Journal of proteome research

pubs.acs.org/jpr

Technical Note

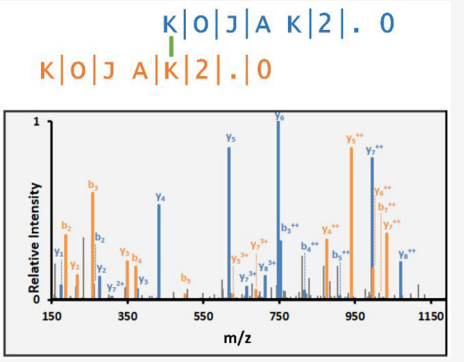
Improved Analysis of Cross-Linking Mass Spectrometry Data with Kojak 2.0, Advanced by Integration into the Trans-Proteomic Pipeline

Michael R. Hoopmann,* David D. Shteynberg, Alex Zelter, Michael Riffle, Andrew S. Lyon, David A. Agard, Qing Luan, Brad J. Nolen, Michael J. MacCoss, Trisha N. Davis, and Robert L. Moritz

Cite This: *J. Proteome Res.* 2023, 22, 647−655

Read Online

ABSTRACT: Fragmentation ion spectral analysis of chemically cross-linked proteins is an established technology in the proteomics research repertoire for determining protein interactions, spatial orientation, and structure. Here we present Kojak version 2.0, a major update to the original Kojak algorithm, which was developed to identify cross-linked peptides from fragment ion spectra using a database search approach. A substantially improved algorithm with updated scoring metrics, support for cleavable cross-linkers, and identification of cross-links between ^{15}N-labeled homomultimers are among the newest features of Kojak 2.0 presented here. Kojak 2.0 is now integrated into the Trans-Proteomic Pipeline, enabling access to dozens of additional tools within that suite. In particular, the PeptideProphet and iProphet tools for validation of cross-links improve the sensitivity and accuracy of correct cross-link identifications at user-defined thresholds. These new features improve the versatility of the algorithm, enabling its use in a wider range of experimental designs and analysis pipelines. Kojak 2.0 remains open-source and multiplatform.



KEYWORDS: *cross-linking mass spectrometry, XL-MS, Kojak, Trans-Proteomic Pipeline, PeptideProphet, iProphet, proteomics, computational proteomics, software tools, protein interaction*

INTRODUCTION

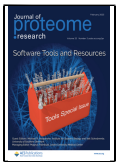
Shotgun mass spectrometry (MS/MS) analysis of chemically cross-linked proteins (XL-MS) has become a versatile tool in the field of proteomics.[1,2] Data analysis of cross-linked proteins has unique challenges; database and spectral library searching algorithms for standard shotgun analyses are not readily extensible to the analysis of chemically cross-linked proteins. In response to this shortcoming, many analytical tools have been developed to meet the unique challenges of cross-linking spectral analysis and have been recently reviewed and evaluated.[3] These tools are diverse in their functionality, capable of analyzing data from static[4−7] and cleavable cross-linkers,[8,9] and can incorporate isotope labeling into the cross-linkers, e.g., as shown in ref 10. Alternatively, the tools must also be able to analyze isotopically labeled proteins.[11] Additional tools incorporate such information into larger analysis pipelines to visualize structure[12−14] and interaction networks.[15]

Kojak was first developed as a modern implementation of a database search algorithm for shotgun MS analysis of chemically cross-linked proteins.[16] At its core, Kojak emulated some of the key features of the Comet algorithm for standard shotgun proteomics analyses,[17] and applying these features to the identification of cross-linked peptide sequences from MS/MS spectra. Foremost, Kojak was designed to be computationally efficient, capable of analysis with many different cross-linkers on both small and large data sets. The simple interface of the Kojak software, combined with adherence to open data standards, enabled its use with a diverse set of experimental conditions, analytical platforms, and pre- and postanalysis pipelines.

Development of Kojak has continued since its publication. These developments have culminated in Kojak version 2.0, which has several improvements and new features. These improvements include support for additional open formats and standards, further refinement to the search algorithm for efficiency, E-values to normalize the scores of the results, support for cleavable cross-linkers, and methods to identify cross-links between homomultimer subunits. Kojak 2.0 is integrated into the Trans-Proteomic Pipeline (TPP), and we use PeptideProphet and iProphet from the TPP to assign a probability value to each cross-linked-spectrum match (CSM) returned from Kojak. These probability values are then used to

Special Issue: Software Tools and Resources 2023

Received: October 17, 2022
Published: January 11, 2023



https://doi.org/10.1021/acs.jproteome.2c00670
J. Proteome Res. 2023, 22, 647−655

estimate error rates and determine which Kojak results are accepted at a user-defined error threshold. We show that this new pipeline developed around Kojak is both sensitive and accurate in the cross-links identified, an area of the field that has recently received additional scrutiny in an effort to improve XL-MS technology.[1,18−21] Here we present results highlighting the latest features of Kojak 2.0, and describe their use in the analysis of chemically cross-linked proteins.

METHODS

Search Database

Kojak 2.0 has multiple features for the creation and use of tailored protein sequence databases in the search analysis. Users provide a FASTA file containing presumed protein sequences to search (see Supplementary File 1). If validation after Kojak analysis requires decoy protein sequences to be searched, the user can opt to provide those sequences with a label in their FASTA file, or request Kojak generate decoys for them. Kojak's decoy generation algorithm fixes in place all digestion enzyme sites and reverses the amino acid sequences between them. This approach produces an equal number of decoy peptide sequences as target peptide sequences and identical masses. Thus, for every target peptide sequence searched against a spectrum, an equivalent decoy sequence is also searched against that spectrum. The decoy generation algorithm is dynamic to any enzyme digestion rule provided to Kojak, so that it can be used regardless of the digestion enzyme used. Palindromic sequences are mitigated through additional amino acid swaps. Kojak also recognizes protein name labels for isotopically labeled proteins, to distinguish these protein sequences from unlabeled protein sequences during the search process. The details of this feature are described in the methods below.

Identification of Linked Peptides

A two-pass approach is used to identify linked peptide pairs, similar to a previously described approach.[22] In the first pass, candidate peptides for the larger peptide mass are obtained through fragment ion matching the peptide sequence to the observed spectrum. In this pass, a candidate peptide is searched against a spectrum if it contains a site for binding to the cross-linker and its mass satisfies the following equation:

$$(m_{pre} - m_{xl})/2 \leq m_{\alpha} \leq (m_{pre} - m_{xl} - m_{\min}) \quad (1)$$

where m_{α} is the mass of the peptide in consideration, m_{pre} is the predicted precursor mass, m_{xl} is the mass of the cross-linker, and m_{min} is the user-defined smallest allowed peptide mass in the analysis. This equation is applicable to all spectra for which m_{pre} is greater than or equal to $2*m_{min} + m_{xl}$.

The mass of the unknown complementary peptide linked to the candidate peptide is computed as

$$m_c = m_{pre} - m_{\alpha} \quad (2)$$

where the complementary peptide mass (m_c) is the difference between the precursor mass (m_{pre}) and the mass of the peptide in consideration (m_{α}). m_c is treated as a modification mass on the potential site of linkage of the peptide. This mass can be moved to other sites on the peptide if it contains additional potential sites of linkage. Additionally, any differential modification masses arising from post-translational modifications (PTMs) or chemical modifications are considered. A dynamic tally of the best scoring peptides (i.e., those peptides with the highest cross-correlation score) in this first pass are maintained for each spectrum at a user-defined size, with recommendations between 5 and 15 peptides.

In the second pass, only peptide sequences whose masses sum to the predicted precursor mass with any of the short list of sequences from the first pass are searched. Additionally, only peptides containing a valid cross-linker binding site are considered. This second stage greatly reduces the number of peptide combinations that must be considered, and further prioritizes those combinations to the most likely candidate sequences based on fragment ion match information obtained in the first pass. This approach differs significantly from previous versions of Kojak, which would search all peptides and attempt to find two peptides among a list of hundreds of candidates that sum to the precursor ion mass. This approach is also algorithmically faster than the previous method used in Kojak. This improvement, combined with other improvements in software engineering has reduced computation time to less than 25% of the previously published version of Kojak.

Improved Scoring Metrics for Candidate Cross-Linked Peptides

Kojak uses a modified form of the Comet fast cross-correlation algorithm.[17,23] Because this cross-correlation score (Xcorr) is influenced by peptide length, expectation values (E-values) are now computed in Kojak 2.0, which are more useful than the cross-correlation score when performing downstream CSM validation. The E-value is computed from a linear least-squares regression of the log transform of the cumulative distribution function of the histogram of all cross-correlation scores to that MS2 spectrum. The user can define a minimum size of the histogram. If insufficient cross-correlation scores were recorded, additional cross-correlation scores to random peptide sequences of the same approximate mass are computed and added to the histogram. Additional E-value calculations are also performed for the individual peptides in the cross-linked sequence pair. In these cases, histograms are generated at a user defined size to include randomized peptide sequences of equivalent mass to the observed sequence, with a randomized site of linkage to the complementary peptide. These E-values better represent how good a Xcorr score is given the peptide length and observed spectrum peaks and are preferred to Xcorr scores when evaluating CSMs.

Isotope Labeling for Identification of Protein Dimer Interactions

Differentiation between self-linked proteins and cross-links between different subunits in homodimers and homomultimers is performed using ^{15}N-labeled techniques.[24,25] Briefly, the protein of interest is purified in both normal and ^{15}N-labeled forms. The resulting mass spectra following cross-linking and acquisition of data by mass spectrometry are then analyzed with Kojak 2.0 using the new *15N_filter* parameter. This parameter specifies a unique identifier word that is added to the FASTA protein identification line for the protein sequence that is to be analyzed with mass adjustments pertaining to the number of nitrogen atoms in each peptide from that protein sequence. For example, if a mixed normal and 15N-labeled "protein X" were cross-linked, the FASTA file would contain (1) the sequence for protein X, identified as ">protein-X", and (2) the sequence repeated for protein X, identified as, ">15n_protein-X". The *15N_filter* parameter would be set to "15n_", indicating that masses from peptides from ">15n_protein-X" are to be adjusted for the heavy nitrogen, while the masses from peptides from ">protein-X" are to be calculated as normal. In this manner,

https://doi.org/10.1021/acs.jproteome.2c00670
J. Proteome Res. 2023, 22, 647−655

peptides of identical sequence will have different masses depending on the protein sequence of origin. Cross-links identified as containing a normal and a ^{15}N-labeled peptide are therefore evidence of interaction between two separate protein subunits. For efficiency, it is not necessary for Kojak to search ^{15}N-labeled peptides for every sequence in the database, instead limiting this step to only the proteins that were labeled in the experiment.

Cleavable Cross-Linker Analysis

Mass spectrometry cleavable cross-linkers contain one or more labile bonds that break during collision-induced dissociation. When performing MS/MS analysis of peptides linked with cleavable cross-linkers, bond breakage can occur both along either peptide and at the labile cross-linker bonds. This creates fragment ion series that contain either the intact cross-linker and complement peptide, or simply a small mass addition equal to the remaining cross-linker after breaking at its labile bond. These predictable fragment ion products can be exploited to improve identification of cross-linked peptide sequences.[26] When a cleavable cross-linker is specified, Kojak considers the additional product masses that occur from breakage of the labile cross-linker bonds, using these masses as additional evidence when scoring CSMs.

Cross-Link Peptide Spectrum Match Validation

The use of the Trans-Proteomic Pipeline (TPP)[27] software tools of PeptideProphet[28] and iProphet[29] to validate the spectrum matches produced by Kojak has been previously suggested.[30] Here, these software tools were updated for the purposes of reading the results and scores as encoded by Kojak in pepXML format and modeling the new types of results produced within the framework of the TPP. The original PeptideProphet method for validation of Kojak CSM results relied on modeling the second-best expectation score in the cross-linked peptide combination as the PeptideProphet f-value, while the second-best Kojak computed score and the top best expectation score were used as additional discriminant models to improve the sensitivity of the PeptideProphet classifier. This approach was not successful in generating accurate or conservative probabilities when applied to Kojak 2.0 results and the validation for the new version of Kojak required a retooling of the models and software applied. The updates to validation software models in PeptideProphet, used to assign probabilities on the level of individual CSMs, consisted of only modeling the second-best expectation score and disabling the additional two discriminant models. This served to improve the accuracy of the estimated probabilities at the cost of reduced sensitivity in the classification of the CSMs. The additional discriminant models described are now optional in the software and can be enabled when running PeptideProphet by options XLSECOND and XLTOPEXP. Using PeptideProphet with the optional EXPECTSCORE flag enabled causes PeptideProphet to use the combined expectation score for the cross-links as the f-value. Additionally, the models in iProphet have been extended for Kojak 2.0 with the integration of an optional model called HETEROXL. This boolean model computes the likelihoods of observing cross-links of two peptides from different proteins (i.e., heteroprotein link), versus cross-links of two peptides from the same protein (i.e., self-protein link), among correct and among incorrect matches, as determined using the protein identifier assigned to each peptide sequence in a cross-link. Like the other models in iProphet, the HETEROXL model is learned by iProphet using the Expectation Maximization algorithm and applied to the results to adjust the probabilities of the CSMs (Supplementary Figure 1). Although, making these changes to the PeptideProphet model reduced the sensitivity of the PeptideProphet analysis, while improving the accuracy; iProphet with its models (including HETEROXL) enabled is able to recover the sensitivity of the classification given the accurate starting probabilities resultant from the updates to the models in PeptideProphet.

Bovine Arp2/3 Complex Sample Preparation

Bovine Arp2/3 complex was purified from calf thymus (Pel-Freez) as previously described[31] with an additional ion exchange column (MonoQ) step used as a final polishing step. After loading on the MonoQ column, the complex was eluted with a gradient of 25−300 mM NaCl in 10 mM Tris pH 8.0, 1 mM dithiothreitol (DTT). Pure fractions were dialyzed into 20 mM Tris pH 8.0, 50 mM NaCl, concentrated, and flash frozen.

Prior to cross-linking, the Arp2/3 complex was exchanged into HB100D (40 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), 100 mM NaCl, 1 mM DTT, pH 7) using Pierce Polyacrylamide Spin Desalting Columns (Catalog number: 89849, Thermo Fisher, Scientific). Cross-linking reactions were 100 μL and consisted of 25.32 uL of Arp2/3 (60 μg total protein) plus 270 μL of HB100D plus 2.34 μL of 25 mM CK-666 Arp2/3 complex inhibitor (Millipore-Sigma) plus 9.96 μL of 14.5 mM BS2 (BS2G-d0, Thermo Fisher Scientific) in HB100D or 9.96 μL of disuccinimidyl suberate (DSS, Thermo Fisher Scientific) in dimethyl sulfoxide (DMSO). Reactions were mixed and incubated for 10 min in an Eppendorf Thermomixer at 21 °C shaking at 1,000 rpm after which 100 μL was quenched by transfer to a fresh 1.5 mL Eppendorf tube containing 10 μL of 1 M ammonium bicarbonate plus 1 μL of 2 M β-mercaptoethanol. Reactions were reduced for 30 min at 37 °C with 10 mM DTT and alkylated for 30 min at room temperature with 15 mM iodoacetamide. Trypsin digestion was performed at 37 °C for 4 h with shaking at a substrate to enzyme ratio of 15:1 prior to acidification by addition of 250 mM HCl. Mass spectrometry was performed as previously described[13] by injection of 3 μL peptide digest onto a fused-silica capillary tip column (75-μm i.d.) packed with 30 cm of Reprosil-Pur C18-AQ (3-μm bead diameter, Dr. Maisch). Peptides were eluted from the column at 0.25 μL/min using an acetonitrile gradient. Mass spectrometry was performed on a QExactive-HF (Thermo Fisher Scientific) in data dependent mode.

Purification of ^{14}N/^{15}N Spc110 Covalent Heterodimers

Saccharomyces cerevisiae Spc110$^{1-276}$-SpyCatcher and -SpyTag were transformed into BL21(DE3) CodonPlus RIL (Agilent). Both constructs bore the Spc110-C225S mutation to prevent disulfide-mediated oligomerization. To generate ^{14}N-Spc110$^{1-276}$-Spc110-SpyTag, cultures were grown in Terrific Broth (Research Products International). For ^{15}N-Spc110$^{1-276}$-SpyCatcher, cultures were grown in the following medium:[32] 50 mM Na_2HPO_4, 25 mM KH_2PO_4, 10 mM NaCl, 5 mM $MgSO_4$, 0.2 mM $CaCl_2$, 1% (w/v) glucose, 0.1% (w/v) $^{15}NH_4Cl$, 0.25× BME vitamins mix (homemade based on the formula in Sigma-Aldrich item B6891), and 0.25× trace metals mixture.[33] Cultures were grown at 30 °C until reaching OD_{600} 0.3−0.4. The temperature was then decreased to 18 °C. Once the culture had reached OD_{600} 0.6−0.8, expression was induced with 0.6 mM isopropyl β-D-1-thiogalactopyranoside for 16−18 h. Cells were harvested by centrifugation then resuspended in lysis buffer (50 mM potassium phosphate pH 8, 300 mM NaCl, 5 mM ethylenediaminetetraacetic acid (EDTA), 1 mM DTT, 0.3%

https://doi.org/10.1021/acs.jproteome.2c00670
J. Proteome Res. 2023, 22, 647−655

Tween-20, 1× cOmplete protease inhibitor, EDTA-free (Millipore-Sigma)). Cells were lysed by Emulsiflex C3 (Avestin). Lysate was cleared by ultracentrifugation at 40,000 rpm for 30 min in a Type 45Ti rotor (Beckman-Coulter). Cleared lysate was applied to cOmplete His-Tag purification resin (Millipore-Sigma) and incubated for 1 h at 4 °C with gentle agitation. The column was then washed with 10 CV lysis buffer followed by 10 CV lysis buffer without Tween-20. Spc110 was then eluted with 4 CV of elution buffer (25 mM Tris pH 8.3, 75 mM NaCl, 5 mM EDTA, 1 mM DTT, 1× cOmplete protease inhibitor, EDTA-free (Millipore-Sigma), and 250 mM imidazole). Eluates were then diluted to <5 mS/cm conductivity with MonoQ buffer A (25 mM Tris pH 8.3, 1 mM DTT). The diluted eluates were then applied separately to a MonoQ 10/100 GL column (GE) pre-equilibrated in 2.5% MonoQ buffer B (25 mM Tris pH 8.3, 1 M NaCl, 1 mM DTT) in MonoQ buffer A. The column was then washed with 2 CV of 2.5% MonoQ buffer B, then eluted with a linear gradient from 2.5 to 50% MonoQ buffer B. $Spc110^{1-276}$ SpyCatcher and -SpyTag typically elute at approximately 17 mS/cm and 9 mS/cm conductivity, respectively. The concentration of the pooled fractions containing $Spc110^{1-276}$-SpyCatcher or -SpyTag were measured using Bradford protein assay reagent (Bio-Rad) using a BSA standard curve, then combined in a 1:1 molar ratio with the addition of TEV protease to cleave the His-tags. After 1 h, the Spc110 covalent adduct was further purified by size exclusion chromatography on S200 HiLoad 16/600 Superdex 200 pg (Cytiva) equilibrated in HB150 + 10% glycerol. Fractions containing undegraded Spc110 covalent adducts were then pooled, centrifugally concentrated, flash frozen in liquid nitrogen, and stored at −80 °C.

Prior to cross-linking, the protein was first buffer exchanged into HB100 buffer (40 mM HEPES, 100 mM NaCl, pH 7) using Pierce Polyacrylamide Spin Desalting Columns (Catalog number: 89849, Thermofisher, Scientific). A 200 μL cross-linking reaction was made by mixing 23.6 uL of desalted Spc110 (42 μg total protein) with 169.6 μL of HB100 and adding 6.8 μL of 14.5 mM BS3 (in HB100). The reaction was allowed to proceed for 2.5 min in an Eppendorf thermomixer at 21 °C shaking at 1,000 rpm after which 50 uL was quenched by transfer to a fresh 1.5 mL Eppendorf tube containing 5 μL of 1 M ammonium bicarbonate plus 1 μL of 2 M β-mercaptoethanol. Then, 18 μL of cross-linked protein was loaded onto an sodium dodecyl sulfate polyacrylamide gel electrophoresis gel (Biorad, Any kD Mini-PROTEAN TGX Precast Protein Gel; catalogue number 4569033) and run according to the manufacturer's instructions. A single band corresponding to the cross-linked Spc110 dimer was excised from the gel and subjected to in gel digestion using the following procedure. The gel band was cut into small pieces and washed with 200 μL water followed by 200 μL 50% acetonitrile 25 mM ammonium bicarbonate for five min followed by 200 μL acetonitrile for 1 min. Solvent was removed and the sample dried on a speed vac and reconstituted in 50 μL of 25 mM ammonium bicarbonate containing 10 mM tris(2-carboxyethyl)phosphine and incubated at 60 °C for 1 h. Excess liquid was removed and 50 μL of 25 mM ammonium bicarbonate containing 10 mM iodoacetamide was added and the sample incubated in the dark for 20 min. The sample was washed with 400 μL followed by 200 μL of 50% acetonitrile 25 mM ammonium bicarbonate for 5 min followed by 200 μL of acetonitrile for 1 min. The solvent was removed and the sample dried on a speed vac. The sample was reconstituted in 20 μL of 0.01 μg/uL promega trypsin in 25 mM ammonium bicarbonate. Additional 25 mM ammonium bicarbonate was added sufficient to cover the gel slice and the sample was digest overnight at room temperature. After digestion excess solution was removed and transferred to a new 1.5 mL Eppendorf tube. Then, 50 μL of acetonitrile was added to the gel slice, vortexed, and removed and combined with the solution in the 1.5 mL tube. Next, 50 μL of 60% acetonitrile 0.1% formic acid was added, vortexed, removed, and combined. The solution in the 1.5 mL tube was then dried in a speed vac and reconstituted in 20 μL of 0.1% formic acid. Finally, 3 μL of this solution was injected onto on a QExactive HF (Thermo Fisher Scientific) mass spectrometer run in data dependent mode as described above.

Cross-Linking Benchmark Standard

Additional analyses were performed using a previously published ground truth cross-linking data set.[34] These data are available via ProteomeXchange with identifier PXD014337.

Computational Data Analyses

All raw mass spectrometer data files were converted to mzML using msconvert (−mzML −zlib −filter "peakPicking true 1-" −filter "zeroSamples removeExtra") from ProteoWizard[35] prior to analysis. All computational analyses were performed using Kojak version 2.0 within the Trans-Proteomic Pipeline using the automated decoy sequence generation described above. Parameters for each tool for each analysis are provided in Supplementary Table 1. Kojak is available in standalone format at http://kojak-ms.org and bundled with the TPP at http://www.tppms.org. Novel data acquired for this study have been deposited to the ProteomeXchange Consortium via the PRIDE[36] partner repository with identifier PXD037492.

RESULTS

The newest features of Kojak 2.0 were explored using a ground truth data set.[34] In this data set, cross-linking was performed between 12 sets of synthesized peptides belonging to *S. pyogenes* Cas9. Cross-linking only occurred between peptides within a set, and sets consisted of seven to nine peptides. By design, all possible cross-links are known beforehand, and incorrect results are CSMs that contain a non-Cas9 peptide, or are Cas9 peptides from two different groups. Peptides were cross-linked with either DSS or DSSO.

Data were searched using Kojak 2.0 for CSM identification and validated using PeptideProphet and iProphet. A CSM probability cutoff of 0.9 was set, which gave an estimated <1% FDR. For the DSS-linked results, the searches of the data from three replicate injections were combined prior to validation, then CSMs were tallied from each replicate following validation. The number of correct and incorrect CSMs found in each replicate are reported in Table 1. These numbers met or exceeded the published results of this data set at this error threshold, with an observed error rate that remained below the estimate. That same study also identified 157−265 CSMs per replicate with StavroX and 312−438 CSMs per replicate with Xi at a 1% FDR threshold.[34] pLink identified 585−644 CSMs at a 1% FDR threshold, but with error rates in excess of the threshold in all replicates, including higher than 4% in one replicate. The CSMs were then grouped by unique peptide combinations, as correct CSMs are likely to be redundantly observed while incorrect CSMs tend to be random pairings. Here the number of unique cross-links (156−166 per replicate) were again similar or better to previously published results, with notably better error rates. Across all three replicates, the observed error rate was 1.62%, which, though higher than the CSM-level error

https://doi.org/10.1021/acs.jproteome.2c00670
J. Proteome Res. 2023, 22, 647−655

Table 1. Kojak Identified DSS-Cross-Linked CSMs at PeptideProphet/iProphet Estimated <1% Error Rate (CSM-Level)

	Correct (CSM)[a]	Incorrect (CSM)	Error Rate (CSM)	Correct (XL[b])	Incorrect (XL)	Error Rate (XL)
R1	405	2	0.49%	156	2	1.27%
R2	539	3	0.55%	166	2	1.19%
R3	490	1	0.20%	165	1	0.60%
Total	1434	6	0.42%	182	3	1.62%

[a]Results are evaluated for each CSM above the 0.9 probability threshold. [b]Multiple CSMs to the same cross-linked pair of peptides are combined and results are evaluated for each unique cross-linked peptide pair.

threshold, is expected and remained low. For comparison, StavroX identified 90−124 unique CSMs per replicate at error rates of 0−3.1% and Xi identified 141−163 unique CSMs at error rates of 1.4−3.0%. pLink identified 189−218 unique CSMs per replicate, but with error rates of 4.0−11.6%, well in excess of the desired 1% threshold. This same study also reported unique CSM results for an earlier version of Kojak, but using a 5% FDR threshold, and found 120−128 CSMs per replicate at error rates of 1.4−3.2%.[34] Overall, Kojak 2.0 with PeptideProphet and iProphet analysis showed high sensitivity and accuracy in its results.

Next, the search was performed on the DSSO-linked data from the same benchmarking standard to showcase the new cleavable cross-linking analysis features of Kojak 2.0. Here two sets of parameters were compared, either using the new cleavable cross-linking settings or not. All other parameters remained identical during the analysis. To replicate the stringency level of the analysis from the original publication, a larger, more challenging sequence database was used that contained the singular *S. pyogenes* Cas9 sequence, plus more than 100 additional sequences from the CRAPome.[37] A CSM probability cutoff of 0.9 was set, which gave an estimated <1% FDR. Using cleavable cross-linker optimizations showed an increase in the number of correct CSMs detected in the analysis, though an increase in the error rate was observed (Table 2). When looking at the unique cross-linked peptide pairs, the cleavable cross-linker optimizations more than doubled the number of correctly identified cross-links. The observed error rate among unique cross-linked peptides was higher than the desired 1% threshold set at the CSM-level, as expected, though the number of correct cross-links identified (216) was greater than the previously published results, and often at lower error rates. For example, XlinkX found 128 unique cross-linked peptides at a 29% error rate, and MeroX-RiseUP mode found 149 unique cross-linked peptides at 11% error rate. MeroX-Rise mode found 124 unique cross-linked peptides, but at a notably low 0.8% error rate.[34]

To illustrate Kojak 2.0 and PeptideProphet/iProphet validation with protein complex analysis, mass spectrometry data from bovine Arp2/3 complex cross-linked with BS2 or DSS were analyzed. A larger protein sequence database consisting of seven Arp2/3 complex subunit sequences and 60 bovine and human contaminant sequences was used in the analysis. Decoy sequences were produced in Kojak using the new decoy database generation feature. Following Kojak 2.0 and PeptideProphet/iProphet analysis, CSMs were uploaded to ProXL[38] and visualized by mapping to PDB structure 3UKU (Figure 1). When viewing the DSS results, at a 1% FDR estimated from the iProphet probability scores on the CSMs, 57 unique cross-linked residue pairs were mapped to six of the seven protein subunits (Figure 1A). All but one were within the expected cross-link distance restraints of 35 Å, with 51 of them falling within 25 Å. A distance density plot (Figure 1B) compares the observed CSM distance restraints compared to the set of all possible distance restraints obtainable from the structure, and shows the validated CSMs to belong solely to the small fraction of all possible CSMs that are within the expected distance restraints for DSS. The same analysis was repeated with the shorter cross-linker, BS2, and 35 unique cross-links were identified (Figure 1C), and all but three were within a 30 Å distance restraint threshold across all seven ARP2/3 subunits. The three outliers were only slightly beyond the threshold (all less than 36 Å), and the distribution of CSMs shows a bias toward short distance restraints when compared to the distribution expected from randomly assigning linked residues (Figure 1D). Together, these results show high conformity between prophet validated CSMs identified with Kojak and crystal structures.

Homodimers, and by extension homomultimers, are often difficult to study by typical XL-MS because it is unclear whether or not the two interacting peptides originate from the same or different subunits. A solution is to mix ^{15}N-labeled and unlabeled forms of the identical subunits and identify XL interactions between subunits as a mix of labeled and unlabeled peptide sequences. The observation of both ^{15}N-labeled and unlabeled fragment ions for two peptides cross-linked together provides spectral evidence that the cross-linked peptides originated from different protein subunits. The *S. cerevisiae* gamma tubulin small complex binds to the nuclear face of the spindle pole body via interaction with the dimeric coiled-coil protein Spc110. Identification of interacting domains between the two Spc110 subunits was performed by mixing unlabeled and ^{15}N-labeled Spc110$^{1-276}$ protein. To maximize the likelihood of forming a heavy-light dimer, the SpyCatcher-SpyTag system[39] was used. Unlabeled Spc110$^{1-276}$-SpyTag was produced and purified. ^{15}N-Spc110$^{1-276}$-SpyCatcher was produced and purified separately. Spc110$^{1-276}$-SpyTag and ^{15}N-Spc110$^{1-276}$-SpyCatcher were then mixed. Dimers subsequently formed between ^{15}N-Spc110$^{1-276}$-SpyCatcher and unlabeled Spc110$^{1-276}$-SpyTag are permanently trapped by formation of a covalent bond between the SpyCatcher and SpyTag. In this way samples were

Table 2. Additional Kojak CSMs Identified from DSSO Cleavable Crosslinker Analysis at PeptideProphet/iProphet Estimated <1% Error Rate (CSM-Level)

	Correct (CSM[a])	Incorrect (CSM)	Error Rate (CSM)	Correct (XL[b])	Incorrect (XL)	Error Rate (XL)
DSSO	760	3	0.39%	88	3	3.30%
DSSO, CLEAVABLE	2261	37	1.61%	216	13	5.68%

[a]Results are evaluated for each CSM above the 0.9 probability threshold. [b]Multiple CSMs to the same cross-linked pair of peptides are combined and results are evaluated for each unique cross-linked peptide pair.

https://doi.org/10.1021/acs.jproteome.2c00670
J. Proteome Res. 2023, 22, 647−655

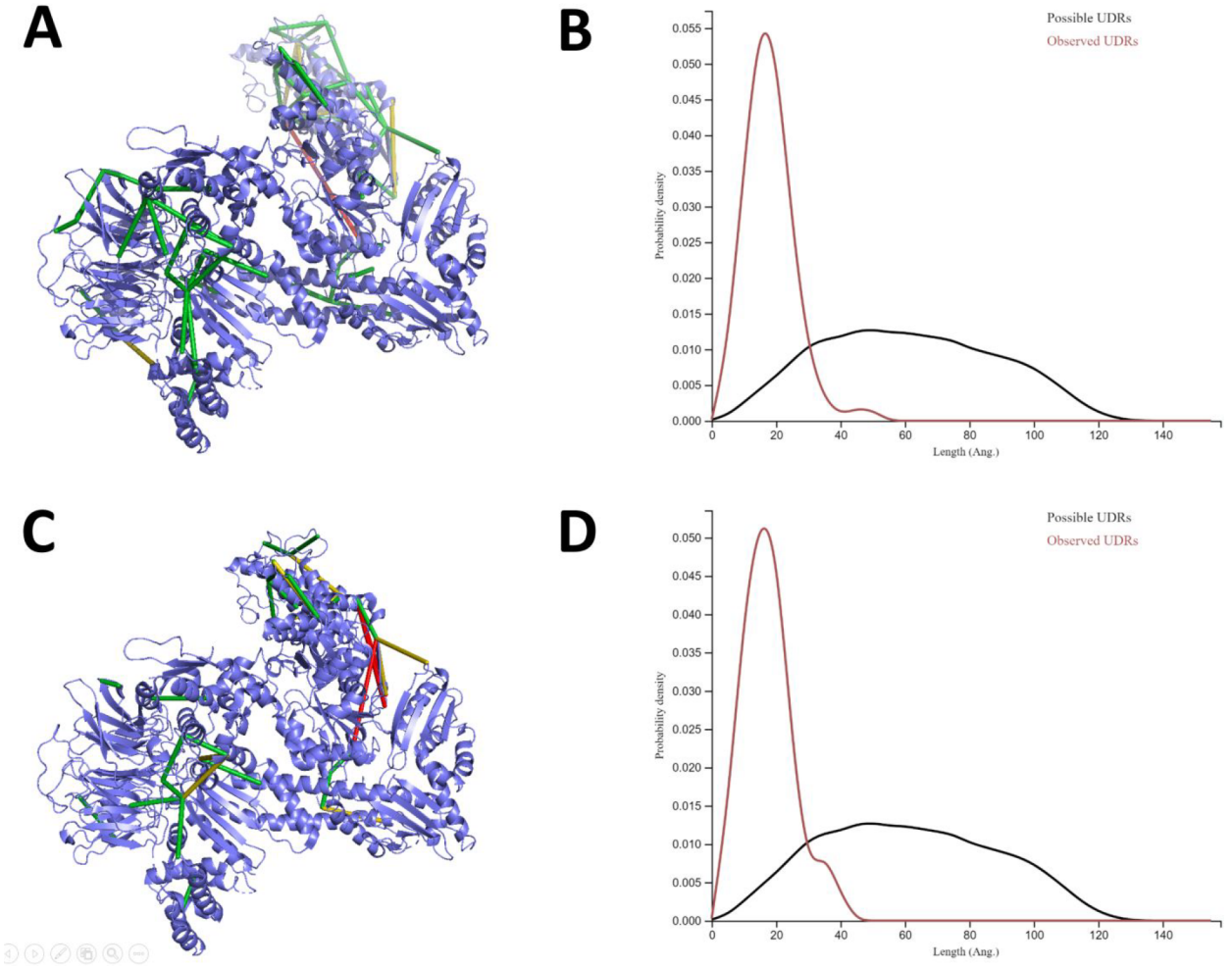


Figure 1. Structural views and cross-link distance restraint distributions for bovine Arp2/3 complex cross-linked with DSS (A and B) and BS2 (C and D). Cross-link distances are colored to represent Cα Lys-Lys distance constraints within 25 Å (green), 35 Å (yellow), and >35 Å (red) for DSS, and 20 Å (green), 30 Å (yellow), and >30 Å (red) for BS2. The distribution of observed unique distance restraints (UDRs) versus all possible UDRs are shown in panel B for DSS and panel D for BS2.

selectively enriched for Spc110$^{1-276}$ dimers between ^{15}N and unlabeled Spc110 subunits. Cross-linked spectra database searching with Kojak 2.0 was performed using a FASTA sequence database with two instances of the Spc110 sequence; however, one was annotated with a unique identifier to indicate all its peptides have additional mass due to the excess of ^{15}N. CSMs were validated with the prophets and uploaded to ProXL for visualization (Figure 2). Cross-links were plotted following a CSM probability threshold of 0.9 (<1% FDR estimation) and high frequency in the coiled-coil region of Spc110$^{1-276}$, which includes residues 164−276 of each subunit (Figure 2A). Self-links were found as well (i.e., unlabeled-to-unlabeled and ^{15}N-to-^{15}N peptides, Figure 2B), as expected when the cross-linker binds to two locations on the same subunit. These types of cross-links were distributed across the entirety of the dimer, in contrast to the mixed-label cross-links that indicate interaction between two subunits.

DISCUSSION

Kojak 2.0 has been improved through code optimization and new feature implementation. Combined within a large data processing suite, it reflects the diversity of new technology and analyses available for XL-MS, particularly cleavable cross-linker and ^{15}N-labeled homomultimer analyses. However, regardless of the nature of the cross-linking study, accurately estimating the error rate of the results is critical to the success of any XL-MS study. Kojak does not report error rates in its results by design, instead requiring use of the existing tools for such tasks. Kojak 2.0 now includes the ability to generate decoy sequences on-the-fly, to facilitate target-decoy validation strategies employed by many tools for error estimation. Previously, we have shown how to use Percolator[40] for CSM validation.[16] Here, we have integrated Kojak 2.0 into the Trans-Proteomic Pipeline to make use of PeptideProphet and iProphet for CSM validation. This versatility is meant to make Kojak extensible to still other validation tools (e.g., XiFDR[19]), and pipelines. These tools have different capabilities that can be tailored to the needs of the study. For example, Percolator is, as of the time of this writing, best for use only at the CSM-level, while iProphet extends error estimation to the peptide level. These levels might be sufficient for small studies, but large-scale analyses such as whole cell linking require even stronger thresholds and tools appropriate for them.[18,21] The plug-and-play nature of Kojak 2.0 makes it easy to adapt to the different tools required for XL-MS analysis at any scale.

It is important to note that beyond the scope of the latest features presented here, Kojak 2.0 remains open source and supporting open formats. Upon initial release, Kojak supported

https://doi.org/10.1021/acs.jproteome.2c00670
J. Proteome Res. 2023, 22, 647−655

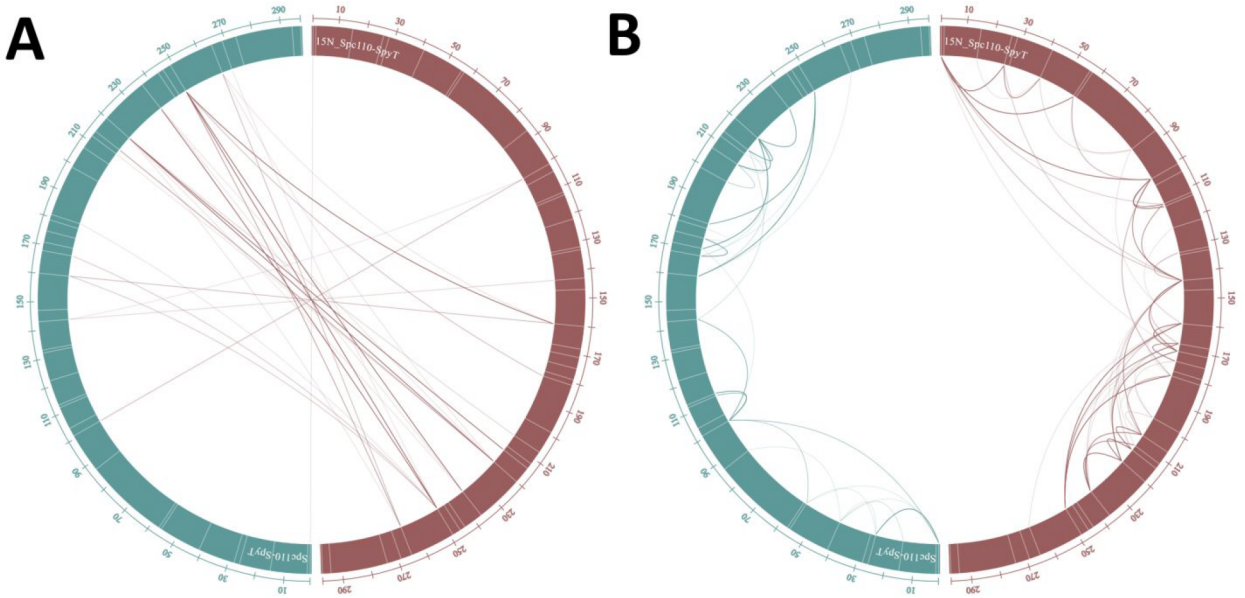


Figure 2. Cross-link distribution between unlabeled Spc-110$^{1-276}$ (teal) and ^{15}N-Spc-110$^{1-276}$ (red). Potential sites of cross-linker binding are marked at each position and observed cross-links as the arcs connecting two positions. Cross-link line weight indicates frequency of observing the interaction. (A) Interprotein cross-links most heavily connect the coiled-coil regions of the homodimer (residues 164−276). (B) Self-cross-links within each subunit are distributed across the entire sequence.

mzXML and mzML for input, and provided tab-delimited output. Because many existing tools for both preprocessing and postprocessing of searched spectral data require specific formats, Kojak 2.0 was extended to support MGF for input and pepXML[30] and mzIdentML[41] for output. These additional formats allow for the integration of Kojak into existing pipelines,[30,38,42,43] as well as facilitate integration to future pipelines. Such adaptations are particularly beneficial over all-in-one software suites, particularly if tools exist elsewhere that offer tangential features, such as spectral preprocessing, CSM validation, and results visualization that exceed the capabilities contained within any single software suite. We expect this versatility empower users in the analysis of particularly difficult data sets, while still providing a fast and simple interface accessible to anyone for use in most cross-linking analyses.

CONCLUSIONS

Kojak 2.0 offers many new features and capabilities for XL-MS data analysis. Because it is open source and adheres to open data formats and standards, it is easily incorporated into computational pipelines. We have demonstrated several of these new features and shown how, through use of the TPP, Kojak can be integrated into a robust pipeline for cross-linked peptide identification and validation.

ASSOCIATED CONTENT

Supporting Information

Supplementary Table 1: Software analysis parameters. Supplementary Figure 1: HETEROXL model of iProphet. Supplementary File 1: Configuration files and FASTA search databases for Kojak. The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.2c00670.

Supplementary Table 1: Software analysis parameters; Supplementary Figure 1: HETEROXL model of iProphet (PDF)

Supplementary File 1: Configuration files and FASTA search databases for Kojak (ZIP)

AUTHOR INFORMATION

Corresponding Author

Michael R. Hoopmann − *Institute for Systems Biology, Seattle, Washington 98109, United States;* orcid.org/0000-0001-7029-7792; Email: michael.hoopmann@isbscience.org

Authors

David D. Shteynberg − *Institute for Systems Biology, Seattle, Washington 98109, United States*

Alex Zelter − *Department of Biochemistry, University of Washington, Seattle, Washington 98195, United States;* orcid.org/0000-0002-5331-0577

Michael Riffle − *Department of Biochemistry, University of Washington, Seattle, Washington 98195, United States;* orcid.org/0000-0003-1633-8607

Andrew S. Lyon − *Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, California 94143, United States;* Present Address: A.S.L.: Department of Biophysics, UT Southwestern Medical Center, Dallas, TX 75390, USA; orcid.org/0000-0001-7681-4981

David A. Agard − *Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, California 94143, United States*

Qing Luan − *Department of Chemistry and Biochemistry, University of Oregon, Eugene, Oregon 97403, United States*

Brad J. Nolen − *Department of Chemistry and Biochemistry, University of Oregon, Eugene, Oregon 97403, United States*

Michael J. MacCoss − *Department of Genome Sciences, University of Washington, Seattle, Washington 98195, United States;* orcid.org/0000-0003-1853-0256

Trisha N. Davis − *Department of Biochemistry, University of Washington, Seattle, Washington 98195, United States*

https://doi.org/10.1021/acs.jproteome.2c00670
J. Proteome Res. 2023, 22, 647−655

Robert L. Moritz − *Institute for Systems Biology, Seattle, Washington 98109, United States;* orcid.org/0000-0002-3216-9447

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.2c00670

Author Contributions

M.R.H. developed the Kojak algorithm and conceptualized this study. M.R.H. and D.D.S. performed the data analysis. D.D.S. and M.R. provided additional software development. A.S.L., D.A.A., Q.L., B.J.N., A.Z., T.N.D., and M.J.M. performed experiments and acquired data for analysis. D.A.A., T.N.D., M.R.M., and R.L.M. provided resources for this study. All authors contributed to the paper.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded in part by the National Institutes of Health grants from the National institute General Medical Sciences R01GM087221, the National Heart, Lung, and Blood Institute R01HL133135, the Office of the Director S10OD026936, the National Science Foundation awards 1920268, Howard Hughes Medical Institute (D.A.A.), National Institute of General Medical Sciences R01 GM031627, R35GM118099, and P01 GM105537 (D.A.A.), National Science Foundation Graduate Research Fellowship Grant No. 1144247 (ASL), UCSF Discovery Fellowship (A.S.L.), and National institute General Medical Sciences P41GM103533 (M.J.M.).

■ REFERENCES

(1) Leitner, A.; Bonvin, A. M. J. J.; Borchers, C. H.; Chalkley, R. J.; Chamot-Rooke, J.; Combe, C. W.; Cox, J.; Dong, M.-Q.; Fischer, L.; Götze, M.; Gozzo, F. C.; Heck, A. J. R.; Hoopmann, M. R.; Huang, L.; Ishihama, Y.; Jones, A. R.; Kalisman, N.; Kohlbacher, O.; Mechtler, K.; Moritz, R. L.; Netz, E.; Novak, P.; Petrotchenko, E.; Sali, A.; Scheltema, R. A.; Schmidt, C.; Schriemer, D.; Sinz, A.; Sobott, F.; Stengel, F.; Thalassinos, K.; Urlaub, H.; Viner, R.; Vizcaíno, J. A.; Wilkins, M. R.; Rappsilber, J. Toward Increased Reliability, Transparency, and Accessibility in Cross-Linking Mass Spectrometry. *Structure* **2020**, *28* (11), 1259−1268.

(2) O'Reilly, F. J.; Rappsilber, J. Cross-Linking Mass Spectrometry: Methods and Applications in Structural, Molecular and Systems Biology. *Nat. Struct Mol. Biol.* **2018**, *25* (11), 1000−1008.

(3) Iacobucci, C.; Piotrowski, C.; Aebersold, R.; Amaral, B. C.; Andrews, P.; Bernfur, K.; Borchers, C.; Brodie, N. I.; Bruce, J. E.; Cao, Y.; Chaignepain, S.; Chavez, J. D.; Claverol, S.; Cox, J.; Davis, T.; Degliesposti, G.; Dong, M.-Q.; Edinger, N.; Emanuelsson, C.; Gay, M.; Götze, M.; Gomes-Neto, F.; Gozzo, F. C.; Gutierrez, C.; Haupt, C.; Heck, A. J. R.; Herzog, F.; Huang, L.; Hoopmann, M. R.; Kalisman, N.; Klykov, O.; Kukačka, Z.; Liu, F.; MacCoss, M. J.; Mechtler, K.; Mesika, R.; Moritz, R. L.; Nagaraj, N.; Nesati, V.; Neves-Ferreira, A. G. C.; Ninnis, R.; Novák, P.; O'Reilly, F. J.; Pelzing, M.; Petrotchenko, E.; Piersimoni, L.; Plasencia, M.; Pukala, T.; Rand, K. D.; Rappsilber, J.; Reichmann, D.; Sailer, C.; Sarnowski, C. P.; Scheltema, R. A.; Schmidt, C.; Schriemer, D. C.; Shi, Y.; Skehel, J. M.; Slavin, M.; Sobott, F.; Solis-Mezarino, V.; Stephanowitz, H.; Stengel, F.; Stieger, C. E.; Trabjerg, E.; Trnka, M.; Vilaseca, M.; Viner, R.; Xiang, Y.; Yilmaz, S.; Zelter, A.; Ziemianowicz, D.; Leitner, A.; Sinz, A. First Community-Wide, Comparative Cross-Linking Mass Spectrometry Study. *Anal. Chem.* **2019**, *91* (11), 6953−6961.

(4) Chu, F.; Baker, P. R.; Burlingame, A. L.; Chalkley, R. J. Finding Chimeras: A Bioinformatics Strategy for Identification of Cross-Linked Peptides. *Mol. Cell Proteomics* **2010**, *9* (1), 25−31.

(5) Götze, M.; Pettelkau, J.; Schaks, S.; Bosse, K.; Ihling, C. H.; Krauth, F.; Fritzsche, R.; Kühn, U.; Sinz, A. StavroX−a Software for Analyzing Crosslinked Products in Protein Interaction Studies. *J. Am. Soc. Mass Spectrom.* **2012**, *23* (1), 76−87.

(6) Rinner, O.; Seebacher, J.; Walzthoeni, T.; Mueller, L. N.; Beck, M.; Schmidt, A.; Mueller, M.; Aebersold, R. Identification of Cross-Linked Peptides from Large Sequence Databases. *Nat. Methods* **2008**, *5* (4), 315−318.

(7) Yang, B.; Wu, Y.-J.; Zhu, M.; Fan, S.-B.; Lin, J.; Zhang, K.; Li, S.; Chi, H.; Li, Y.-X.; Chen, H.-F.; Luo, S.-K.; Ding, Y.-H.; Wang, L.-H.; Hao, Z.; Xiu, L.-Y.; Chen, S.; Ye, K.; He, S.-M.; Dong, M.-Q. Identification of Cross-Linked Peptides from Complex Samples. *Nat. Methods* **2012**, *9* (9), 904−906.

(8) Mohr, J. P.; Perumalla, P.; Chavez, J. D.; Eng, J. K.; Bruce, J. E. Mango: A General Tool for Collision Induced Dissociation-Cleavable Cross-Linked Peptide Identification. *Anal. Chem.* **2018**, *90* (10), 6028−6034.

(9) Götze, M.; Pettelkau, J.; Fritzsche, R.; Ihling, C. H.; Schäfer, M.; Sinz, A. Automated Assignment of MS/MS Cleavable Cross-Links in Protein 3D-Structure Analysis. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (1), 83−97.

(10) Leitner, A.; Walzthoeni, T.; Aebersold, R. Lysine-Specific Chemical Cross-Linking of Protein Complexes and Identification of Cross-Linking Sites Using LC-MS/MS and the XQuest/XProphet Software Pipeline. *Nat. Protoc* **2014**, *9* (1), 120−137.

(11) Jaiswal, M.; Crabtree, N.; Bauer, M. A.; Hall, R.; Raney, K. D.; Zybailov, B. L. XLPM: Efficient Algorithm for the Analysis of Protein-Protein Contacts Using Chemical Cross-Linking Mass Spectrometry. *BMC Bioinformatics* **2014**, *15* (S11), S16.

(12) Leitner, A.; Joachimiak, L. A.; Bracher, A.; Mönkemeyer, L.; Walzthoeni, T.; Chen, B.; Pechmann, S.; Holmes, S.; Cong, Y.; Ma, B.; Ludtke, S.; Chiu, W.; Hartl, F. U.; Aebersold, R.; Frydman, J. The Molecular Architecture of the Eukaryotic Chaperonin TRiC/CCT. *Structure* **2012**, *20* (5), 814−825.

(13) Zelter, A.; Bonomi, M.; Kim, J. O.; Umbreit, N. T.; Hoopmann, M. R.; Johnson, R.; Riffle, M.; Jaschob, D.; MacCoss, M. J.; Moritz, R. L.; Davis, T. N. The Molecular Architecture of the Dam1 Kinetochore Complex Is Defined by Cross-Linking Based Structural Modelling. *Nat. Commun.* **2015**, *6*, 8673.

(14) 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* **2016**, *32* (2), 306−308.

(15) 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.

(16) Hoopmann, M. R.; Zelter, A.; Johnson, R. S.; Riffle, M.; MacCoss, M. J.; Davis, T. N.; Moritz, R. L. Kojak: Efficient Analysis of Chemically Cross-Linked Protein Complexes. *J. Proteome Res.* **2015**, *14* (5), 2190−2198.

(17) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: An Open-Source MS/MS Sequence Database Search Tool. *Proteomics* **2013**, *13* (1), 22−24.

(18) Lenz, S.; Sinn, L. R.; O'Reilly, F. J.; Fischer, L.; Wegner, F.; Rappsilber, J. Reliable Identification of Protein-Protein Interactions by Crosslinking Mass Spectrometry. *Nat. Commun.* **2021**, *12* (1), 3564.

(19) Fischer, L.; Rappsilber, J. Quirks of Error Estimation in Cross-Linking/Mass Spectrometry. *Anal. Chem.* **2017**, *89* (7), 3829−3833.

(20) Trnka, M. J.; Baker, P. R.; Robinson, P. J. J.; Burlingame, A. L.; Chalkley, R. J. Matching Cross-Linked Peptide Spectra: Only as Good as the Worse Identification. *Mol. Cell Proteomics* **2014**, *13* (2), 420−434.

(21) de Jong, L.; Roseboom, W.; Kramer, G. Towards Low False Discovery Rate Estimation for Protein-Protein Interactions Detected by Chemical Cross-Linking. *Biochim Biophys Acta Proteins Proteom* **2021**, *1869* (7), 140655.

(22) Chen, Z.-L.; Meng, J.-M.; Cao, Y.; Yin, J.-L.; Fang, R.-Q.; Fan, S.-B.; Liu, C.; Zeng, W.-F.; Ding, Y.-H.; Tan, D.; Wu, L.; Zhou, W.-J.; Chi, H.; Sun, R.-X.; Dong, M.-Q.; He, S.-M. A High-Speed Search Engine PLink 2 with Systematic Evaluation for Proteome-Scale Identification of Cross-Linked Peptides. *Nat. Commun.* **2019**, *10* (1), 3404.
(23) Eng, J. K.; Fischer, B.; Grossmann, J.; Maccoss, M. J. A Fast SEQUEST Cross Correlation Algorithm. *J. Proteome Res.* **2008**, *7* (10), 4598−4602.
(24) Taverner, T.; Hall, N. E.; O'Hair, R. A. J.; Simpson, R. J. Characterization of an Antagonist Interleukin-6 Dimer by Stable Isotope Labeling, Cross-Linking, and Mass Spectrometry. *J. Biol. Chem.* **2002**, *277* (48), 46487−46492.
(25) Lima, D. B.; Melchior, J. T.; Morris, J.; Barbosa, V. C.; Chamot-Rooke, J.; Fioramonte, M.; Souza, T. A. C. B.; Fischer, J. S. G.; Gozzo, F. C.; Carvalho, P. C.; Davidson, W. S. Characterization of Homodimer Interfaces with Cross-Linking Mass Spectrometry and Isotopically Labeled Proteins. *Nat. Protoc* **2018**, *13* (3), 431−458.
(26) Kolbowski, L.; Lenz, S.; Fischer, L.; Sinn, L. R.; O'Reilly, F. J.; Rappsilber, J. Improved Peptide Backbone Fragmentation Is the Primary Advantage of MS-Cleavable Crosslinkers. *Anal. Chem.* **2022**, *94* (22), 7779−7786.
(27) Deutsch, E.; Mendoza, L.; Shteynberg, D.; Hoopmann, M.; Sun, Z.; Eng, J.; Moritz, R.Trans-Proteomic Pipeline: Robust Mass Spectrometry-Based Proteomics Data Analysis Suite. *ChemRxiv***2022**, DOI: 10.26434/chemrxiv-2022-3c75n.
(28) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem.* **2002**, *74* (20), 5383−5392.
(29) Shteynberg, D.; Deutsch, E. W.; Lam, H.; Eng, J. K.; Sun, Z.; Tasman, N.; Mendoza, L.; Moritz, R. L.; Aebersold, R.; Nesvizhskii, A. I. IProphet: Multi-Level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates. *Mol. Cell Proteomics* **2011**, *10* (12), M111.007690.
(30) 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.
(31) Doolittle, L. K.; Rosen, M. K.; Padrick, S. B. Purification of Native Arp2/3 Complex from Bovine Thymus. *Methods Mol. Biol.* **2013**, *1046*, 231−250.
(32) Sivashanmugam, A.; Murray, V.; Cui, C.; Zhang, Y.; Wang, J.; Li, Q. Practical Protocols for Production of Very High Yields of Recombinant Proteins Using Escherichia Coli. *Protein Sci.* **2009**, *18* (5), 936−948.
(33) Studier, F. W. Protein Production by Auto-Induction in High Density Shaking Cultures. *Protein Expr Purif* **2005**, *41* (1), 207−234.
(34) Beveridge, R.; Stadlmann, J.; Penninger, J. M.; Mechtler, K. A Synthetic Peptide Library for Benchmarking Crosslinking-Mass Spectrometry Search Engines for Proteins and Protein Complexes. *Nat. Commun.* **2020**, *11* (1), 742.
(35) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M.-Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. *Nat. Biotechnol.* **2012**, *30* (10), 918−920.
(36) Perez-Riverol, Y.; Bai, J.; Bandla, C.; García-Seisdedos, D.; Hewapathirana, S.; Kamatchinathan, S.; Kundu, D. J.; Prakash, A.; Frericks-Zipper, A.; Eisenacher, M.; Walzer, M.; Wang, S.; Brazma, A.; Vizcaíno, J. A. The PRIDE Database Resources in 2022: A Hub for Mass Spectrometry-Based Proteomics Evidences. *Nucleic Acids Res.* **2022**, *50* (D1), D543−D552.
(37) Mellacheruvu, D.; Wright, Z.; Couzens, A. L.; Lambert, J.-P.; St-Denis, N. A.; Li, T.; Miteva, Y. V.; Hauri, S.; Sardiu, M. E.; Low, T. Y.; Halim, V. A.; Bagshaw, R. D.; Hubner, N. C.; Al-Hakim, A.; Bouchard, A.; Faubert, D.; Fermin, D.; Dunham, W. H.; Goudreault, M.; Lin, Z.-Y.; Badillo, B. G.; Pawson, T.; Durocher, D.; Coulombe, B.; Aebersold, R.; Superti-Furga, G.; Colinge, J.; Heck, A. J. R.; Choi, H.; Gstaiger, M.; Mohammed, S.; Cristea, I. M.; Bennett, K. L.; Washburn, M. P.; Raught, B.; Ewing, R. M.; Gingras, A.-C.; Nesvizhskii, A. I. The CRAPome: A Contaminant Repository for Affinity Purification-Mass Spectrometry Data. *Nat. Methods* **2013**, *10* (8), 730−736.
(38) 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.
(39) Li, L.; Fierer, J. O.; Rapoport, T. A.; Howarth, M. Structural Analysis and Optimization of the Covalent Association between SpyCatcher and a Peptide Tag. *J. Mol. Biol.* **2014**, *426* (2), 309−317.
(40) The, M.; MacCoss, M. J.; Noble, W. S.; Käll, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (11), 1719−1727.
(41) Vizcaíno, J. A.; Mayer, G.; Perkins, S.; Barsnes, H.; Vaudel, M.; Perez-Riverol, Y.; Ternent, T.; Uszkoreit, J.; Eisenacher, M.; Fischer, L.; Rappsilber, J.; Netz, E.; Walzer, M.; Kohlbacher, O.; Leitner, A.; Chalkley, R. J.; Ghali, F.; Martínez-Bartolomé, S.; Deutsch, E. W.; Jones, A. R. The MzIdentML Data Standard Version 1.2, Supporting Advances in Proteome Informatics. *Mol. Cell Proteomics* **2017**, *16* (7), 1275−1285.
(42) Riffle, M.; Jaschob, D.; Zelter, A.; Davis, T. N. Proxl (Protein Cross-Linking Database): A Public Server, QC Tools, and Other Major Updates. *J. Proteome Res.* **2019**, *18* (2), 759−764.
(43) Kertesz-Farkas, A.; Adoquaye Acquaye, F. L. N.; Bhimani, K.; Eng, J. K.; Fondrie, W. E.; Grant, C.; Hoopmann, M. R.; Lin, A.; Lu, Y. Y.; Moritz, R. L.; MacCoss, M. J.; Noble, W. S.Crux Toolkit for Analysis of Bottom-up Tandem Mass Spectrometry Proteomics Data. *J. Proteome Res.***2023**, DOI: 10.1021/acs.jproteome.2c00615.