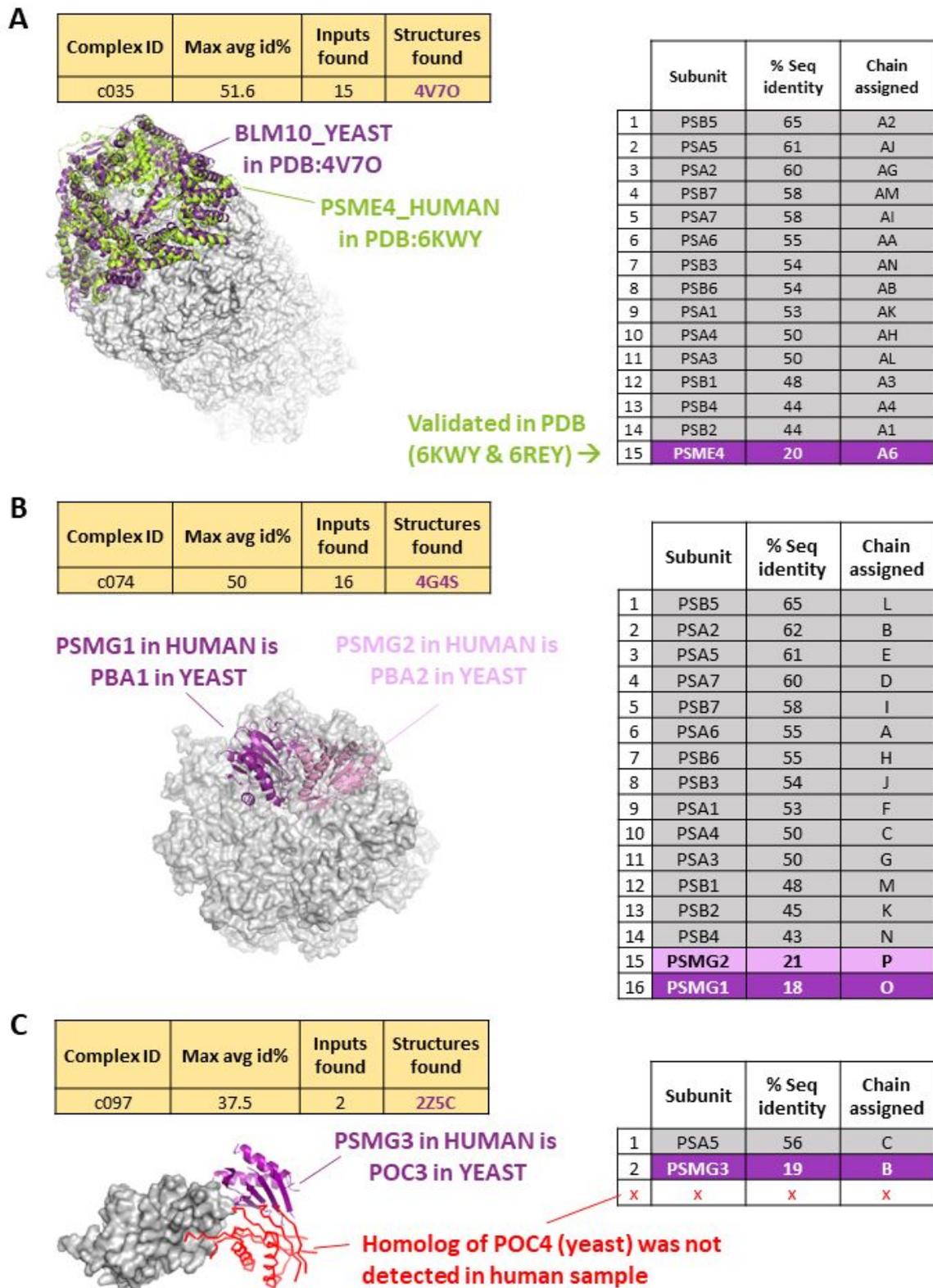
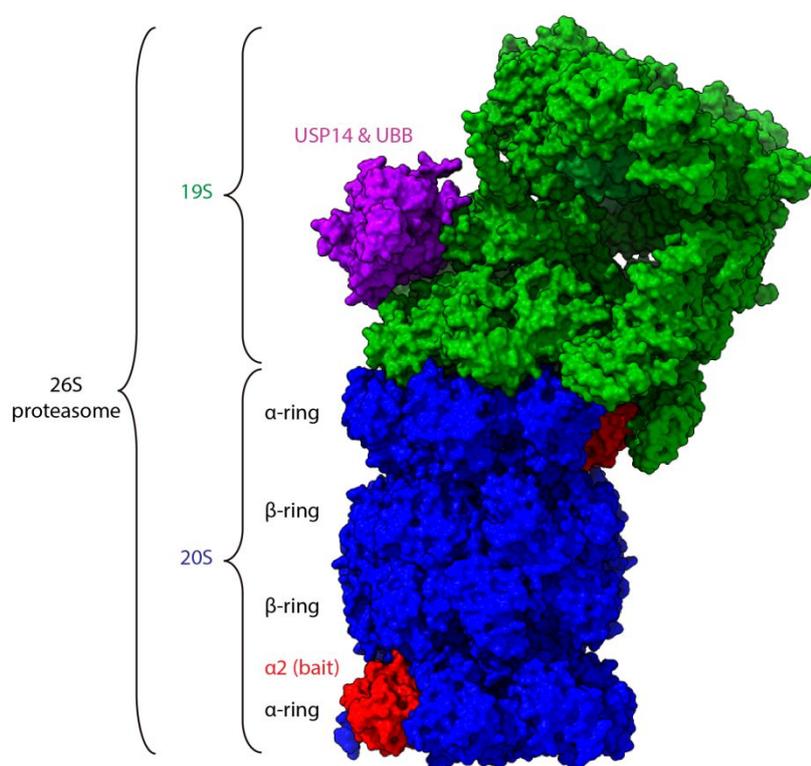


Figure 3





1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Figure 5

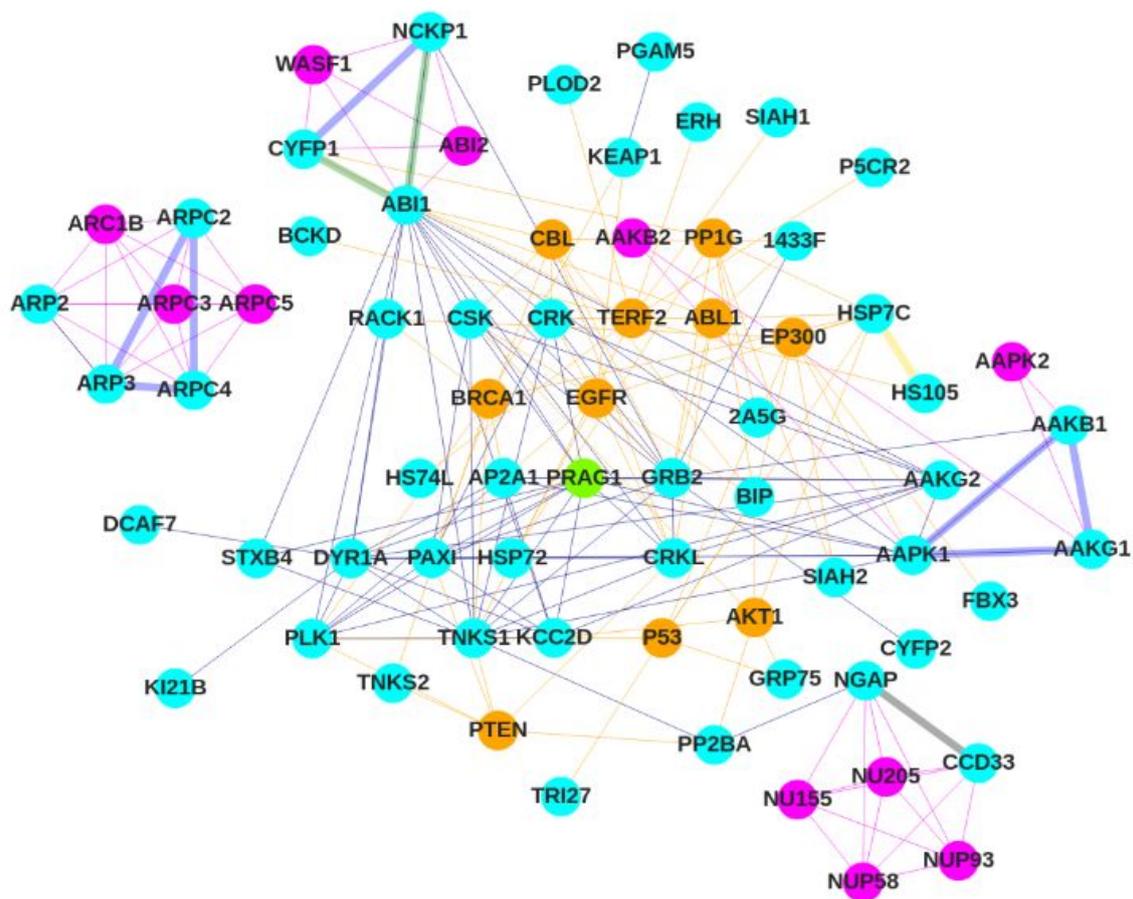


Figure 6

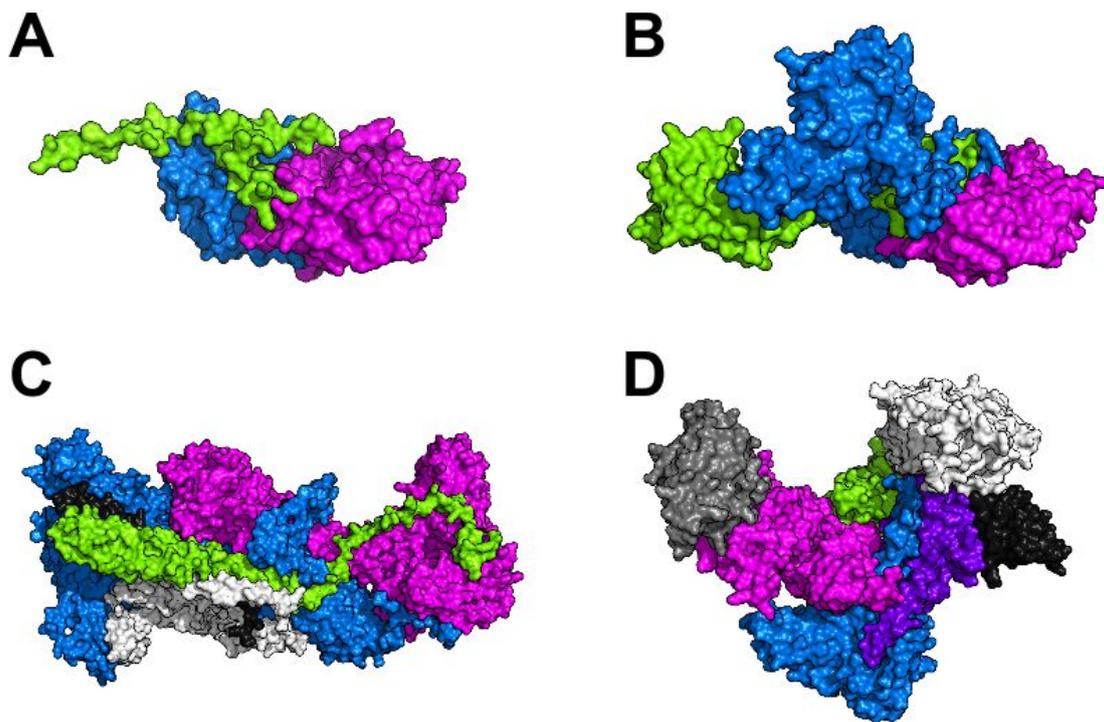
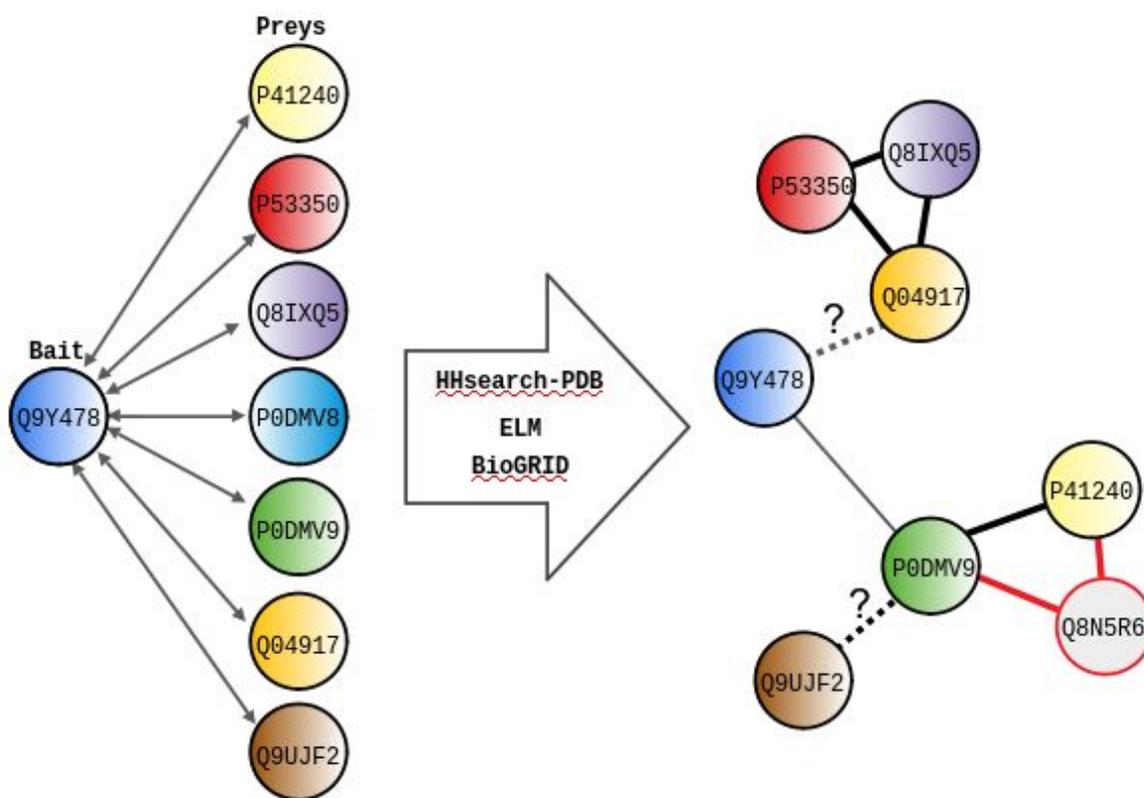


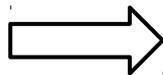
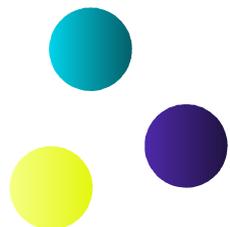
Figure 7

## Table of Contents graphic (TOC)

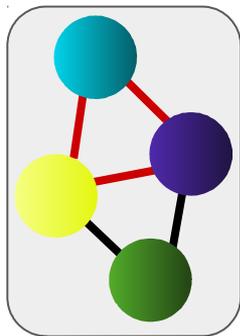


**Case 1**

Input:  
candidate partners



Experimental  
3D structure

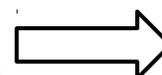
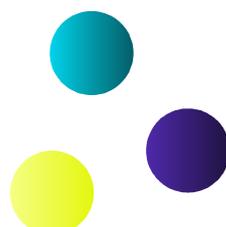


Output:

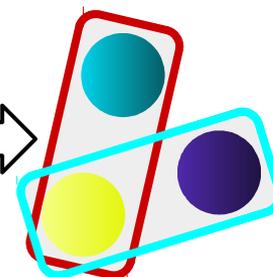
- **Interactions confirmed:** true positives (TP)
- **Undetected partners:** false negatives (FN)

**Case 2**

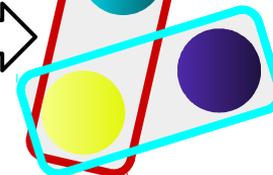
Input:  
candidate partners



**Structure 1**



**Structure 2**



Output:

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

Figure 1

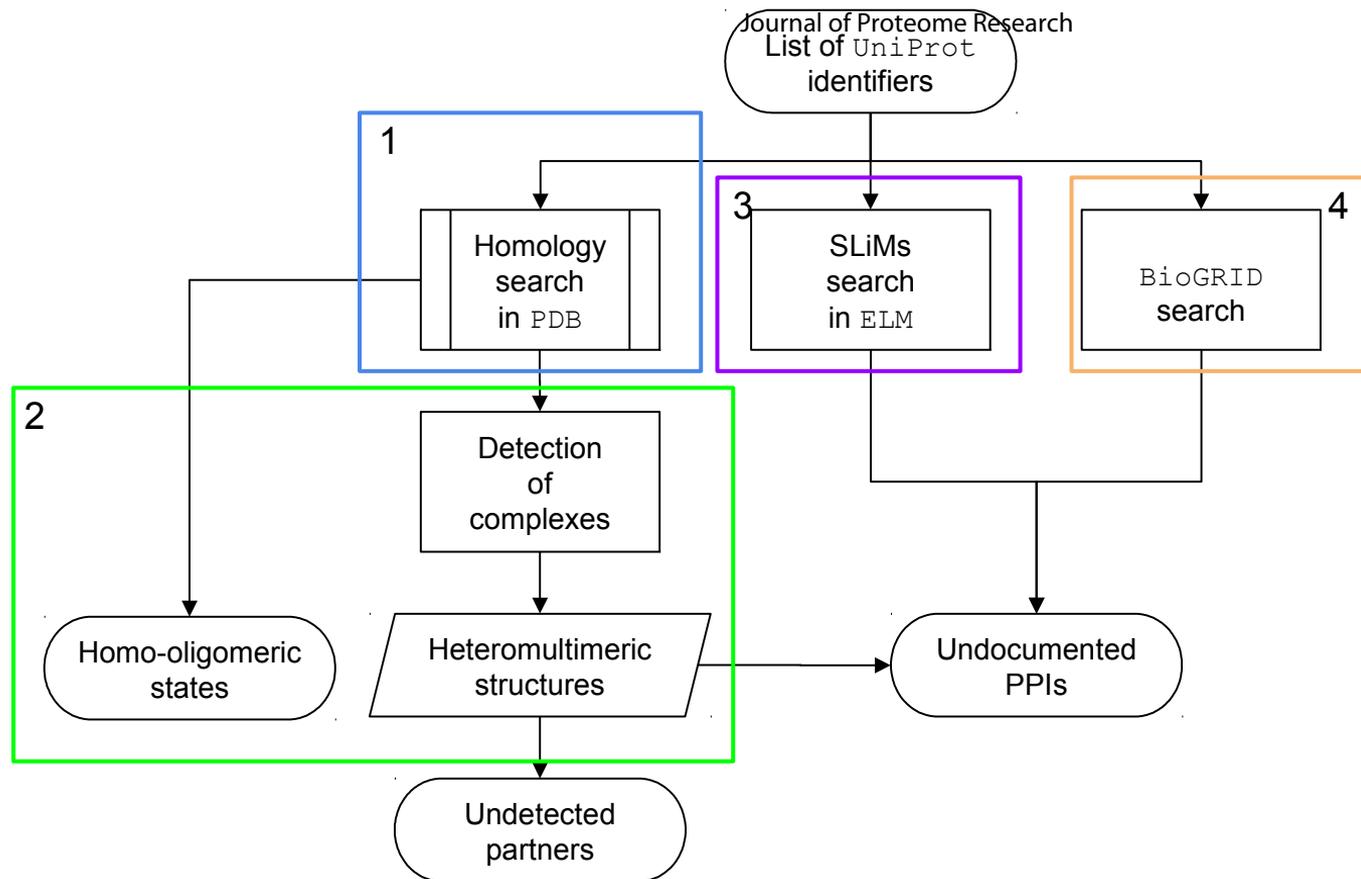


Figure 2

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30

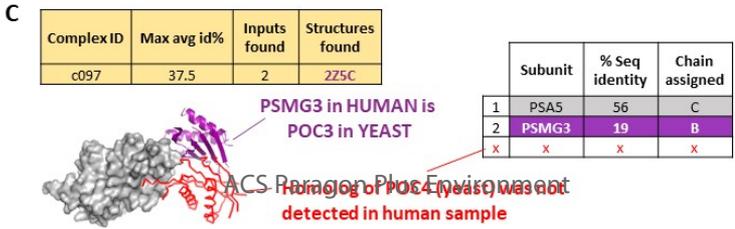
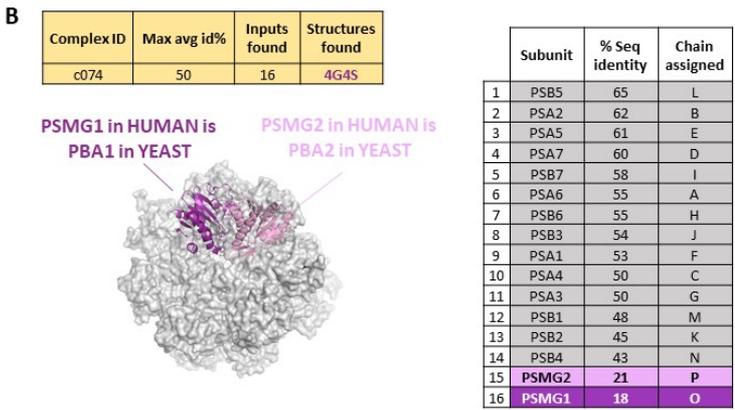
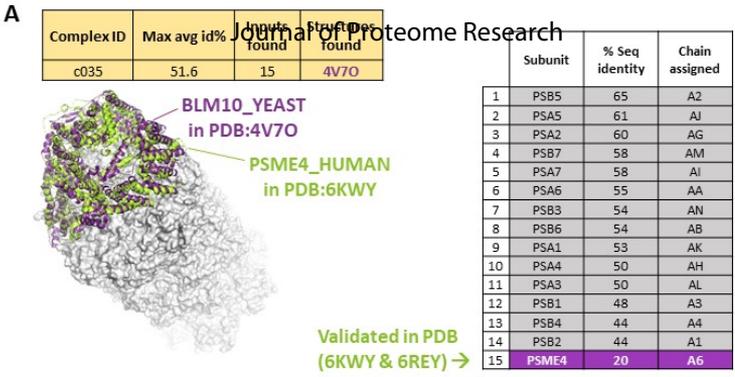


Figure 3

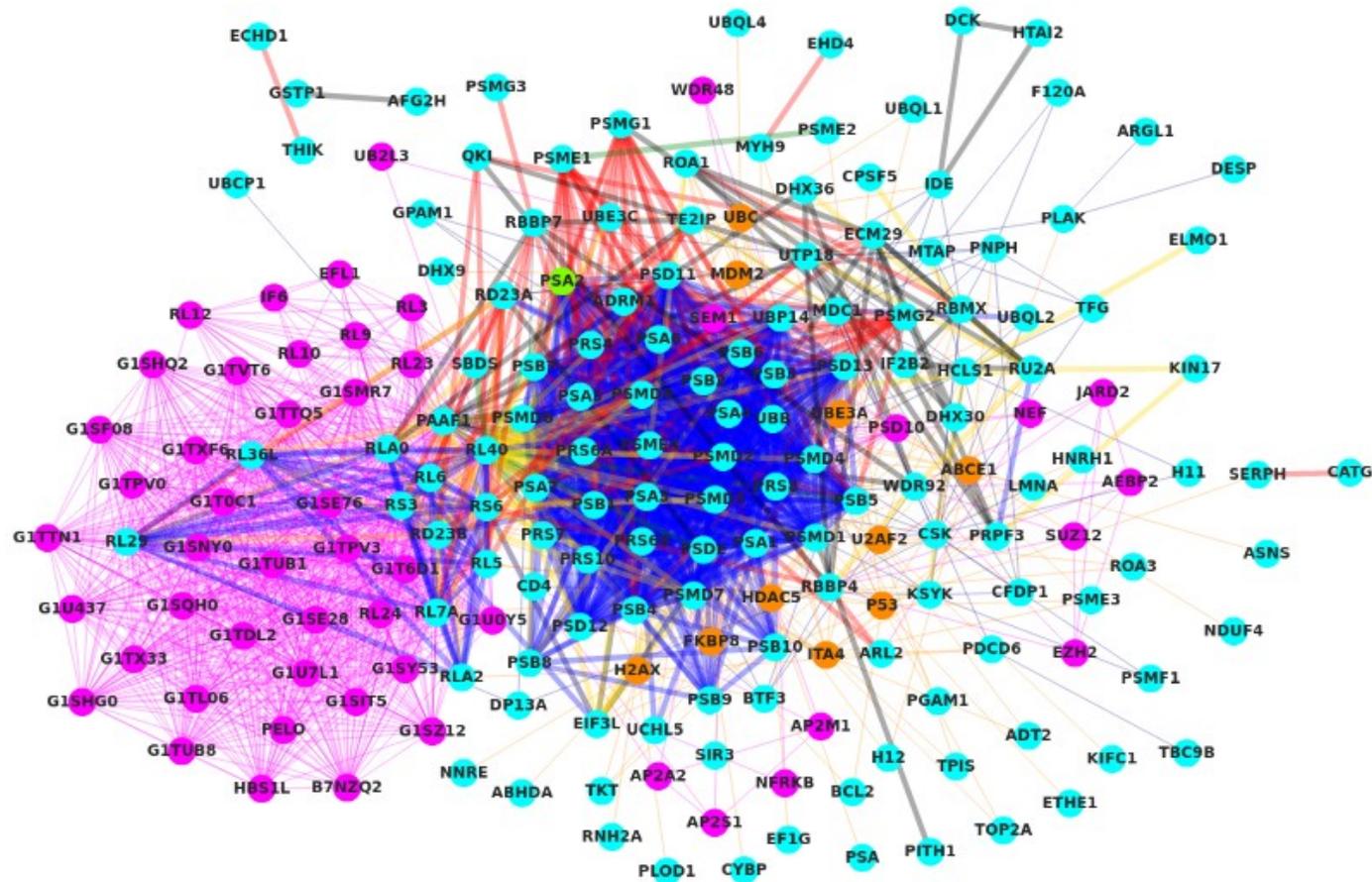


Figure 4

ACS Paragon Plus Environment

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30

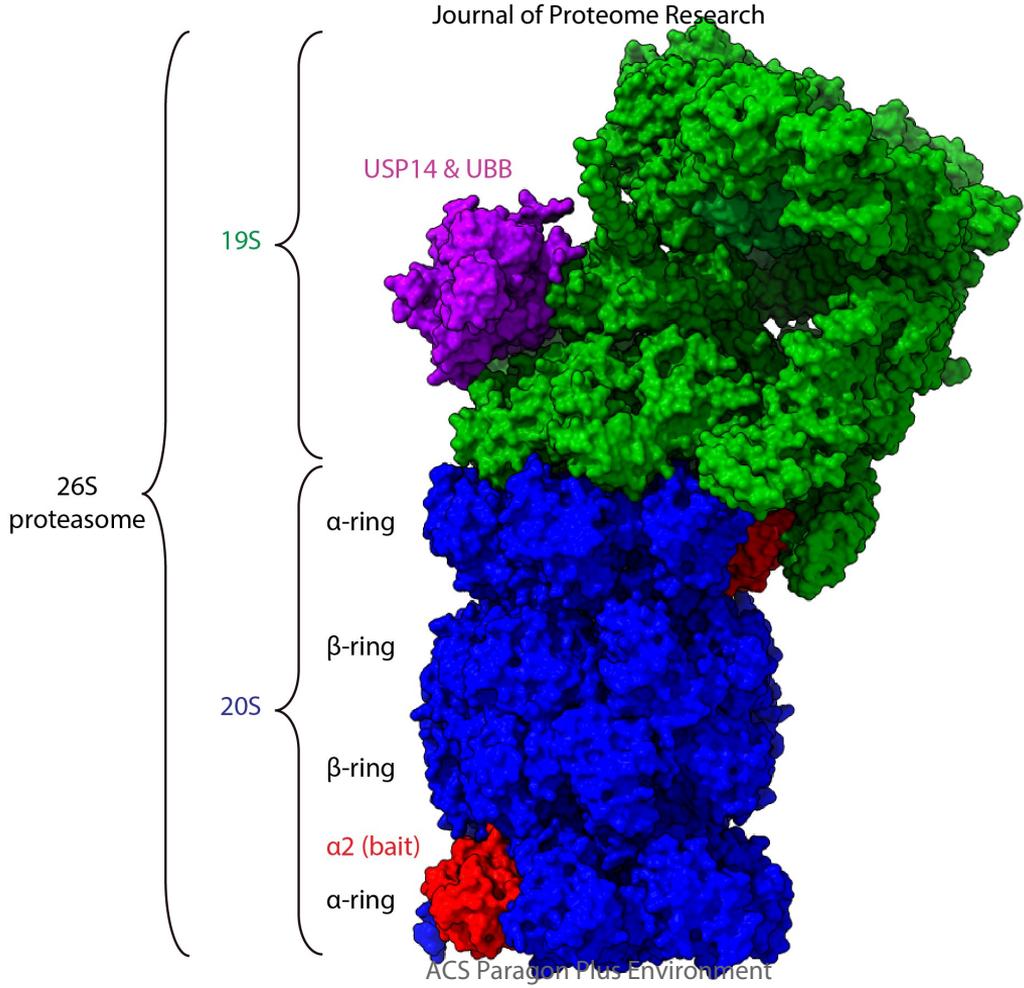


Figure 1

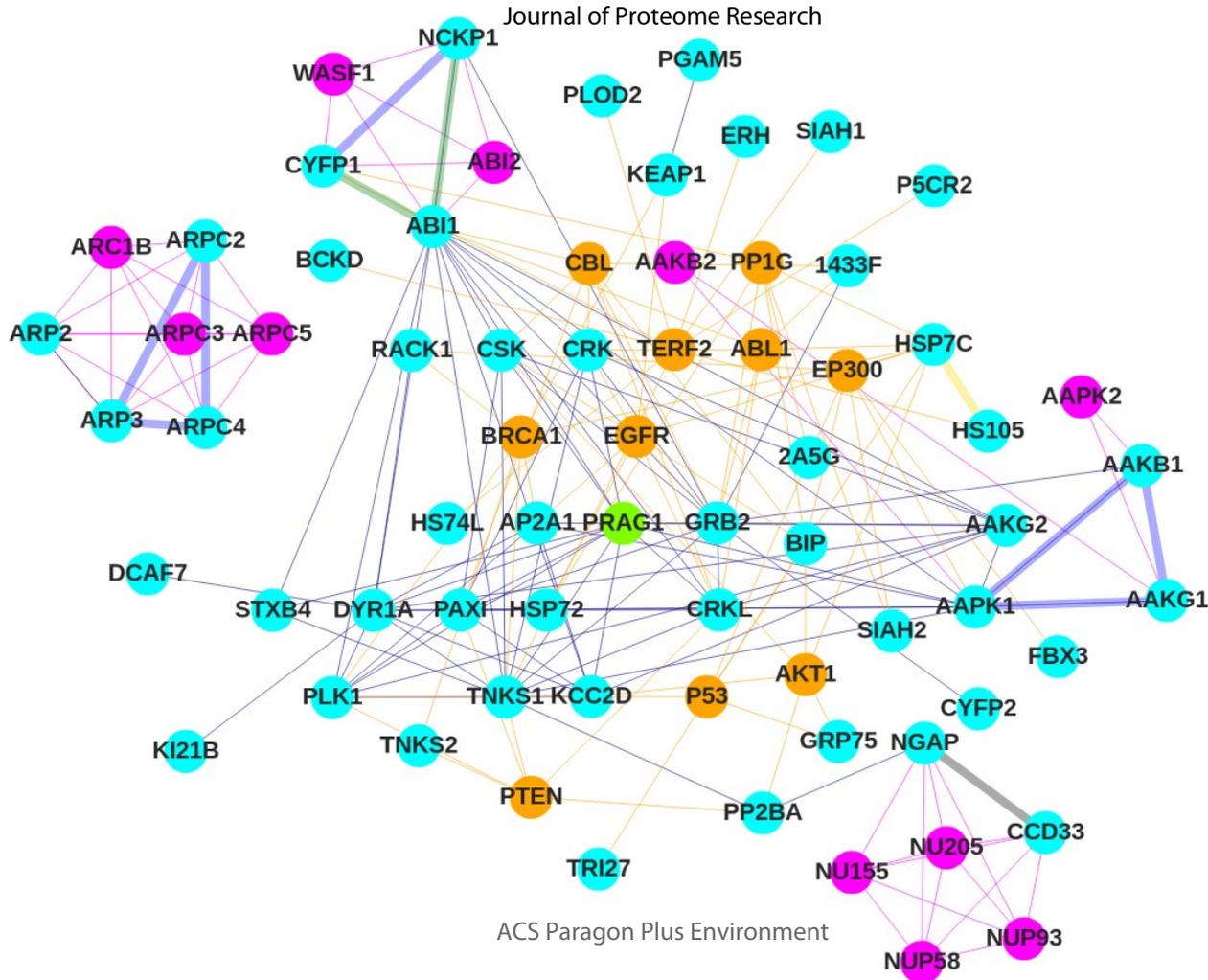
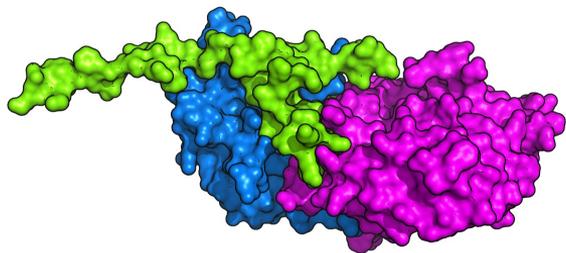


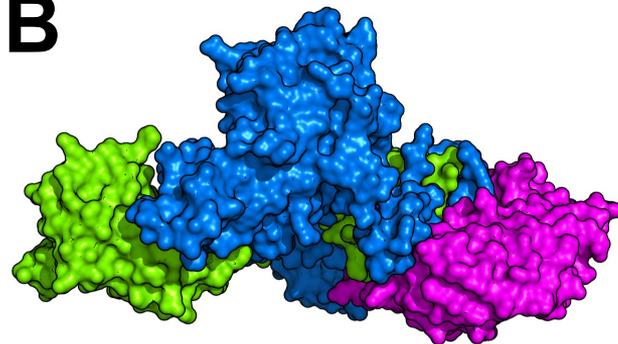
Figure 6

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30

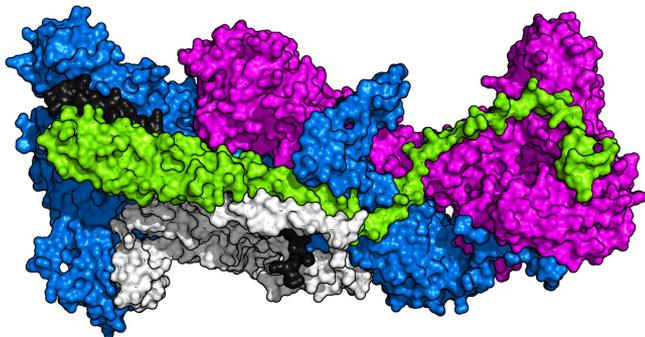
**A**



**B**



**C**



**D**

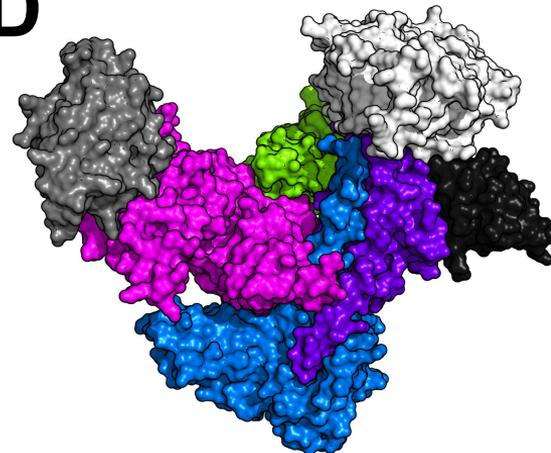


Figure 7

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30