

# Cross-ID: Analysis and Visualization of Complex XL–MS-Driven Protein Interaction Networks

Sebastiaan C. de Graaf,<sup>†,‡,§</sup> Oleg Klykov,<sup>†,‡,§</sup> Henk van den Toorn,<sup>†,‡</sup> and Richard A. Scheltema<sup>\*,†,‡</sup>

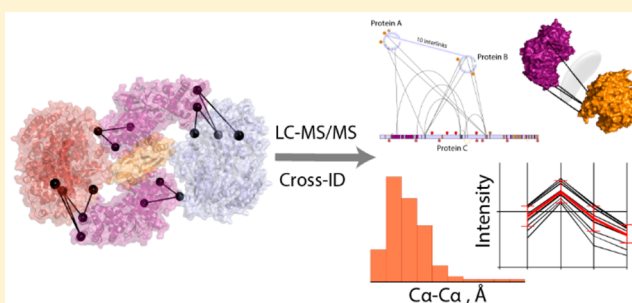
<sup>†</sup>Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research, and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

<sup>‡</sup>Netherlands Proteomics Centre, Padualaan 8, 3584 CH Utrecht, The Netherlands

## S Supporting Information

**ABSTRACT:** Protein interactions enable much more complex behavior than the sum of the individual protein parts would suggest and represents a level of biological complexity requiring full understanding when unravelling cellular processes. Cross-linking mass spectrometry has emerged as an attractive approach to study these interactions, and recent advances in mass spectrometry and data analysis software have enabled the identification of thousands of cross-links from a single experiment. The resulting data complexity is, however, difficult to understand and requires interactive software tools. Even though solutions are available, these represent an agglomerate of possibilities, and each features its own input format, often forcing manual conversion. Here we present Cross-ID, a visualization platform that links directly into the output of XlinkX for Proteome Discoverer but also plays well with other platforms by supporting a user-controllable text-file importer. The platform includes features like grouping, spectral viewer, gene ontology (GO) enrichment, post-translational modification (PTM) visualization, domains and secondary structure mapping, data set comparison, previsualization overlap check, and more. Validation of detected cross-links is available for proteins and complexes with known structure or for protein complexes through the DisVis online platform (<http://milou.science.uu.nl/cgi/services/DISVIS/disvis/>). Graphs are exportable in PDF format, and data sets can be exported in tab-separated text files for evaluation through other software.

**KEYWORDS:** cross-linking mass spectrometry, XL–MS, complex protein mixtures, proteome-wide cross-linking, XlinkX, protein–protein interactions, DisVis, XL-TMT



## INTRODUCTION

Protein interactions represent a level of cellular complexity that is essential for almost all biological processes. The protein assemblies they represent are highly dynamic and orchestrate cellular processes by regulating enzymes and forming macromolecular clusters capable of more complex behavior than the sum of their parts would suggest. Cross-linking mass spectrometry (XL–MS) has emerged as an attractive approach to elucidate protein–protein interactions (PPIs) by mass spectrometry. It uses small reagents with two reactive moieties capable of forging a covalent bond between two amino acids in close proximity. Upon application to proteins and protein–protein complexes followed by their proteolytic digestion, four distinct peptide products are formed: nonmodified, mono-linked, loop-linked, and cross-linked peptides.<sup>1</sup> The first three product groups consist of single peptides in various forms that yield limited or no structural information. The fourth group consists of two peptides captured by the cross-linking reagent; this yields valuable distance information for the elucidation of protein tertiary structure (the two peptides originate from the same protein) or protein quaternary structure (the two peptides originate from different proteins). The identification

of both peptides by mass spectrometry allows for the localization of the cross-link within the proteins of interest. Although several well-established methods like affinity purification mass spectrometry (AP–MS) are available for studying PPIs at high speeds,<sup>8</sup> most of these are limited to stable interactions or provide little to no structural information.<sup>2–7</sup> XL–MS, on the contrary, has the potential to capture weak and transient interactions complete with structural information. With recent advances in mass spectrometry, cross-linker chemistry, prefractionation techniques, and data analysis software, XL–MS can now routinely detect thousands of cross-linked peptides from a single experiment.<sup>9–12</sup> Even in the case of single proteins, XL–MS can yield hundreds of detected distance restraints.<sup>13</sup>

An attractive means to obtain a bird's-eye view of the cross-linking results is network graphs.<sup>14–16</sup> This type of visualization, however, also becomes cumbersome to read for

**Special Issue:** Software Tools and Resources 2019

**Received:** September 14, 2018

**Published:** December 21, 2018

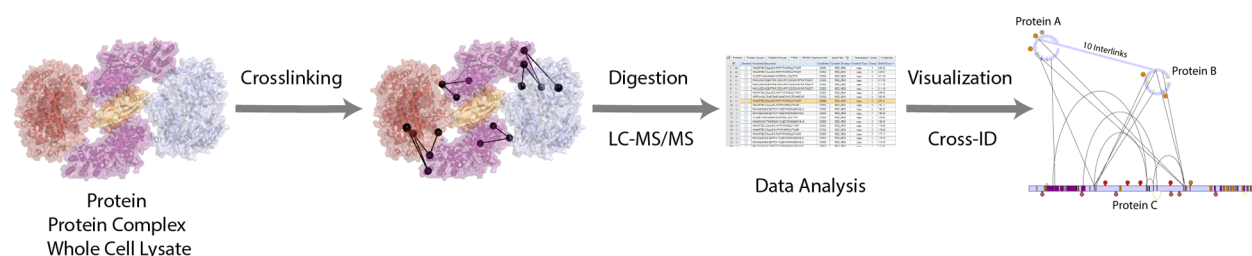


Figure 1. Visualization with Cross-ID.

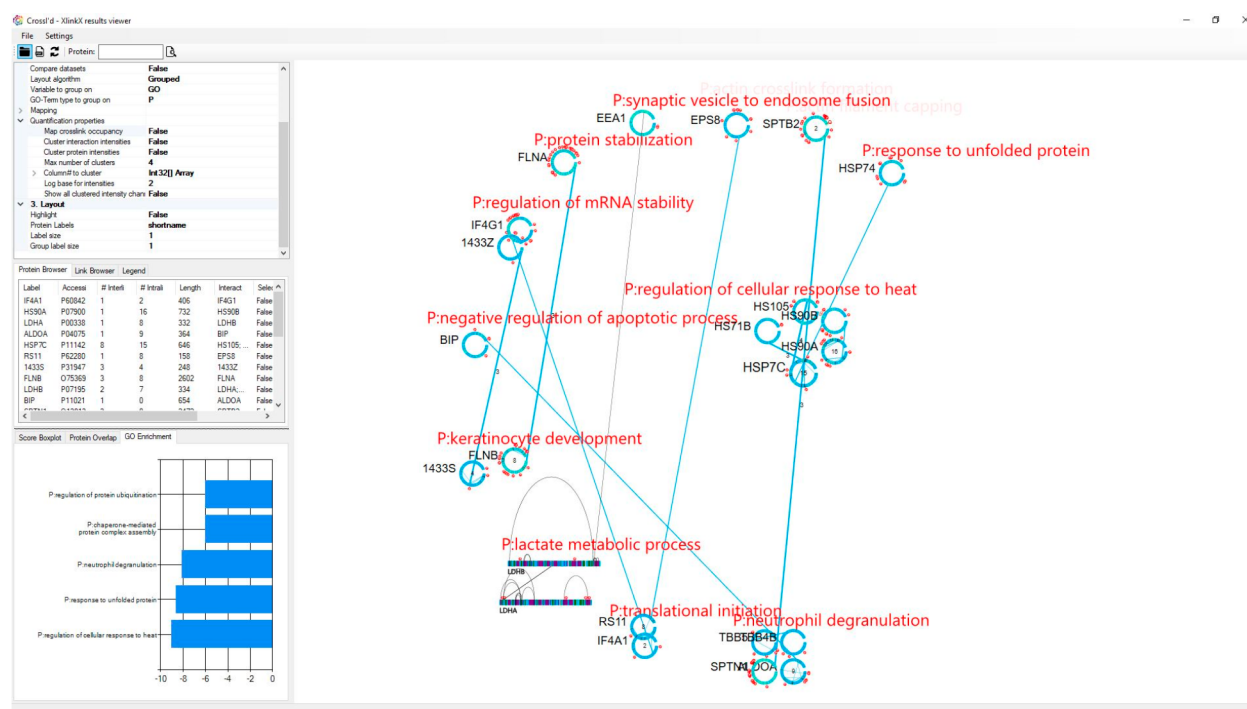


Figure 2. Screenshot of Cross-ID.

increasing sizes of the depicted data sets, where the large number of nodes and edges can easily obfuscate the view.<sup>17</sup> Additionally, when no connection between the visualized elements and the initial input data sets exists, it remains very difficult for the user to browse the data and check for validity. To circumvent these obstacles, advanced software allowing the network to be visualized, organized, and filtered in real time is needed. Several software platforms partly supporting such features exist,<sup>18–26</sup> with varying degrees of specificity toward XL–MS. Prominent examples include xiNet<sup>27</sup> and xVis.<sup>28</sup> Each of the tools has a unique set of features, and each offers a different subset of visualization options, which tailors them for particular applications (e.g., Xwalk<sup>20</sup> calculates solvent-accessible surface distances or XlinkAnalyzer<sup>21</sup> fits distance restraints to a given 3D model). However, when it comes to visualization of large-scale cross-linking data sets like whole-cell lysates, a combination of software solutions is often required. Proteome-wide interactomes can, for example, be visualized with biological network builders such as Cytoscape,<sup>14,15</sup> but there are no tools specifically tailored toward in-depth analysis of large proteome-wide XL–MS data sets. Added to this, relatively few tools support generic input formats from multiple software platforms (prominent examples that do include this are xiNET<sup>27</sup> and ProXL<sup>23</sup>); however, most are tightly linked to a specific search engine or define their own data format

requiring cumbersome file format changes to compare results between different data sets.

Here we present Cross-ID, a standalone solution for the visualization of XL–MS data as network graphs (Figures 1 and 2). It provides a direct connection to the output of XlinkX for Proteome Discoverer 2.3 but also supports an importer for comma-separated text output generated by any XL–MS search engine. In addition to cross-linking data, Cross-ID can display any data containing connection or distance restraints (e.g., small-angle X-ray scattering (SAXS) data<sup>29–31</sup>) as long as it is available in a tabular form. The importer uses natural language processing to predict the use of each column header in the output file and allows the user to make adjustments where required. The generated graphs are highly interactive and can be explored by filtering, expanding, repositioning, highlighting, mapping, or altering the graph directly. Ultimately, this will enable the user to draw meaningful conclusions from the graphs edited inside Cross-ID and without the need for editing the input data set each time before uploading. It is also possible to group proteins based on detected interlinks or according to other parameters (e.g., their GO enrichment coefficient), significantly simplifying the data analysis. A number of site-specific findings from the UniProt database<sup>32</sup> (among others glycosylation, disulfide bridges, and phosphorylation sites) can be mapped onto all protein representations as well as residues

of interest. In addition, it is also possible to depict specific detected modifications and quantitation values by various methods. Cross-ID also supports the validation of cross-links for a single protein or protein complex using available structures in Protein Data Bank (PDB) format. Alternatively, Cross-ID provides a direct link to DisVis for validating potentially interacting partners based on the detected cross-links.<sup>33,34</sup> As a showcase study, we provide a whole-cell lysate data set with 2754 cross-link spectra matches (CSMs), obtained from PC9 cells. To showcase the quantitation functionalities we used tandem mass tag (TMT) labelling to trace the behavior of protein kinase A (PKA), activated upon the addition of cAMP as a model system.

## MATERIALS AND METHODS

### DSSO Protein–Protein Cross-Linking

Cross-linked cell lysates have been prepared as previously described.<sup>35</sup> In brief, PC9 (Sigma-Aldrich, Steinheim, Germany) cells were collected and washed three times with PBS (Lonza, Basel, Switzerland). After centrifugation, the cell pellet was resuspended in cross-linking buffer consisting of 50 mM HEPES, 150 mM NaCl, and 1.5 mM MgCl<sub>2</sub> (all from Sigma-Aldrich). Protease inhibitors (Roche, Basel, Switzerland) and 0.5 mM DTT (Sigma-Aldrich) were added right before use. After the cells were lysed with a Bioruptor (Diagenode SA, Seraing, Belgium), freshly dissolved disuccinimidyl sulfoxide or DSSO in DMSO (Sigma-Aldrich) was added to a final concentration of 2 mM. The cross-linking reaction was quenched after 30 min with Tris-HCl at a final concentration of 20 mM. The cross-linked proteins were denatured and reduced and alkylated in a mixture of 8 M urea, TCEP, and CAA. Proteolytic digestion was performed in two steps: for 30 min with LysC (Wako, Tokyo) at r.t. and overnight with trypsin (Promega, Madison, WI) at 37 °C. Digested peptides were desalted with a Sep-Pak cartridge and dried prior to fractionation.

### Fractionation of Cross-Linked Peptides

Strong cation-exchange (SCX) chromatography was performed on an Agilent 1200 HPLC system (Agilent Technologies, Waldbronn, Germany). The setup was previously described<sup>36</sup> but, in short, consists of an Opti-Lynx trap column connected to a PolyLC SCX-separation column (PolyLC, Columbia, MD). Peptide mixtures were reconstituted in 5% DMSO/10% formic acid/85% water (v/v/v) and separated over a gradient of 120 min, resulting in 50 collected fractions. A total of 15 cross-link-rich fractions were chosen for analysis and prior to further analysis were dried and stored at –80 °C.

### LC–MS/MS Analysis

Peptide mixtures were reconstituted in 5% DMSO/10% formic acid/85% water (v/v/v) and analyzed on an Orbitrap Fusion Lumos (Thermo Fisher Scientific, San Jose, CA) coupled online to an Agilent 1290 UPLC (Agilent Technologies). Peptides were trapped on a double-frit C<sub>18</sub> precolumn (Reprosil C<sub>18</sub>, Dr. Maisch, 100 μm × 2 cm, 3 μm; packed in-house) for 5 min with buffer A (0.1% formic acid) and separated on a single-frit analytical column (Poroshell 120 EC C18, Agilent Technologies, 50 μm × 50 cm, 2.7 μm) over 155 min with a linear gradient from 10 to 40% B (B: 0.1% formic acid, 80% acetonitrile). Optimized MS settings were previously described.<sup>11,35</sup> Acquired data were analyzed with the Proteome Discoverer software suite 2.3 (Thermo Fisher Scientific) with

incorporated XlinkX nodes. Spectra were matched against the *Homo sapiens* database from SwissProt (version 2018\_06, 20 349 sequences, downloaded from UniProt). The protease was set to “Trypsin”, and the maximum number of missed cleavages was defined as 2. Carbamidomethylation of cysteines was set as a fixed modification and oxidation of methionine and protein N-terminal acetylation were set as variable modifications. For the linear peptide search, precursor mass tolerance was defined as 20 ppm, and fragment mass tolerance was defined as 0.5 Da for ion trap readout or 20 ppm for the Orbitrap readout. For the cross-linked peptides search, the minimum peptide length was set to 5 and minimum peptide mass was set to 300, whereas the maximum peptide mass was set to 7000. The precursor mass tolerance was set to 10 ppm, FTMS fragment mass tolerance was set at 20 ppm, and ITMS fragment mass was set at 0.5 Da. FDR threshold was set to 0.01 (1%), and FDR strategy was set as “Percolator”.

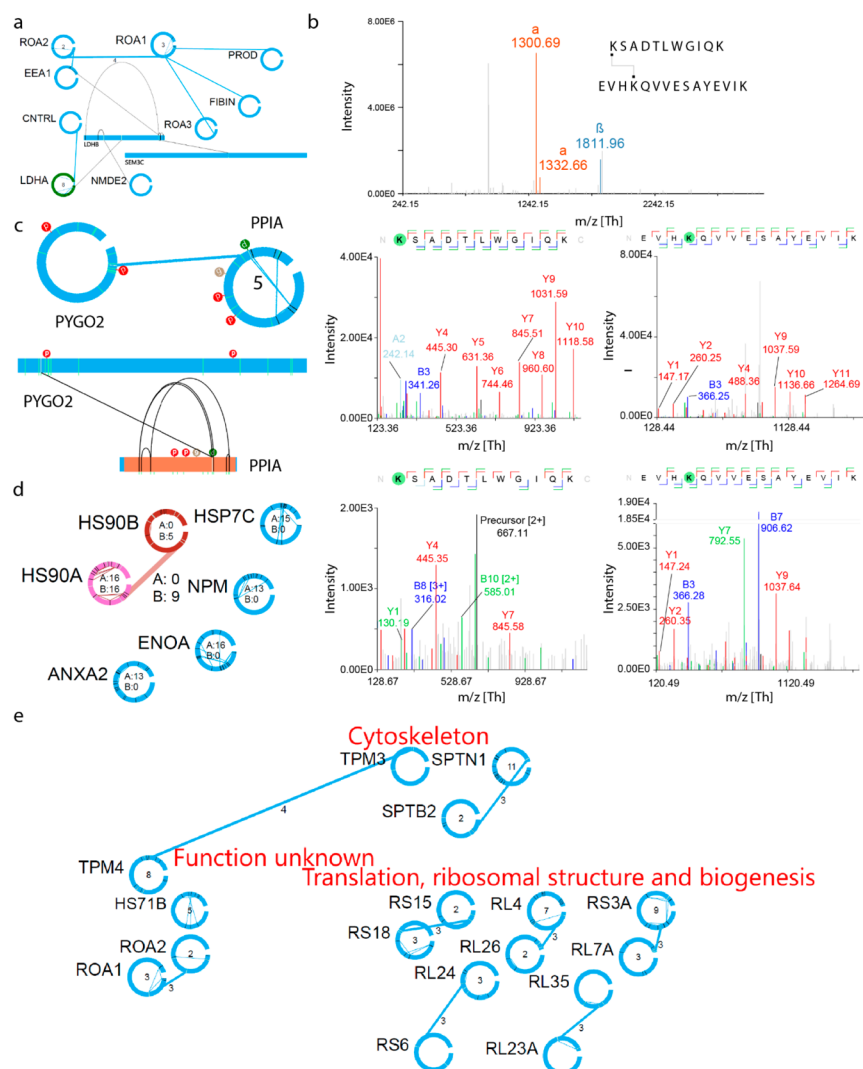
### TMT Experiments

TMT labels were purchased from Thermo Fisher Scientific, and the labeling protocol was performed according to supplier instructions after desalting of the cross-linked peptides. Ten channels were used to label 10 samples of model system PKA (Sigma-Aldrich, Steinheim, DE) solubilized at a concentration of 5.74 μM with added cAMP (Sigma-Aldrich) ligand to the final concentration of 0–8 and 10 μM respectively. Digestion, fractionation, and LC–MS/MS analysis were performed according to the procedure described above, except for alterations to the LC gradient consisting of increasing the starting point from 5 to 36% of buffer B. For these data, the Orbitrap Fusion (Thermo Fisher Scientific) with tune page version 3.1.2412.14 was used for data acquisition with the standard template for TMT-labeled cross-linking samples. For the data analysis, the TMT-specific nodes were added to the standard cross-linking data acquisition protocol<sup>35</sup> after the “Precursor Ion Exclusion” node; namely, “Isobaric Tag Loss” was set to TMT, “Precursor Selection Range” with mass range 400–1200 *m/z*, followed by 10 SPS scans with HCD at 65% NCE, and resolution of 50 000 in the Orbitrap. Recorded data were searched against PKA protein complex proteins with 200 Human proteins as decoys taken from the reviewed Swiss-Prot database. In addition to the standard XlinkX processing workflow, the “Reporter Ions Quantifier” node was added with “Integration Tolerance” set to 0.03 Da and “Centroid With Smallest Delta Mass” as “Integration Method”. For the consensus workflow, the “Reporter Ions Quantified” node was included with standard settings.

### Software and Data Availability

Cross-ID was developed in Microsoft Visual Studio 2017 as a C# WinForms application using Windows Presentation Foundation elements. The GraphX .NET library was used as the foundation for the network visualizations. For running the tool, a minimum of .NET version 4.7 needs to be installed. The software can be downloaded from <https://www.hecklab.com/software/xlinkx/> together with an instruction video. The raw data and all of the associated output and databases used in this study have been deposited to the Proteome-Xchange Consortium<sup>37</sup> via the PRIDE partner repository with the identifier PXD008418 (already published and openly accessible) for the whole-proteome data set and PXD011077 for the TMT data set.





**Figure 3.** Visualization with Cross-ID on a PC9 whole proteome data set. (a) Snapshot of the generated protein interaction network. (b) Spectral viewer for selected top-scored cross-links. (c) Comparison of bar and circle views for filtered proteins with depicted phosphorylation, glycosylation, and DSSO monolinks together with the known protein domain. (d) Comparison of cross-links filtered by XlinkX score at 50 with 13 intralinks and 7 interlinks. (e) Clustering according to the EggNOG database for filtered proteins.

## RESULTS AND DISCUSSION

### Data Import

Cross-ID provides a direct link to the output generated by the XlinkX nodes integrated in the Proteome Discoverer data analysis environment.<sup>35</sup> The files with extension “.pdResult” contain all information required to build the visualization of the network, including the spectra and protein information, together with the tables generated by XlinkX. by loading these files directly, correctness and access to all required information are ensured. To work with output from other search engines, Cross-ID provides a convenient import interface for tab- or comma-delimited text files, with column names on the first row. Because column names are not fixed between different search engines, or even in some cases between different versions of the same search engine, Cross-ID assists in manually selecting the correct columns. It provides a prediction of the purpose of each column by calculating a Levenshtein distance<sup>38</sup> to predefined column names.

The full UniProt database<sup>32</sup> is supported by Cross-ID and is used to provide additional information about the identified

proteins, like known post-translational modifications (PTMs) and secondary/tertiary structure information. It can, however, only do so when the cross-linked peptides contain valid UniProt accessions from the proteins they derive from (e.g., when the RAW data were analyzed against a protein FASTA file extracted from UniProt). For those cases where UniProt accessions are not available, Cross-ID automatically provides the opportunity to load the appropriate protein FASTA file for basic visualization and validation tasks described below.

### Basic Protein Visualization

To show the basic functionality of Cross-ID, we provide a whole-cell lysate data set with 2754 CSMS obtained from PC9 cells (Supplementary Table 1). Individual proteins are visualized as either a horizontal bar or a circular view, with the addition of their short or full protein name or the UniProt accession number in the form of an editable label. For a “clean” view, these labels can be removed completely or resized. At any time the visualization style can be altered from circular to bar or vice versa by mouse right-click for each individual protein (Figure 3a,c). Both the circular and horizontal bar protein visualizations represent the amino acid sequence in a clockwise

fashion or from left to right, respectively. In the horizontal bar, the width of the bar represents the length of the amino acid sequence, helping to get insight into the relative sizes of the different proteins and the exact positions of the detected cross-links. To provide initial insight into the potential of PTM-driven interactions, both representations can be annotated, with PTMs visualized as spherical tags containing the first letter describing the modification, both from UniProt and detected by the search engine. Uniquely for the circular view, gray lines on the circle depict residues involved in interlinks. Interlinks are connected by a single line between circles at the positions from the cross-link with the highest score. The number of cross-links between two proteins is shown above the connecting line, something that is also reflected by the thickness of the line. Black lines on the circle depict residues involved in intralinks, which are also connected by a line inside the circle. A number inside the circle depicts the number of unique intralinks. To assist in locating proteins with a high degree of interconnectivity, the size of both the circular display and the label are scaled according to the amount of interlinks detected for that protein. In addition, switching from circular to the horizontal bar view provides insight into both domain and secondary structure information extracted from UniProt.

Using a search bar, individual proteins and cross-links can easily be located within the graphs by full name, abbreviated name, or accession. All proteins involved in cross-links are displayed in the protein browser tab, and detected cross-links are displayed in the link browser. Here the user can center the graph on selected proteins/cross-links and sort and filter based on the source data set, number of inter- or intralinks, associated GO term (when grouped by GO term), and whether the protein has been selected. Browsers can be sorted on a column by clicking the column name, while clicking the right mouse button on column names opens a filtering menu. For example, in the link browser, interactions can be filtered and sorted through both protein names ("source" and "target"), the number of cross-links representing the interaction, the maximum score, the data set origin, and the cross-link type by mouse clicking on these column names. The protein browser can be filtered and sorted in a similar manner. To assess the data underlying the visualization, both the proteins and connecting lines can be clicked to access a list of all associated cross-links and their properties. Selecting an individual cross-link in this list provides another list of associated CSMs, and selecting a CSM shows the associated spectra in an integrated spectrum viewer together with information about the linked peptides and all detected modifications (Figure 3b). This option will, however, only work when the path to the folder containing the (Thermo) raw file has been correctly specified.

### Graph Visualization Options

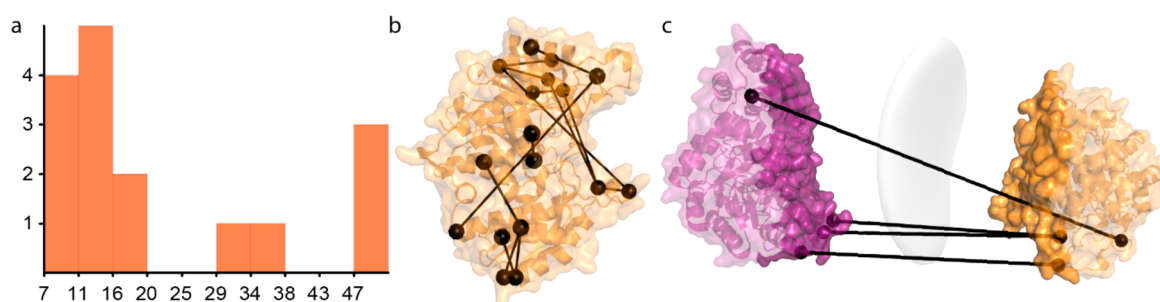
The network graph can be laid out in three different fashions: circular layout, Lin Log layout,<sup>39</sup> or via one of the various grouping options. The Lin Log algorithm positions the largest groups of interconnected nodes in the center of the graph and places groups of interconnected nodes increasingly further from the center the smaller they are, thereby minimizing the "energy" of the graph.<sup>40</sup> The grouping algorithm can group on GO terms, on source data set, by protein function according to the EggNOG database,<sup>41</sup> by creating hubs of equally interconnected proteins, or by user-defined groups (Figure 3e). In all cases, the largest group of proteins is placed at the

center as a circle of nodes, and the rest of the groups are smaller circles around it. The GO-term grouping is determined by comparing the frequency of the associated terms to either occurrence in a reference data set (provided in the form of a list of accession numbers) or in the whole genome of the organism under investigation by performing a Fisher's Exact test.<sup>42</sup> The term with the lowest resulting *p* value for each type of GO term chosen by user ("P" for Pathway, "F" for Function, or "C" for Compartment) is assigned as the term of interest for a given protein, and grouping can be done based on a term of interest for any of these three types. When clustered by connectivity, proteins are localized according to the number of interacting partners providing interaction hubs.

To make the graph more clear, cross-links or proteins can be hidden through several mechanisms. For example, the display of inter- or intralinks can be turned off, and a minimum score can hide potentially lower quality cross-link identifications. Alternatively, displayed proteins can be filtered based on the minimum number of inter- or intralinks. Within the link or protein browser, more intricate filters can be assembled as well (right-click the column of interest, select filtering, and implement the desired filter). The filtered data sets, as displayed in the protein browser or link browser, can consequently be exported as a .CSV file and, if required, loaded again into Cross-ID, enabling the creation of more compact graphs. Cross-ID also implements functionality to easily compare two data sets (e.g., controls vs experimental groups). When comparing multiple data sets, proteins and interactions are colored based on which data set they occur in: data set 1, dark red; data set 2, blue; or both, pink (Figure 3d). Additionally, the number of cross-links from each data set is provided, and a filter-responsive Venn-diagram is included as well, indicating overlap for shown proteins. To support replicates, there is an overlap-check function for multiple data sets that requires as additional input a minimum number of data sets in which a cross-link must occur before it is included in the final data set. Additionally, the fraction of cross-links included in the final data set is shown as well as a Venn diagram if more than one input file was provided. As before, the filtered data set can be exported to .CSV or directly used as a data set for further processing.

### Mapping Detected Cross-Links to Existing Structures

An often time-consuming task when analyzing cross-linking data is mapping the detected cross-links on existing structures. A major hurdle here is that the sequences in structures in PDB format tend to not precisely match those in standard databases like UniProt, usually caused by truncations, point mutations, or the exclusive availability of a structure from another organism. Such differences require a lot of time-consuming and error-prone manual work to locate the correct position for each cross-link. In particular, for large structures like the ribosome, this task quickly becomes infeasible. Automation is therefore desirable, and a number of separate solutions are available. One of the notable examples is Xlink Analyzer,<sup>21</sup> a Chimera<sup>43</sup> module that requires only structure and distance restraints as inputs for mapping. Similar input information is required for the R package XLmap,<sup>24</sup> which also generates overlaid plots of cross-linked sites on contact maps and assigns a score to each model. Cross-ID incorporates extensive automation for this cumbersome task. It aligns the sequences used for analysis and those encapsulated within the crystal structure file using the Smith–Waterman local sequence alignment together with the



**Figure 4.** Validation of cross-links detected for alpha-enolase. (a) Distance distribution of mapped cross-links on alpha-enolase. (b) Detected cross-links on crystal structure. (c) Interaction interface generated by DisVis based on indicated restraints (grey surface) in comparison to the existing dimeric interface (dark purple and dark orange).

**Table 1. List of Crosslinks Detected for Alpha-Enolase**

accession	name	score	position A	sequence A	position B	sequence B	distance
P06733	alpha-enolase	124.34	4	SILKIHAR	71	AVEHINKTIAPALVSK	12.0
P06733	alpha-enolase	109.28	54	DNDKTR	233	TAIGKAGYTDK	48.2 (14.0 on dimer)
P06733	alpha-enolase	106.76	80	TIAPALVSKK	89	LNVTEQEKIDK	10.7
P06733	alpha-enolase	102.98	80	TIAPALVSKK	92	IDKLMIEDGTENK	12.8
P06733	alpha-enolase	97.87	197	NVIKEK	233	TAIGKAGYTDK	13.8
P06733	alpha-enolase	90.32	54	DNDKTR	60	YMGKGVSK	7.9
P06733	alpha-enolase	89.08	4	SILKIHAR	81	KLNVTQEK	10.3
P06733	alpha-enolase	84.61	60	YMGKGVSK	202	YGKDATNVGDEGGFAPNILENK	33.0 (21.2 on dimer)
P06733	alpha-enolase	83.55	193	IGAEVYHNLKNVIK	202	YGKDATNVGDEGGFAPNILENK	9.2
P06733	alpha-enolase	81.34	330	LAKAVNEK	326	FTASAGIQVVGDDLTVNPKR	6.5
P06733	alpha-enolase	77.94	81	KLNVTQEK	92	IDKLMIEDGTENK	14.7
P06733	alpha-enolase	77.91	197	NVIKEK	60	YMGKGVSK	35.5 (20.0 on dimer)
P06733	alpha-enolase	65.9	256	SGKYDLDFK	126	AGAVEKGVPLYR	51.8 (64.2 on dimer)
P06733	alpha-enolase	65.06	92	IDKLMIEDGTENK	358	VNQGSVTESLQACKLAQANGWGMVSHR	17.6
P06733	alpha-enolase	43.63	256	SGKYDLDFK	281	YISPDQLADLYKSFIF	17.7
P06733	alpha-enolase	41.96	199	EKYGK	233	TAIGKAGYTDK	14.0

BLOSUM62 substitution matrix.<sup>44</sup> Non-natural amino acids such as pyrrolysine and selenocysteine are automatically substituted with standard lysine and cysteine, respectively, prior to alignment. The minimum sequence similarity for this step can be defined by the user but is set by default to 60%, which works well in most cases. This initial alignment step is used to determine which proteins' structures are represented in the provided PDB structure and select these proteins and their cross-links as candidates for validation. Next, the alignment of the selected cross-linked peptide by the same protocol is performed. Again, a minimum sequence similarity can be defined, but the default set to 88%. To guide the process, the residues involved in the cross-link can be defined, set by default to lysine. In case another residue is matched after alignment, the software automatically verifies whether this residue is characterized by similar chemistry (e.g., arginine instead of lysine). In those cases where this is not so, the cross-link is flagged and the user can decide on a case-to-case basis how to proceed. Afterward, the cross-link positions are mapped to the structure, and Euclidian distances between the  $C_{\alpha}$  atoms of linked residues are calculated and presented in a filter-responsive list. This list also contains the last eight characters of the PDB filename and the detected distances as well as amino acid sequences of cross-linked peptides with the highest sequence overlap. Upon completion, the user is presented with a dialogue summarizing the validation by detailing the amount of unvalidated intra- and interlinks, substituted residues, and flagged residues. The distribution of the found distances is automatically shown in a histogram (Figure 4a).

Another major hurdle is the preparation of the existing structures and cross-linking data for automated docking procedures. Cross-ID also provides far-reaching automation for these purposes by integrating with the DisVis/HADDOCK computational structural docking environment.<sup>33,34,45,46</sup> For this purpose, Cross-ID currently provides automated access to DisVis,<sup>33,34</sup> although we intend to add more options in future releases. Given 2 structures, DisVis is able to predict prospective interaction interfaces based on user-supplied distance restraints and has already been applied successfully to XL-MS data sets.<sup>35</sup> Restraints that are violated in the predicted interface will be marked as false-positives and can be omitted prior to further modeling steps. A score indicating the probability of occurrence of each of the restraints between the submitted structures is calculated as well. All required files are automatically prepared and uploaded based on the results from the sequence alignment step previously described. Prior to upload, the minimum and maximum restraint lengths can be changed manually. As a model for validation, we used alpha-enolase, which has a known PDB structure (PDB ID: 2PSN; resolution 2.2 Å). For this protein, XlinkX detected a total of 28 cross-links (Supplementary Table 2), of which 16 are on enolase alone (Table 1). Of these, 12 restraints are within the DSSO cross-linking distance of 30 Å, whereas 4 exceed this (Figure 4b). Enolase, however, exists in solution as a dimer, meaning that the violated restraints are potentially cross-links between the two subunits. To verify this, we submitted chains A and B from a known PDB structure, with only the outliers to DisVis (Supplementary Table 3). Three out of four restraints

Table 2. Recommended Input Format for Cross-ID<sup>a</sup>

column name	value type	example value	priority	function
cross-linker	string	DSSO	recommended	spectral viewer
strategy	string	MS2_MS2_MS3	recommended	spectral viewer
charge	integer	4	optional	
score	double	100.01	recommended	filter
M/Z (Da)	double	365.95	optional	
retention time	double	131.21	optional	
modifications A–B	strings	K7(DSSO); C8(Oxidation)	recommended	mapping
sequence A– <sup>b</sup>	string	MKIVDVIGEK	recommended	validation
cross-link position A–B <sup>b</sup>	integer	6	recommended	validation
identifier A–B <sup>b</sup>	string	P02671	essential	visualization
leading protein position A–B	integer	166	recommended	mapping/visualization
all scans	integers	494; 495; 496; 497; 498; 499	recommended	spectral viewer
M/Z A–B <sup>b</sup>	doubles	830.48938; 846.475586	recommended	spectral viewer
charges A–B <sup>b</sup>	integers	2; 2	recommended	spectral viewer
first scan	integer	49349	recommended	spectral viewer
spectrum file	string	rawFile.raw	recommended	spectral viewer
cross-link occupancy	double	0.66	recommended	mapping
intensity 1–20 <sup>c</sup>	double	3709.23	recommended	intensity clustering

<sup>a</sup>Optional columns were kept for the purpose of providing the user with additional information for inspection in tabular format and for exporting.

<sup>b</sup>Columns with a name indicating a range (e.g., charges A–B) indicate multiple columns with the same requirements. Columns needing an array of values (e.g., “all scans”) require those values to be separated by semicolons. <sup>c</sup>Intensity columns are imported through a separate mechanism to avoid cluttering of the importer form. This means that the intensity columns should be named: intensity1, intensity2, and so on. Other column names do not require specific formatting as long as the importer is used.

were detected as valid and indeed could be mapped on a dimer structure with distances of 14.0, 20.0, and 21.2 Å (Figure 4c). The remaining restraint has been detected by DisVis as a false-positive and can be mapped on a dimer with a distance of 64.2 Å (Supplementary Table 4).

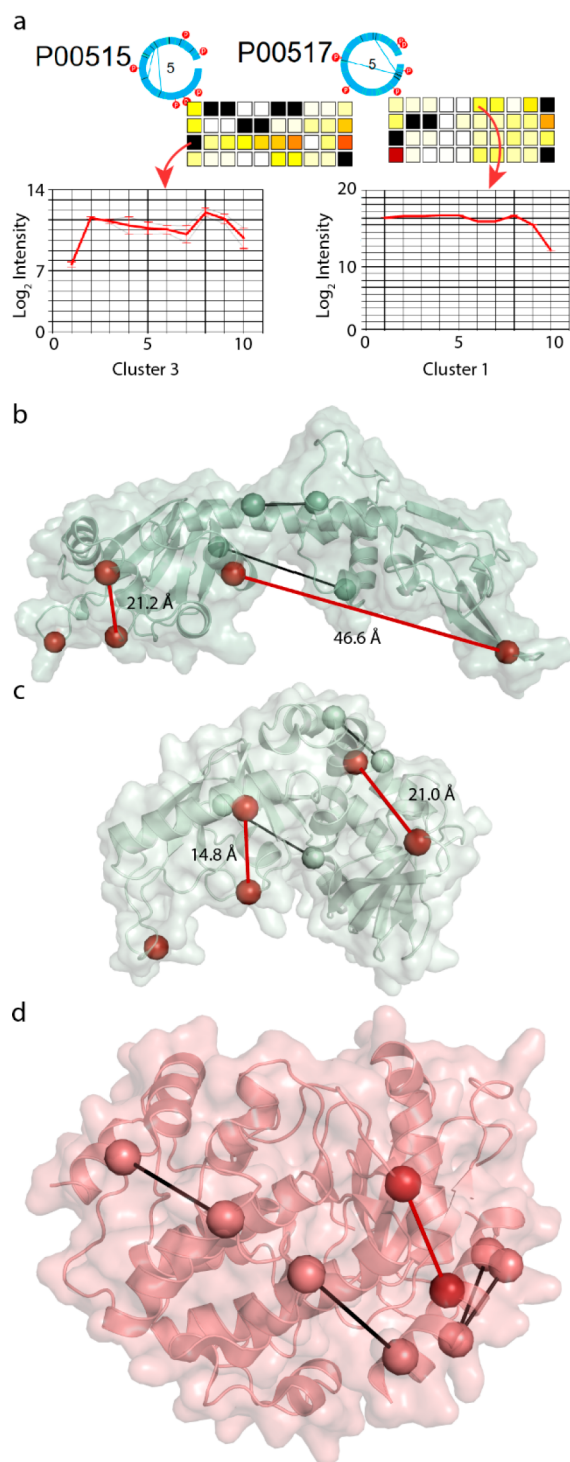
### Quantitation

Quantitation of cross-links is rapidly becoming an important facet for cross-linking analyses, providing insight into the structural rearrangement of proteins upon stimulation. To support quantitation coming from cross-link analysis, Cross-ID offers two quantitation parameters in the generated graphs: intensities and cross-link occupancy (representing how often a pair of residues was actually cross-linked as opposed to not modified or monolinked). The latter value is mapped as a heat-colored circle on the cross-link line (black for 0, white for 1, and a scale from red to yellow in between). If intensities are provided, the column names in the input file have to be edited accordingly (Table 2). Measured CSM intensities (from label-free or labeling experiments like TMT) are clustered using the *k*-means algorithm.<sup>47</sup> When several identified spectra for the same cross-link positions are quantified, the median value is taken for further analysis. The number of clusters is set by default to 4. The clustered intensities are subsequently visualized in table format within the graph, using heat-colored squares, the colors of which are determined by their log-transformed intensity relative to the rest of the cluster. The number of columns of this table can be set to match the number of clustered intensity channels. The intensity values are automatically log-transformed before clustering, and the base of this log transformation can be set by the user. Within the table representation, a column represents the experiment (e.g., in the case of TMT labeling, the first column represents channel 1, etc.). The clustered values can be accessed for a more detailed overview by pressing the “C” key while clicking on either the edge (for interaction clustering) or the protein (for protein intensity clustering). The cross-links sorted by

cluster are returned, as well as a line graph for the selected cluster showing the median intensities as a thick red line with error bars and all of the individual intensities as faded-out gray thin lines.

To demonstrate the ability of Cross-ID to rapidly leverage quantitation information, we performed TMT labeling experiments on the bovine PKA complex. PKA is a tetramer composed of two regulatory subunits and two catalytic subunits. Each regulatory subunit is able to bind two molecules of cyclic AMP (cAMP); upon binding, the catalytic subunits are released. We used TMT 10-plex to measure the structural behavior for increasing concentrations of cAMP. There are two types of regulatory subunits; for each type, the alpha and beta forms are present, and a complex can be formed either by combination of the alpha and beta forms or by one a single form. Catalytic subunits are also present as alpha and beta forms, but only one of the forms is present in the PKA complex. Because the structure for the bovine type II alpha regulatory subunit is missing, we modeled this protein from residues 97 to 402 with I-TASSER<sup>48</sup> (Supplementary Structure 1) using the available template from mouse (PDB ID 3TNP with resolution 2.3 Å, chain B). As structure with the bound cAMP ligand, we used a previously modeled structure from the SWISS-MODEL repository<sup>49</sup> (Supplementary Structure 2). We detect five intralinks for the type II regulatory subunit alpha-form (UniProt accession P00515, Figure 5a) and no cross-links for the beta-form. The catalytic subunit is represented by the alpha subunit with five intralinks (UniProt accession P00517, Figure 5a). In both cases, it was possible to group the behavior of all detected cross-links into four clusters even though the number of maximum clusters was set to five (see Supplementary Table 5 and Supplementary Figure 1A–D). It is known that the regulatory subunit undergoes conformational changes upon binding of cAMP. Clusters 3 and 4 contain cross-links with increasing intensities for increasing concentrations of cAMP. Cross-link 187–269 is mapped as





**Figure 5.** Quantitation with Cross-ID. (a) Detected cross-links for the PKA complex in circular representation (left: regulatory subunit; right: catalytic subunit alpha). (b) Ligand-free and (c) cAMP-bound structures of bovine alpha type II regulatory subunit. (d) Structure of bovine catalytic subunit alpha with mapped cross-links. Cross-links mapped in black do not change intensity across TMT channels, whereas residues and cross-links mapped in dark red are changing their intensity across the TMT channels.

46.6 Å when no ligand is present (see Figure 5b) and 21.0 Å with the ligand present (see Figure 5c); for this cross-link, we detect a 10-fold increase in intensity. Cross-link 342–376 is mapped as 21.2 Å on the holoenzyme regulatory subunit and

quantified with relatively low intensity in the control experiment. On the folded conformation, the same restraint is 30% shorter and shows an intensity increase upon cAMP addition of almost four-fold. There is one unmapped cross-link between lysine residues 315, which is located on the surface-exposed flexible loop and might belong to an alternative folded conformation of the complex. Notably, the remaining cross-links are mapped within the DSSO cross-linking range for at least one of the conformations of the regulatory subunit (Supplementary Table 6). The catalytic subunit is released upon cAMP binding and with this release creates a highly dynamic protein with domains involved in hinge and shear motions.<sup>50</sup> Even though it is expected that this protein is very flexible, all detected intralinks can be mapped on the available apoenzyme structure (PDB ID 5VI9 with resolution 1.9 Å, chain A) within the DSSO maximum cross-linking distance (Supplementary Table 6). Cross-link 24–193 is located in cluster 1 (Supplementary Figure 1E) and shows a drastic decrease in intensity for higher concentrations of cAMP. This behavior is not readily explainable, but we hypothesize that upon substrate binding the protein becomes structurally less flexible through the formation of a salt-bridge between one of these lysines and Asp'162 (see Figure 5d).

## CONCLUSIONS

Cross-linking mass-spectrometry experiments tend to produce such large amounts of data that processing rapidly becomes impractical, especially in the case of whole proteome experiments. To alleviate this, we present Cross-ID, a tool that produces graph presentations of cross-linking data and offers several tools to bring the detected cross-linking data into structural data like crystal structures. It offers optimal integration with the XlinkX data analysis pipeline<sup>11,55</sup> but also supports the import of data in CSV format from other search engines with partial automation through a natural language importer. Various forms of grouping of the protein network are supported for gaining optimal insights into the detected data, with support for grouping on external data such as, for example, GO annotations. To support analyses of the detected cross-links on existing crystal structure data, Cross-ID implements automated sequence alignment to bridge the differences between the used sequences and those in the crystal structures. As further support in this direction, it also offers a convenient interface to the structural modeling pipeline of DisVis/HADDOCK. With support for various quantitation options with automated clustering, the tool provides a very detailed look at structures from a cross-linking point-of-view. Cross-ID was developed with extensibility in mind and as part of the XlinkX data analysis pipeline will see continued development and support. Future functionalities currently already under development include integration of PPI databases such as String<sup>51</sup> and CORUM<sup>52</sup> to groups based on known complexes, further integration with the HADDOCK software for structural modeling, implementation of true distance measures like, for example, Xwalk implements,<sup>20</sup> integration with standardization efforts like mzIdentML, and many others. To further integrate with other software, we aim to add support for non CSV/text-based output formats like pepXML<sup>53</sup> or pepXMLTab.<sup>54</sup>



## ■ ASSOCIATED CONTENT

### ■ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.8b00725.

Supplementary Figure S1, DisVis in-/output files, I-TASSER and SWISS-MODEL models, and XlinkX output together with detected cross-link distances (ZIP)

## ■ AUTHOR INFORMATION

### Corresponding Author

\*Tel: +31 30 253 6793. E-mail: r.a.scheltema@uu.nl.

### ORCID

Henk van den Toorn: 0000-0002-0270-5763

Richard A. Scheltema: 0000-0002-1668-0253

### Author Contributions

<sup>§</sup>S.C.d.G. and O.K. contributed equally. R.A.S. conceived of the study. O.K. performed the XL-MS experiments and provided ideas for features. S.C.d.G., H.v.d.T., and R.A.S. programmed Cross-ID. S.C.G., O.K., and R.A.S. wrote the paper, whereafter all authors critically read and edited the manuscript.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

We thank all Heck group members for their helpful contributions and for enduring early testing. From the Bonvin lab, we thank Alexandre Bonvin and Jörg Schaarschmidt for integration with the DisVis platform. From Thermo Fisher Scientific, we thank Bernard Delanghe, Kai Fitzemeier, and Frank Berg for their collaboration on incorporating the XLinkX cross-link search engine into the Proteome Discoverer software and Rosa Viner for her collaborative work on DSSO cross-linking and support in mass-spectrometry method development. We acknowledge financial support by the large-scale proteomics facility Proteins@Work (Project 184.032.201) embedded in The Netherlands Proteomics Centre and supported by The Netherlands Organization for Scientific Research (NWO). Additional support came through the European Union Horizon 2020 program FET-OPEN project MSmed (Project 686547) and the European Union Horizon 2020 program INFRAIA project Epic-XS (Project 823839).

## ■ REFERENCES

- (1) Schilling, B.; Row, R. H.; Gibson, B. W.; Guo, X.; Young, M. M. MS2Assign, Automated Assignment and Nomenclature of Tandem Mass Spectra of Chemically Crosslinked Peptides. *J. Am. Soc. Mass Spectrom.* **2003**, *14* (8), 834–850.
- (2) Fagerlund, R. D.; Wilkinson, M. E.; Klykov, O.; Barendregt, A.; Pearce, F. G.; Kieper, S. N.; Maxwell, H. W. R.; Capolupo, A.; Heck, A. J. R.; Krause, K. L.; et al. Spacer Capture and Integration by a Type I-F Cas1–Cas2–3 CRISPR Adaptation Complex. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (26), 5122–5128.
- (3) Benda, C.; Ebert, J.; Scheltema, R. A.; Schiller, H. B.; Baumgärtner, M.; Bonneau, F.; Mann, M.; Conti, E. Structural Model of a CRISPR RNA-Silencing Complex Reveals the RNA-Target Cleavage Activity in Cmr4. *Mol. Cell* **2014**, *56* (1), 43–54.

- (4) Joachimiak, L. A.; Walzthoeni, T.; Liu, C. W.; Aebersold, R.; Frydman, J. The Structural Basis of Substrate Recognition by the Eukaryotic Chaperonin TRiC/CCT. *Cell* **2014**, *159* (5), 1042–1055.
- (5) Herzog, F.; Kahraman, A.; Boehringer, D.; Mak, R.; Bracher, A.; Walzthoeni, T.; Leitner, A.; Beck, M.; Hartl, F.-U.; Ban, N.; et al. Structural Probing of a Protein Phosphatase 2A Network by Chemical Cross-Linking and Mass Spectrometry. *Science (Washington, DC, U. S.)* **2012**, *337* (6100), 1348–1352.
- (6) Chen, Z. A.; Jawhari, A.; Fischer, L.; Buchen, C.; Tahir, S.; Kaminski, T.; Rasmussen, M.; Lariviere, L.; Bukowski-Wills, J.-C.; Nilges, M.; et al. Architecture of the RNA Polymerase II-TFIIF Complex Revealed by Cross-Linking and Mass Spectrometry. *EMBO J.* **2010**, *29* (4), 717–726.
- (7) Armony, G.; Jacob, E.; Moran, T.; Levin, Y.; Mehlman, T.; Levy, Y.; Fass, D. Cross-Linking Reveals Laminin Coiled-Coil Architecture. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (47), 13384–13389.
- (8) Hosp, F.; Scheltema, R. A.; Eberl, H. C.; Kulak, N. A.; Keilhauer, E. C.; Mayr, K.; Mann, M. A Double-Barrel Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS) System to Quantify 96 Interactomes per Day. *Mol. Cell. Proteomics* **2015**, *14* (7), 2030–2041.
- (9) Kao, A.; Chiu, C.; Vellucci, D.; Yang, Y.; Patel, V. R.; Guan, S.; Randall, A.; Baldi, P.; Rychnovsky, S. D.; Huang, L. Development of a Novel Cross-Linking Strategy for Fast and Accurate Identification of Cross-Linked Peptides of Protein Complexes. *Mol. Cell. Proteomics* **2011**, *10* (1), M110.002212.
- (10) Liu, F.; Rijkers, D. T. S.; Post, H.; Heck, A. J. R. Proteome-Wide Profiling of Protein Assemblies by Cross-Linking Mass Spectrometry. *Nat. Methods* **2015**, *12* (12), 1179–1184.
- (11) Liu, F.; Lössl, P.; Scheltema, R.; Viner, R.; Heck, A. J. R. Optimized Fragmentation Schemes and Data Analysis Strategies for Proteome-Wide Cross-Link Identification. *Nat. Commun.* **2017**, *8*, 15473.
- (12) Schweppe, D. K.; Chavez, J. D.; Lee, C. F.; Caudal, A.; Kruse, S. E.; Stuppard, R.; Marcinek, D. J.; Shadel, G. S.; Tian, R.; Bruce, J. E. Mitochondrial Protein Interactome Elucidated by Chemical Cross-Linking Mass Spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (7), 1732–1737.
- (13) Belsom, A.; Mudd, G.; Giese, S.; Auer, M.; Rappsilber, J. Complementary Benzophenone Cross-Linking/Mass Spectrometry Photochemistry. *Anal. Chem.* **2017**, *89* (10), 5319–5324.
- (14) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13* (11), 2498–2504.
- (15) Cline, S. M.; Smoot, M.; Cerami, E.; Kuchinsky, A.; Landys, N.; Workman, C.; Christmas, R.; Avila-campilo, I.; Creech, M.; Gross, B.; et al. Integration of Biological Networks and Gene Expression Data Using Cytoscape. *Nat. Protoc.* **2007**, *2* (10), 2366–2382.
- (16) Bastian, M.; Heymann, S.; Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Third Int. AAAI Conf. Weblogs Soc. Media* **2009**, 361–362.
- (17) Agapito, G.; Guzzi, P. H.; Cannataro, M. Visualization of Protein Interaction Networks: Problems and Solutions. *BMC Bioinf.* **2013**, *14* (Suppl 1), S1.
- (18) Courcelles, M.; Coulombe-Huntington, J.; Cossette, É.; Gingras, A. C.; Thibault, P.; Tyers, M. CLMSVault: A Software Suite for Protein Cross-Linking Mass-Spectrometry Data Analysis and Visualization. *J. Proteome Res.* **2017**, *16* (7), 2645–2652.
- (19) Heymann, M.; Paramelle, D.; Subra, G.; Forest, E.; Martinez, J.; Geourjon, C.; Deléage, G. MSX-3D: A Tool to Validate 3D Protein Models Using Mass Spectrometry. *Bioinformatics* **2008**, *24* (23), 2782–2783.
- (20) Kahraman, A.; Malmström, L.; Aebersold, R. Xwalk: Computing and Visualizing Distances in Cross-Linking Experiments. *Bioinformatics* **2011**, *27* (15), 2163–2164.
- (21) Kosinski, J.; von Appen, A.; Ori, A.; Karius, K.; Müller, C. W.; Beck, M. Xlink Analyzer: Software for Analysis and Visualization of Cross-Linking Data in the Context of Three-Dimensional Structures. *J. Struct. Biol.* **2015**, *189* (3), 177–183.

- (22) Lima, D. B.; de Lima, T. B.; Balbuena, T. S.; Neves-Ferreira, A. G. C.; Barbosa, V. C.; Gozzo, F. C.; Carvalho, P. C.; et al. SIM-XL: A Powerful and User-Friendly Tool for Peptide Cross-Linking Analysis. *J. Proteomics* **2015**, *129*, 51–55.
- (23) Riffle, M.; Jaschob, D.; Zelter, A.; Davis, T. N. ProXL (Protein Cross-Linking Database): A Platform for Analysis, Visualization, and Sharing of Protein Cross-Linking Mass Spectrometry Data. *J. Proteome Res.* **2016**, *15* (8), 2863–2870.
- (24) Schweppe, D. K.; Chavez, J. D.; Bruce, J. E. XLmap: An R Package to Visualize and Score Protein Structure Models Based on Sites of Protein Cross-Linking. *Bioinformatics* **2015**, *32* (2), btv519.
- (25) Schweppe, D. K.; Zheng, C.; Chavez, J. D.; Navare, A. T.; Wu, X.; Eng, J. K.; Bruce, J. E. XLinkDB 2.0: Integrated, Large-Scale Structural Analysis of Protein Crosslinking Data. *Bioinformatics* **2016**, *32* (17), 2716–2718.
- (26) Solis-Mezarino, V.; Herzog, F. CompleXView: A Server for the Interpretation of Protein Abundance and Connectivity Information to Identify Protein Complexes. *Nucleic Acids Res.* **2017**, *45* (W1), W276–W284.
- (27) Combe, C. W.; Fischer, L.; Rappsilber, J. XiNET: Cross-Link Network Maps with Residue Resolution. *Mol. Cell. Proteomics* **2015**, *14* (4), 1137–1147.
- (28) Grimm, M.; Zimniak, T.; Kahraman, A.; Herzog, F. XVis: A Web Server for the Schematic Visualization and Interpretation of Crosslink-Derived Spatial Restraints. *Nucleic Acids Res.* **2015**, *43* (W1), W362–W369.
- (29) Gutierrez, C. B.; Block, S. A.; Yu, C.; Soohoo, S. M.; Huszagh, A. S.; Rychnovsky, S. D.; Huang, L. Development of a Novel Sulfoxide-Containing MS-Cleavable Homobifunctional Cysteine-Reactive Cross-Linker for Studying Protein-Protein Interactions. *Anal. Chem.* **2018**, *90* (12), 7600–7607.
- (30) Gutierrez, C. B.; Yu, C.; Novitsky, E. J.; Huszagh, A. S.; Rychnovsky, S. D.; Huang, L. Developing a Novel Acidic Residue Reactive and Sulfoxide-Containing MS-Cleavable Homobifunctional Cross-Linker for Probing Protein-Protein Interactions. *Anal. Chem.* **2016**, *88* (3), 8315–8322.
- (31) Morimoto, Y.; Nakagawa, T.; Kojima, M. Small-Angle X-Ray Scattering Constraints and Local Geometry like Secondary Structures Can Construct a Coarse-Grained Protein Model at Amino Acid Residue Resolution. *Biochem. Biophys. Res. Commun.* **2013**, *431* (1), 65–69.
- (32) Bateman, A.; Martin, M. J.; O'Donovan, C.; Magrane, M.; Alpi, E.; Antunes, R.; Bely, B.; Bingley, M.; Bonilla, C.; Britto, R.; et al. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D158–D169.
- (33) van Zundert, G. C. P.; Trellet, M.; Schaarschmidt, J.; Kurkuoglu, Z.; David, M.; Verlato, M.; Rosato, A.; Bonvin, A. M. J. J. The DisVis and PowerFit Web Servers: Explorative and Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **2017**, *429* (3), 399–407.
- (34) van Zundert, G. C. P.; Bonvin, A. M. J. J. DisVis: Quantifying and Visualizing Accessible Interaction Space of Distance-Restrained Biomolecular Complexes. *Bioinformatics* **2015**, *31* (19), 3222–3224.
- (35) Klykov, O.; Steigenberger, B.; Pektaş, S.; Fasci, D.; Heck, A. J. R.; Scheltema, R. A. Efficient and Robust Proteome-Wide Approaches for Crosslinking Mass Spectrometry. *Nat. Protoc.* **2018**, *13* (12), 2964–2990.
- (36) Hennrich, M. L.; Groenewold, V.; Kops, G. J. P. L.; Heck, A. J. R.; Mohammed, S. Improving Depth in Phosphoproteomics by Using a Strong Cation Exchange-Weak Anion Exchange-Reversed Phase Multidimensional Separation Approach. *Anal. Chem.* **2011**, *83* (18), 7137–7143.
- (37) Vizcaino, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Rios, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; et al. ProteomeXchange Provides Globally Coordinated Proteomics Data Submission and Dissemination. *Nat. Biotechnol.* **2014**, *32* (3), 223–226.
- (38) Levenstein, I. V. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Cybern. Control Theory* **1966**, *10* (8), 707–710.
- (39) Noack, A. Energy Models for Graph Clustering. *J. Graph Algorithms Appl.* **2007**, *11* (2), 453–480.
- (40) Pajntar, B. Overview of Algorithms for Graph Drawing. *Knowledge Creation, Diffusion, and Utilization* **2006**, *3* (6), 1–4.
- (41) Huerta-Cepas, J.; Szklarczyk, D.; Forslund, K.; Cook, H.; Heller, D.; Walter, M. C.; Rattei, T.; Mende, D. R.; Sunagawa, S.; Kuhn, M.; et al. EGGNOG 4.5: A Hierarchical Orthology Framework with Improved Functional Annotations for Eukaryotic, Prokaryotic and Viral Sequences. *Nucleic Acids Res.* **2016**, *44*, D286–D293.
- (42) Rivals, I.; Personnaz, L.; Taing, L.; Potier, M. C. Enrichment or Depletion of a GO Category within a Class of Genes: Which Test? *Bioinformatics* **2007**, *23* (4), 401–407.
- (43) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605–1612.
- (44) Henikoff, S.; Henikoff, J. G. Amino Acid Substitution Matrices from Protein Blocks. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89* (22), 10915–10919.
- (45) Dominguez, C.; Boelens, R.; Bonvin, A. M. J. J. HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information. *J. Am. Chem. Soc.* **2003**, *125* (7), 1731–1737.
- (46) van Zundert, G. C. P.; Rodrigues, J. P. G. L. M.; Trellet, M.; Schmitz, C.; Kastiris, P. L.; Karaca, E.; Melquiond, A. S. J.; Van Dijk, M.; De Vries, S. J.; Bonvin, A. M. J. J. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **2016**, *428* (4), 720–725.
- (47) Jain, A. K. Data Clustering: 50 Years beyond K-Means. *Pattern Recognit. Lett.* **2010**, *31* (8), 651–666.
- (48) Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: Protein Structure and Function Prediction. *Nat. Methods* **2015**, *12* (1), 7–8.
- (49) Bienert, S.; Waterhouse, A.; De Beer, T. A. P.; Tauriello, G.; Studer, G.; Bordoli, L.; Schwede, T. The SWISS-MODEL Repository: New Features and Functionality. *Nucleic Acids Res.* **2017**, *45*, D313–D319.
- (50) Akamine, P.; Madhusudan; Wu, J.; Xuong, N. H.; Eyck, L. F. T.; Taylor, S. S. Dynamic Features of CAMP-Dependent Protein Kinase Revealed by Apoenzyme Crystal Structure. *J. Mol. Biol.* **2003**, *327* (1), 159–171.
- (51) Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K. P.; et al. STRING V10: Protein-Protein Interaction Networks, Integrated over the Tree of Life. *Nucleic Acids Res.* **2015**, *43*, D447–D452.
- (52) Ruepp, A.; Brauner, B.; Dunger-Kaltenbach, I.; Frishman, G.; Montrone, C.; Stransky, M.; Waegle, B.; Schmidt, T.; Doudieu, O. N.; Stümpflen, V.; et al. CORUM: The Comprehensive Resource of Mammalian Protein Complexes. *Nucleic Acids Res.* **2007**, *36*, D646–D650.
- (53) Hoopmann, M. R.; Mendoza, L.; Deutsch, E. W.; Shteynberg, D.; Moritz, R. L. An Open Data Format for Visualization and Analysis of Cross-Linked Mass Spectrometry Results. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (11), 1728–1734.
- (54) Wang, X. *PepXMLTab: Parsing PepXML Files and Filter Based on Peptide FDR*, 2018.