analytical chemistry

ACS AUTHORCHOICE

pubs.acs.org/ac Article

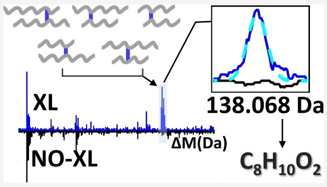
Open Search Strategy for Inferring the Masses of Cross-Link Adducts on Proteins

Moriya Slavin, Tamar Tayri-Wilk, Hala Milhem, and Nir Kalisman*

Cite This: *Anal. Chem.* 2020, 92, 15899−15907 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Development of new reagents for protein cross-linking is constantly ongoing. The chemical formulas for the linker adducts formed by these reagents are usually deduced from expert knowledge and then validated by mass spectrometry. Clearly, it would be more rigorous to infer the chemical compositions of the adducts directly from the data without any prior assumptions on their chemistries. Unfortunately, the analysis tools that are currently available to detect chemical modifications on linear peptides are not applicable to the case of two cross-linked peptides. Here, we show that an adaptation of the open search strategy that works on linear peptides can be used to characterize cross-link modifications in pairs of peptides. We benchmark our approach by correctly inferring the linker masses of two well-known reagents, DSS and formaldehyde, to accuracies of a few parts per million. We then investigate the cross-linking chemistries of two poorly characterized reagents: EMCS and glutaraldehyde. In the case of EMCS, we find that the expected cross-linking chemistry is accompanied by a competing chemistry that targets other amino acid types. In the case of glutaraldehyde, we find that the chemical formula of the dominant linker is C_5H_4, which indicates a ringed aromatic structure. These results demonstrate how, with very little effort, our approach can yield nontrivial insights to better characterize new cross-linkers.



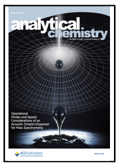
■ INTRODUCTION

Mass spectrometry is the main tool for the discovery and characterization of new types of protein modifications.[1] In peptide digests of protein mixtures, modifications manifest as unexplained differences between the measured and theoretical masses of certain peptides. For novel modifications, these mass differences are not known a priori and, therefore, it is not straightforward to infer them from the data. Different computational approaches were developed to address this challenge,[2] with the majority of them adopting an "open search" strategy (Inspect,[3] pMOD,[4] PepNovo,[5] modi,[6] modA,[7] MaxQuant,[8] MSFragger,[9] GPTMD,[10] and Open-pFind[11]). Open search uses MS/MS fragmentation spectra to identify the most likely peptide to be assigned to each MS/MS spectrum. The theoretical mass of the peptide is then subtracted from the measured mass of the precursor ion, thereby revealing the mass of the putative modification. Further information on the modification, such as the exact residue on which it occurs, can also be inferred from the analysis of the MS/MS spectrum.

Cross-linking is a special class of protein modifications that occur either within one protein or between two different protein subunits. At the level of the peptide digest, a cross-link modification manifests as a pair of peptides connected by a linker adduct. The approach of cross-linking coupled to mass spectrometry (XL-MS) makes use of such covalent links to study protein−protein interactions[12] and probe protein structures.[13] In XL-MS, intact proteins are reacted with a bifunctional cross-linking reagent that creates stable covalent links between side chains that are spatially close to each other. The proteins are then denatured and digested by a protease to yield a mixture of linear peptides and linked peptide pairs. This digest is analyzed by mass spectrometry, and dedicated applications can then identify which peptides underwent cross-linking. An identified cross-linked peptide pair implies that the peptides were spatially close in the context of the intact protein. The cross-links can then be converted into distance constraints for structural modeling. The identification process of cross-links in XL-MS relies on detection of ions with masses that correspond to the total mass of two peptides plus the mass of the linker adduct.[14] Clearly, for XL-MS to work, one must know the exact mass of this adduct.

New cross-linking reagents are constantly being added to the chemical toolbox of XL-MS.[15−18] The studies that describe these new reagents calculate the mass of the linker adduct from expert knowledge of the chemical reaction and then proceed to successfully detect it in the mass spectrometry data. It would have been preferable if the mass of the adduct could have been inferred directly from the data. This direct inference serves as an independent validation and helps in cases for which the prediction of the chemical reaction is not straightforward. Unfortunately, the excellent software tools that are available for identifying de novo modifications on linear peptides are not

Received: August 4, 2020
Accepted: November 9, 2020
Published: November 25, 2020



https://dx.doi.org/10.1021/acs.analchem.0c03292

applicable to the identification of cross-links between two peptides.

Here, we show that the open search framework, which works well for linear peptides, can be extended also to cross-linked peptide pairs. The approach works by identifying occurrences of two overlapping fragmentation patterns in the MS/MS fragmentation spectra, which are assumed to originate from a pair of cross-linked peptides. An implementation of the algorithm correctly infers the adduct masses of two well-known cross-linking reagents (DSS and formaldehyde) from benchmark mass spectrometry measurements. We then proceed to identify the adduct masses of less characterized cross-linking reagents: EMCS and glutaraldehyde.

■ MATERIALS AND METHODS

Protein Cross-Linking. A mixture solution of three purified proteins was prepared by reconstituting lyophilized protein powder in PBS (pH = 7.5). The proteins were bovine serum albumin (BSA), ovotransferrin, and α-amylase with respective final molarities in the mixture of 10, 10, and 20 μM. Each cross-linking experiment was conducted in 108 μL solution comprising a total protein mass of 260 μg. For formaldehyde cross-linking, a formalin solution (37% formaldehyde and 10% methanol, Sigma) was used, which was diluted with the proteins to a final formaldehyde concentration of 2%. The cross-linking reaction was incubated at RT for 20 min with agitation at 600 rpm. The reaction was quenched by adding 0.5 M ammonium bicarbonate. For DSS cross-linking, a 250 mM solution of DSS (Sigma) in DMSO was used, which was diluted with the proteins to a final DSS concentration of 3 mM. The cross-linking reaction was incubated at RT for 20 min with agitation at 600 rpm. The reaction was quenched by adding 30 mM ammonium bicarbonate. For glutaraldehyde cross-linking, a 70% glutaraldehyde solution (Sigma) was used, which was diluted with the proteins to the final glutaraldehyde concentrations: 0.0001, 0.001, 0.01, and 0.1%. The cross-linking reaction was incubated at RT for 20 min with agitation at 600 rpm. The reaction was quenched by adding 100 mM ammonium bicarbonate. We noted that quenching with higher concentrations of ammonium bicarbonate yielded yellow aggregates. For EMCS cross-linking, we only used BSA dissolved in PBS to a final concentration of 10 μM. A 10 mM solution of EMCS (Sigma) in PBS was prepared freshly prior to each experiment. The EMCS solution was diluted with the protein to the final EMCS concentrations of 0.1, 0.3, 1, and 3 mM. The cross-linking reaction was incubated at 30 °C for 30 min on a thermomixer at 600 rpm. The reaction was quenched by adding 30 mM ammonium bicarbonate and 5 mM DTT. In parallel to the above cross-linking experiments, we performed control experiments ("no XL" in the figures) in which the cross-linking step was omitted (but the quenching step was performed).

Mass Spectrometry. The proteins were precipitated in acetone at −80 °C for 1 h, followed by centrifugation at 13,000*g*. The pellet was resuspended in 20 μL of 8 M urea with 10 mM DTT. After 30 min, iodoacetamide was added to a final concentration of 25 mM and the alkylation reaction proceeded for 30 min. The urea was diluted by adding 200 μL digestion buffer (25 mM TRIS pH = 8.0; 10% acetonitrile), trypsin (Promega) was added at a 1:100 protease-to-protein ratio, and the protein was digested overnight at 37 °C under agitation. Following digestion, the peptides were desalted on C18 stage-tips and eluted by 55% acetonitrile. The eluted peptides were dried in a SpeedVac, reconstituted in 0.1% formic acid, and measured in a mass spectrometer. The samples were analyzed by a 120 min 0−40% acetonitrile gradient on a liquid chromatography system (Acquity M UPLC, Waters) coupled to a Q-Exactive Plus mass spectrometer (Thermo). We were careful not to increase the temperature of the sample above 40 °C through all of the preparation stages (alkylation, digestion, desalting, and in the analytical column of the LC) so as not to break the formaldehyde and glutaraldehyde cross-links. The RAW data files from the mass spectrometer were converted into MGF format in a Proteome Discoverer (Thermo), which was the input format for our analysis pipeline. The method parameters of the runs were as follows: data-dependent acquisition; full MS resolution, 70,000; MS1 AGC target, 1e6; MS1 maximum IT, 200 ms; scan range, 450−1800; dd-MS/MS resolution, 35,000; MS/MS AGC target, 2e5; MS2 maximum IT, 300 ms; loop count, top 12; isolation window, 1.1; fixed first mass, 130; HCD energy (NCE), 26; MS2 minimum AGC target, 800; charge exclusion: unassigned, 1, 2, 3, 8, >8; peptide match, off; exclude isotope, on; and dynamic exclusion, 45 s.

Implementation of the Open Search Pipeline. The pipeline described in Figure 1 was implemented into a MATLAB application. The application includes a graphical user interface that should allow any user to run it without difficulties. The parameters that were used in this work are set as the default parameters in the user interface. The following is a more detailed explanation of the working of the application. The sequence database is digested in silico into tryptic peptides. We allow two miscleavages and require a minimal peptide length of six residues. The mass spectrometry data comprise two data sets (in MGF format): one of the cross-linked sample and one of the same sample without the cross-linker. The MGF files are read and the MS/MS spectra are de-isotoped. The next computation is an estimate of the systematic MS1 offsets of the mass measurements. To this end, the application identifies all of the MS/MS events that can be confidently assigned to linear nonmodified peptides from the in silico digest. From these assignments, the systematic MS1 offset is calculated, and all subsequent steps of the application then compensate for it.

The next steps are repeated for every MS/MS event. All of the peptides are compared against the MS/MS spectrum of the event. The score of each peptide is set to be the total number of *b*- and *y*-fragments that match the spectrum within a tolerance of 8 ppm. We then take only peptides that have a score of 6 or higher. The number of peptides passing this threshold is usually small (less than 10), and consequently, the main computational burden is the calculation of the scores in the previous step. All of the possible pairs of the high-scoring peptides are enumerated and the total mass of each pair is calculated. The mass difference (ΔM) between the precursor mass of the event and the calculated mass of each pair is evaluated. These mass difference values are then binned into a histogram that adds up over all of the MS/MS events. This concludes the loop over all of the MS/MS events.

In the final step, the user is presented with a butterfly plot that compares the resulting ΔM histograms of the cross-linked and no XL samples. The analysis of this plot is manual, and the user has to zoom in on various peaks of interest and verify their relevance to the cross-linking. A bona fide linker mass will be high in the cross-linked sample and nearly zero for the no XL sample (e.g., Figure 2B).

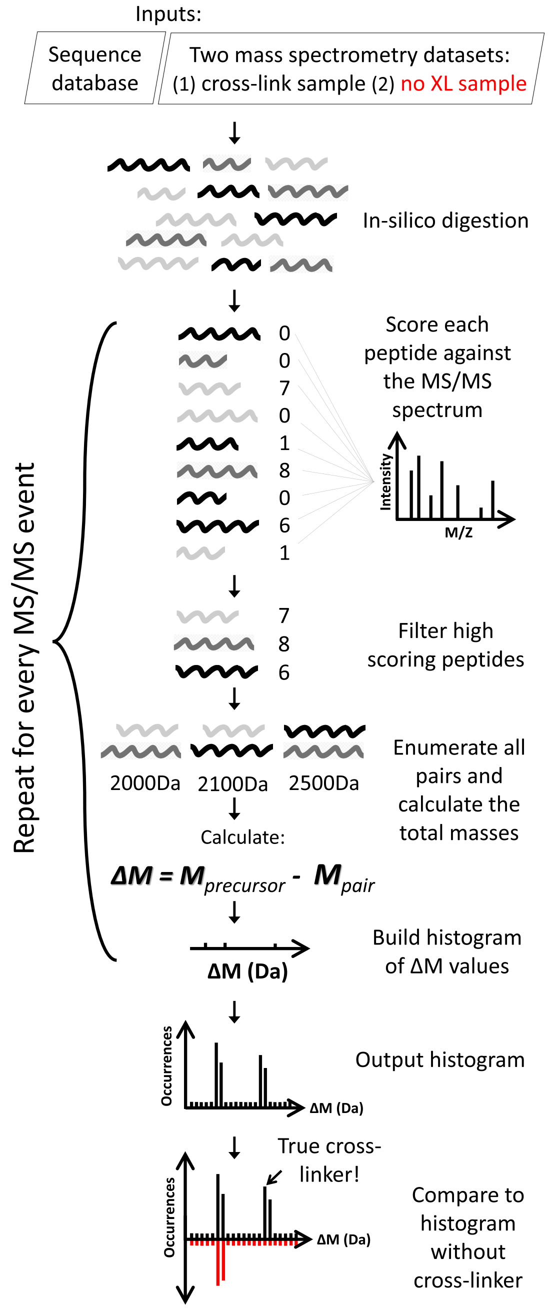

Inputs:
Sequence database
Two mass spectrometry datasets:
(1) cross-link sample (2) no XL sample
In-silico digestion
Score each peptide against the MS/MS spectrum
0
0
7
0
1
8
0
6
1
Intensity
M/Z
Filter high scoring peptides
7
8
6
2000Da 2100Da 2500Da
Enumerate all pairs and calculate the total masses
Calculate:
ΔM = M_precursor - M_pair
Build histogram of ΔM values
ΔM (Da)
Repeat for every MS/MS event
Occurrences
ΔM (Da)
Output histogram
True cross-linker!
Occurrences
ΔM (Da)
Compare to histogram without cross-linker

Figure 1. Computational pipeline to infer the mass of the linker adduct without any prior assumption on the chemistry. The pipeline is run once on mass spectrometry data from a cross-linked protein sample and once on data from the same sample but without the cross-linking (red). The two conditions are compared at the last step. The underlying assumption is that ΔM values that correspond to bona fide linker adducts will occur frequently and generate peaks only in the cross-link histogram.

Detailed Search for EMCS Cross-Links. We used an established search application[19] that exhaustively enumerates all of the possible peptide pairs. The search parameters were as follows: sequence database—the sequence of BSA; protease—trypsin, allowing up to three miscleavage sites; fixed modification of cysteine by iodoacetamide; and variable modification of methionine by oxidation. For the 193.0739 Da search, cross-linking must occur between a lysine and a cysteine. For the 540.1600 Da search, cross-linking must occur between two lysine residues; the cross-linker is never cleaved. MS/MS fragments to consider are as follows: b-ions and y-ions; MS1 tolerance, 6 ppm; MS2 tolerance, 8 ppm; cross-linker mass—one of three possible masses: 193.0739, 193.0739 + 1.00335, and 193.0739 + 2.0067. The three cross-linker masses were considered in turn in the calculation of the theoretical mass of the two cross-linked peptides. These masses address the occasional incorrect assignment of the monoisotopic mass by the mass spectrometer.[20]

A cross-link was identified as a match between a measured MS/MS event and a peptide pair if it fulfilled five conditions: (1) the mass of the precursor ion is within the MS1 tolerance of the theoretical mass of the linked peptide pair (with either of the three possible cross-link masses), (2) at least four MS/MS fragments were identified within the MS2 tolerance on each peptide, (3) the fragmentation score of the cross-link (defined as the number of all matching MS/MS fragments divided by the combined length of the two peptides) is 0.6 or higher, (4) the peptides are not overlapping in the protein sequence, and (5) there is no other peptide pair or linear peptide that matches the data with an equal or better fragmentation score.

Given the small size of the sequence database, we estimated the false detection rate (FDR) in the following way. The analysis of data from the 0.1 mM experiment was repeated 10 times with erroneous cross-linker masses of 61.0, 62.0, 63.0, ... 70.0 Da. This led to bogus identifications with fragmentation scores that were much lower than the scores obtained with the correct cross-linker mass. On average, one bogus identification had a fragmentation score above 0.5 in each decoy run, and none had a fragmentation score above 0.6. The run with the correct cross-linker mass (193.0739 Da) identified 19 cross-links above the 0.6 threshold. We therefore estimate the false detection rate (FDR) to be considerably less than 1 in 19 cross-links or 5%.

Detailed Search for a Glutaraldehyde Cross-Link. The search was similar to the one used above for EMCS, except for the following parameters: sequence database—sequences of BSA, ovotransferrin, and α-amylase; cross-linking can occur on any residue type; cross-linker is always cleaved; MS/MS fragments to consider are as follows: b-ions, y-ions, *b-ions (b-ions plus 64.0313 Da), and *y-ions (y-ions plus 64.0313 Da); and cross-linker mass—one of three possible masses: 64.0313, 65.0346, or 66.0380. Again, these three cross-linker masses address the assignment of the monoisotopic mass. The list was cut arbitrarily at a fragmentation score of 0.7. Given the small number of cross-links that were identified, no attempt was made to estimate the false detection rate.

Code Availability. The open search application is available for download at http://biolchem.huji.ac.il/nirka/software.html. The application is implemented in MATLAB and therefore requires MATLAB 2015 (or newer) to be installed before it can be run. The MATLAB implementation means that it can be run on any platform for which MATLAB is available (Windows, MACOS, and Linux). The run time is approximately 30 min on a standard desktop PC for MGF files with 15,000 MS/MS spectra.

■ RESULTS AND DISCUSSION

Our goal is to establish a workflow that will provide the user with the exact masses of the linker adducts formed by any cross-linking reagent of interest. The workflow comprises two simple cross-linking and mass spectrometry experiments,

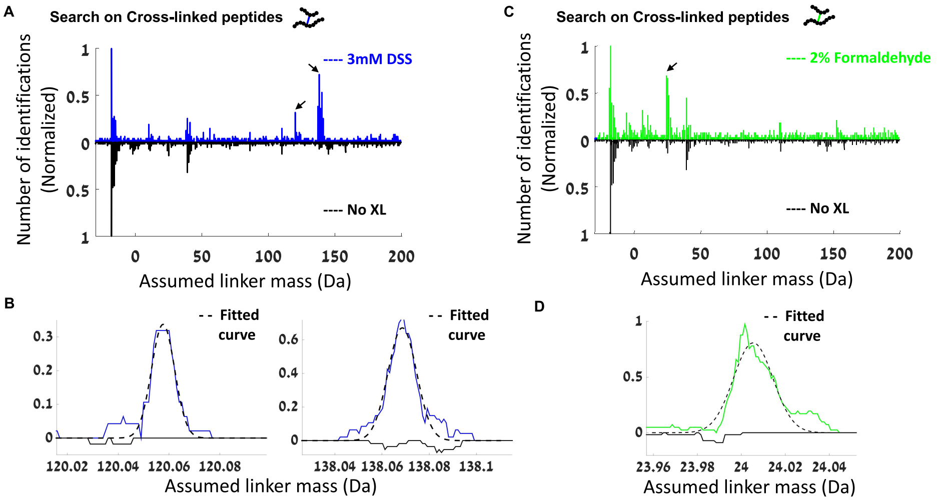


Figure 2. Applications of the search pipeline to data from XL-MS experiments with DSS and formaldehyde. (A) Butterfly plot contrasting the condition with the DSS cross-linker (top, blue) against the condition without the cross-linker (bottom, black). Shown are histograms of the number of identifications (y-axis) as a function of the assumed mass of the linker adduct (x-axis). Masses corresponding to true cross-links manifest as peaks that occur only in the top histogram (arrows). The plots are normalized by the number of identifications at −18 Da, which corresponds to nonmodified linear peptides with a miscleavage site. (B) Enlargements of the peaks at 120.05 and 138.06 Da, with fitted Gaussian models. (C, D) Similar butterfly plots for the formaldehyde cross-linking data.

followed by a dedicated computational analysis. We first discuss the experimental requirements and then proceed to describe the computational analysis in detail. Experimentally, the user should prepare a mixture of a few purified proteins on which to test the cross-linker. In this work, we mixed bovine serum albumin, ovotransferrin, and α-amylase. The low complexity of the sample is important later to ensure the fast run time of the computational analysis. The user should then perform two mass spectrometry analyses on the protein mixture: one on a sample that was cross-linked and a second on the same sample without cross-linking (no XL control). The preparation of both samples for mass spectrometry, as well as the mass spectrometry measurements, should follow standard protocols for XL-MS.[21,22] The mass spectrometry measurements provide two data-dependent MS/MS data sets that are the inputs for the computational analysis. The data sets are a series of MS/MS fragmentation spectra and the mass of the precursor ion for each spectrum.

Our open search computational approach is outlined in Figure 1. The inputs are the two mass spectrometry data sets and an in silico digest of the protein sequences comprising the mixture. A search is performed separately for each of the two data sets, and the results from the two searches are only compared at the very last step. The search scheme revolves around a series of steps that repeat for every MS/MS event. For each event, the computer assigns a score for every peptide from the in silico digest. The score equals the total number of b- and y-fragments from the peptide that can be matched to measured masses in the MS/MS fragmentation spectrum. We then filter the scores and keep only the peptides that scored above a certain threshold (six fragment matches in this study). We assume that if the MS/MS event indeed represents a pair of cross-linked peptides, then both peptides should be present in the filtered list. Accordingly, we enumerate all of the possible pairs and calculate the total mass of each pair. We then subtract these masses from the mass of the precursor ion and add these differences to a growing list of values. After all of the MS/MS events are processed, we calculate a final histogram from all of these values. The underlying assumption is that mass differences corresponding to the mass of the linker adduct will occur frequently. Consequently, they will appear as peaks in the cross-link histogram but will be absent in the histogram calculated for the control sample without the cross-linking.

We demonstrate the feasibility of the proposed approach by inferring the masses of the linker adducts for two well-characterized reagents: DSS (disuccinimidyl suberate) and formaldehyde. To this end, we cross-linked a mixture of three proteins with either 3 mM DSS or 2% formaldehyde and analyzed it by mass spectrometry. For control, we also analyzed the protein mixture without any cross-linking. Figure 2 shows the butterfly plots of the histograms that resulted under these conditions. For DSS, we see two very pronounced peaks (Figure 2A, arrows) that occur only for the cross-linking condition. Fitting the two peaks with a Gaussian model (Figure 2B) reveals them to be centered around 138.0686 and 120.0576 Da. These values are very close to 138.0681 and 120.05753 Da, which are the known adduct masses for DSS cross-links of type 2 (between two peptides) and type 1 (a loop within a single peptide), respectively. The absolute error in the inference of the adduct mass is 0.0005 Da. This corresponds to a relative error of 0.2 parts per million when assuming a typical mass of 2500 Da for the cross-linked peptides. For formaldehyde, we see one very pronounced peak (Figure 2C, arrow) that occurs only for the cross-linking condition. Fitting the peak with a Gaussian model (Figure 2D) reveals it to be centered around 24.0054 Da. This value is very close to 24.000 Da, which was recently shown to be the adduct mass of formaldehyde cross-links.[23] The absolute error in the formaldehyde case is 0.0054 Da, which corresponds to a relative error of 2.2 parts per million.

The highest peaks in the butterfly plots (Figure 2A,C) occur at −18.01057 Da (water loss) for both the cross-linked and control samples. This is an expected artifact corresponding to the numerous cases of unmodified single linear peptides with a trypsin miscleavage site. These peptides may be erroneously identified (in terms of total mass) as two separate peptides that are cross-linked by a linker of −18 Da. The difference

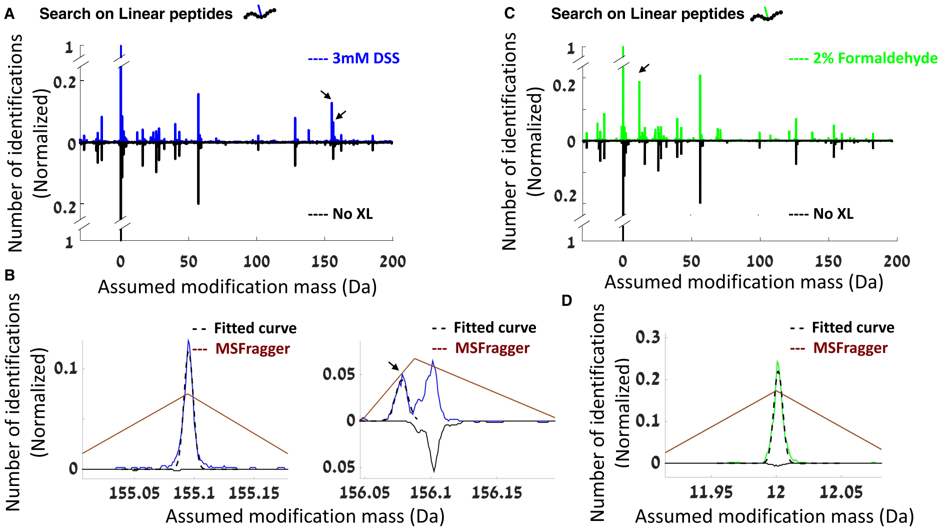


Figure 3. Variant of the pipeline that searches for modifications on linear peptides. (A) Butterfly plot contrasting the condition with the DSS cross-linker (top, blue) against the condition without the cross-linker (bottom, black). Masses corresponding to modifications induced by the cross-linker manifest as peaks that occur only in the top histogram (arrows). The plots are normalized by the number of identifications at 0 Da, which corresponds to linear peptides without any modifications. (B) Enlargements of the peaks at 155.09 and 156.08 Da, with fitted Gaussian models. An arrow indicates the true peak at 156.08 Da. We also show the results from MSFragger, which generally agreed closely with our analysis. Only in one case did our results differ from the output of MSFragger (right panel). (C, D) Similar butterfly plots for the formaldehyde cross-linking data.

corresponds to the water mass that trypsin would have added if the cleavage site was in fact cut. One may argue that our analysis should filter out all cases pertaining to consecutive peptides,[24] as they have very little value for the structural biology aspects of XL-MS. However, we think that for the purpose of characterizing the chemistry, short-range cross-links may be informative and therefore do not exclude them from this analysis.

Our open search approach infers the linker masses with relative accuracies of less than 3 ppm, which are expected, given the specifications of our mass spectrometer (Q-Exactive Plus). These accuracies are sufficient to correctly and easily infer the chemical formulas using one of several available web-based tools. We chose to work with ChemCalc,[25] which finds possible chemical formulas that match a given monoisotopic mass within a user-defined tolerance. For example, for the calculated mass of 138.0686 Da, the top hit of ChemCalc is $C_8H_{10}O_2$, which is indeed the correct chemical formula for the DSS linker.

A major strength of our pipeline is that it does not require any knowledge of the chemical properties of the cross-link. For example, we did not have to assume which residue types participate in the cross-linking. We also did not assume whether the link is cleavable (as is the case for formaldehyde) or not (as is the case for DSS). Also of note is the fact that the open search is not limited to any particular mass range. Figure 2 shows the results from −30 to 200 Da for the purpose of clarity. This is in fact an excerpt from a wider search in the −100 to 700 Da range, which did not find any additional peaks that were exclusive to the cross-link condition. The search is also not limited in its resolution. Here, the binning of the histogram was chosen to be 0.001 Da, which is appropriate for the accuracy of our mass spectrometer. Yet, coarser or finer binning can be set by the user with the tradeoff of faster or slower run times, respectively.

The open search pipeline has been implemented in a MATLAB environment and can therefore be run on any platform and operating system that supports MATLAB (see Code Availability). The run time is approximately 30 min on a standard desktop computer for processing both mass spectrometry files (XL and no XL) with ~15,000 MS/MS spectra each.

Search for Modifications on Linear Peptides. The computational pipeline in Figure 1 can also search for modifications on linear peptides if the peptide pairing step is skipped. The focus of this work is not linear peptides because excellent tools, such as MSFragger,[9] are available for this purpose. Yet, we thought it appropriate to include a short report on the performance of our application on linear peptides, if only for the purpose of quality assurance. To this end, we used the same protein samples as above and searched for modifications that are induced by either DSS or formaldehyde on linear peptides. Figure 3 shows the butterfly plots of the histograms that resulted under each of the conditions. For DSS, we see two very pronounced peaks (Figure 3A, arrows) that occur only for the cross-linking condition. Fitting the two peaks with a Gaussian model (Figure 3B) reveals them to be centered around 155.0953 and 156.0780 Da. These values are very close to 155.0946 and 156.0786 Da, which are the known dead-end modifications of DSS. They correspond to a DSS molecule that is attached to a peptide on one side, while its other side is either neutralized by ammonium bicarbonate or hydrolyzed, respectively. For formaldehyde, we see one very pronounced peak (Figure 3C, arrow) that occurs only for the cross-linking condition. Fitting the peak with a Gaussian model (Figure 3D) reveals it to be centered around 12.0015 Da. This value is very close to 12.000 Da, which is the well-known mass for the Schiff base modification induced by formaldehyde.[23]

We see that also in the case of linear peptides our application is able to infer the masses of the modifications with relative accuracies of 1−2 parts per million. In general, MSFragger gave the same results as our application, while being much faster in its run time. However, we observed one case in which the results of our application and MSFragger disagreed (Figure 3B, right panel). In the case of two close peaks (of which, only the

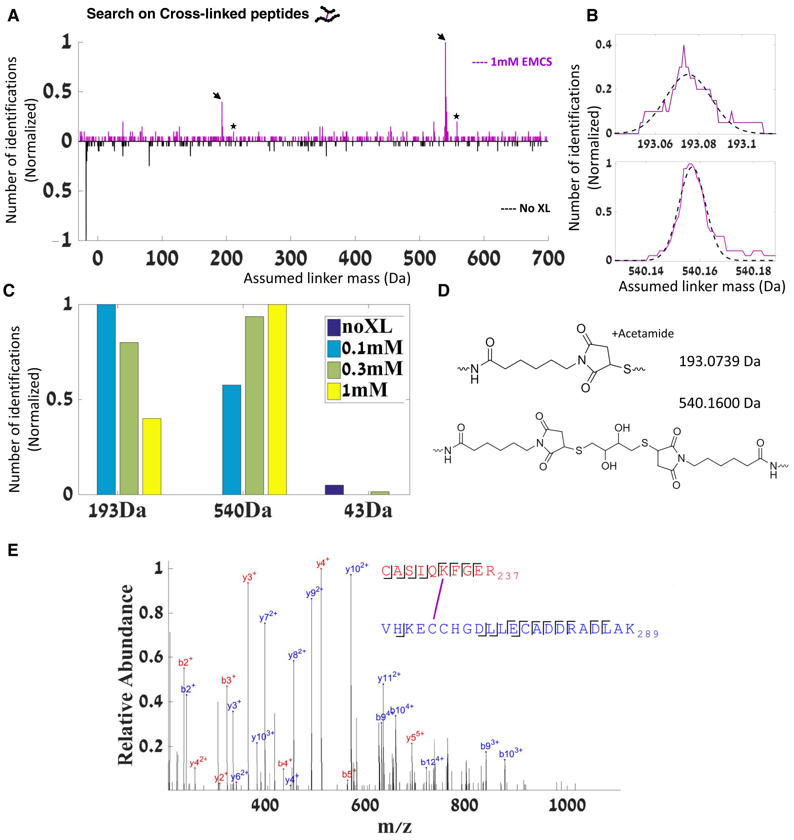


Figure 4. (A) Butterfly plot contrasting the cross-linking with EMCS (top) against the control without the cross-linker (bottom). Arrows point to prominent peaks that occur only in the cross-linked sample. Stars mark peaks that are likely the result of hydrolysis of the melamide moiety of EMCS.[26] (B) Enlargements of the peaks at 193 and 540 Da, with fitted Gaussian models. (C) The heights of the two peaks show opposite trends with increasing EMCS concentration. The height of a peak that is not related to EMCS cross-linking (43 Da) is shown as a control. (D) Proposed chemical structures for the two adducts. The top structure is the expected linker chemistry between the side chains of a lysine and a cysteine. The bottom structure comprises two EMCS molecules at the end of lysine side chains bridged by DTT. (E) Annotated MS/MS spectrum of a 193 Da cross-link between two peptides of BSA.

left one is a DSS modification), we see that MSFragger reports a single value that falls between the peaks. This is probably due to an optimization step in MSFragger that aims to increase the accuracy but fails in scenarios of two adjacent peaks.

Mass Analysis of EMCS Cross-Links in Proteins. The sulfo-EMCS (*N*-(*ε*-malemidocaproyl)-sulfosuccinimide ester) reagent is a hetero-bifunctional protein cross-linker. It targets sulfhydryl (−SH) functional groups with a melamide moiety at one end and primary amines (−NH_2) with an NHS ester moiety at the other end. The predicted chemistry for this reagent will give a linker adduct with a mass of 193.0739 Da ($C_{10}H_{11}NO_3$) between the side chains of a cysteine and a lysine. Our original aim was to verify this prediction and better characterize its usage for XL-MS. To this end, we cross-linked bovine serum albumin with EMCS under physiological buffer conditions (pH = 7.5). We next added ammonium bicarbonate and dithiothreitol (DTT) to quench both functionalities. We then denatured the protein with urea, alkylated it with iodoacetamide, digested it with trypsin, and analyzed it by mass spectrometry.

Figure 4 shows the results of our computational pipeline for the EMCS data. They reveal two dominant peaks (around 193.0753 and 540.1572) that are not observed in samples that were not cross-linked (Figure 4A,B). Both peaks are surprising and deviate from the expected chemistry of EMCS. While the 193 Da peak seemingly validates the predicted chemistry, we note that we performed our search under the premise that cysteine side chains are fixedly modified with acetamide. Therefore, the 193 peak implies that the linker adduct has an attached acetamide moiety resulting from the alkylation step. Given the current data, we cannot determine where exactly this acetamide is located. Yet, we can conclude that iodoacetamide is highly reactive toward the cysteine−melamide region even after the cross-linking step was completed.

The 540 Da peak is likewise unexpected and corresponds best to the chemical formula of $C_{24}H_{32}N_2O_8S_2$ (exact mass of 540.1600 Da). We further observed that this second linker can occur between peptides that do not have any cysteine residue. We suggest that this linker comprises two EMCS molecules attached to the ends of two lysine side chains and bridged by a DTT molecule (Figure 4D). Because DTT was not present in the initial buffer, we assume that this reaction is finalized only at the quenching stage. Interestingly, the two cross-linking reactions appear to be somewhat competing (Figure 4C), with the 540 Da reaction becoming considerably more dominant at higher EMCS concentrations.

Boyatzis et al.[26] showed that melamide might undergo several modifications during the standard preparation of samples for mass spectrometry. One of the modifications is the hydrolysis of the melamide ring, resulting in the addition of 18 Da to that moiety. Indeed, we see minor peaks corresponding to 193 + 18 and 540 + 18 Da also in our butterfly plot (Figure 4A, stars). Ammonium bicarbonate was shown to enhance the formation of hydrolyzed species,[26] and is probably the causative agent also in our samples. All of these findings demonstrate that the actual EMCS linking chemistry is rather different from the one predicted, thus highlighting the importance of an unbiased open search.

Although our pipeline successfully finds the masses of the linkers, it is not intended to function as a search application. To this end, we coded the two masses (193 and 540 Da) into a dedicated search application for cross-links[19] (Methods). We were able to identify 33 and 112 cross-links for the two masses, respectively (Tables S1 and S2). The linker of EMCS appears to be noncleavable and explains well the resulting MS/MS fragmentation spectra (Figure 4E). The cross-links can be mapped onto the crystallographic structure of bovine serum albumin, and show good agreement with the structure. Interestingly, the median $C\alpha{-}C\alpha$ distance between residues that are cross-linked by the 193 Da linker is 15 Å, whereas the median distance spanned by the 540 Da linker is 21 Å. This is in accord with the longer chemical structure of the 540 Da linker.

Mass Analysis of Glutaraldehyde Cross-Links in Proteins. Glutaraldehyde is one of the most effective reagents for protein cross-linking. It has numerous applications in medicine, histochemistry, microscopy, enzyme technology, chemical sterilization, and pharmaceutical sciences.[27] Despite this extensive usage, glutaraldehyde is not used in XL-MS. A partial explanation for this discrepancy may be the insufficient understanding of the cross-linking chemistry, which in turn prevents search programs from identifying glutaraldehyde cross-links. Glutaraldehyde is known to form many modifications and chemical intermediates on proteins.[27] However, the dominant form of long-range cross-links (i.e., cross-links that bridge residues that are far from each other on the protein sequence) is poorly understood. Here, we aim to better characterize the cross-link adducts that glutaraldehyde forms on structured proteins under physiological conditions. To this end, we cross-linked the above mixture of three proteins with glutaraldehyde and analyzed its digest by mass spectrometry. Interestingly, we observed that glutaraldehyde required significantly lower working concentrations than formaldehyde. Already a 0.5% glutaraldehyde solution caused significant aggregation that abolished the potential of trypsin to digest the proteins. Therefore, the similarity in chemical groups between formaldehyde and glutaraldehyde does not necessarily imply similarity in the reactivity.

We observed that glutaraldehyde forms a very prominent modification on linear peptides centered around 64.0350 Da (Figure S1). This modification strongly correlates with glutaraldehyde concentration and fits the chemical formula of C_5H_4 (theoretical mass of 64.0313 Da). Because the chemical formula of glutaraldehyde is $C_5H_8O_2$, the small number of hydrogens in the modification suggests that it is likely a heterocyclic aromatic ring, perhaps a pyridine at the end of a lysine side chain. The open search results for linker adducts are presented in Figure 5. They reveal three dominant peaks that are not observed in a sample that was not cross-linked (Figure 5A,B). Two of the peaks are centered around 64.03465 and 128.0614 Da and likely correspond to a single or double occurrence of the above modification (C_5H_4). A third peak is centered around 32.0005 Da, which does not correspond to any chemical formula within a reasonable tolerance. A deeper examination of the data showed that all of the peptide pairs that contributed to the 32 Da peak in the histogram originated from the same stretch of 40 amino acids in the sequence of bovine serum albumin. In contrast, the peptide pairs that contributed to the 64 and 128 Da peaks originated from many different locations in the sequences. We are forced to conclude that the 32 Da peak does not represent a true cross-link, but rather a very localized artifact that we cannot explain.

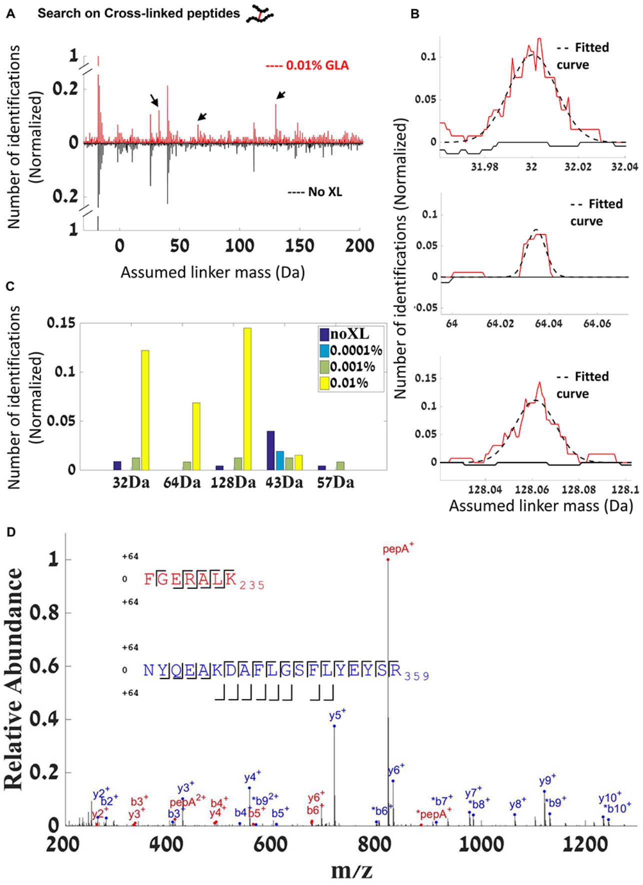


Figure 5. (A) Butterfly plot contrasting the condition of cross-linking with glutaraldehyde (top) against the control without the cross-linker (bottom). Arrows point to prominent peaks that occur only in the cross-linked sample. (B) Enlargements of the peaks at 32, 64, and 128 Da, with fitted Gaussian models. (C) The heights of the three peaks increase with the glutaraldehyde concentration. Peaks that are not related to glutaraldehyde cross-linking (43 and 57 Da) do not exhibit such a trend. (D) Annotated MS/MS spectrum of a cross-link between two peptides of BSA. PepA marks peaks matching the total mass of the top peptide. Peaks annotated with *b or *y match the masses of the corresponding *b*- and *y*-fragments plus a 64 Da mass shift. *PepA marks a peak matching the total mass of the top peptide plus a 64 Da mass shift. The fragmentation pattern suggests that the lysine residue is the link site in the bottom peptide.

We next tested in turn each of the three masses (32, 64, and 128 Da) in a dedicated search application for cross-links[23] (Methods). For the 32 and 128 Da masses, the application identified zero and seven cross-links above the score threshold, respectively. All seven of the 128 Da cross-links comprise pairs of peptides that were consecutive in the protein sequences. Such cross-links hold little value for structural modeling. In contrast, for the 64 Da mass, the application identified 11 cross-links, of which only three were consecutive in their peptide sequences (Table S3). We concluded that the 64 Da adduct is the dominant chemistry of long-range glutaraldehyde cross-links, and proceeded to inspect its mass spectrometry characteristics. We found that during MS/MS fragmentation the 64 Da cross-link is completely cleaved (Figure 5D). In other words, fragment ions that included parts from both peptides together with the linker intact could not be observed in the MS/MS spectra. This is similar to the fragmentation behavior of formaldehyde that also exhibits complete cleavage.[23] However, unlike formaldehyde, the glutaraldehyde adduct breaks in such a way that the entire 64 Da moiety is carried off on either of the peptides.

In summary, it appears that glutaraldehyde is not a suitable reagent for XL-MS. We managed to identify very few glutaraldehyde cross-links, compared with the many tens and hundreds that can typically be identified with formaldehyde and DSS, respectively. This does not reflect the potential chemical reactivity of glutaraldehyde as a cross-linker, but rather the susceptibility of its cross-links to degrade during the sample preparation for mass spectrometry. We also note that the computational pipeline may show prominent peaks that eventually turn out to be artifacts. One should, therefore, run each peak value separately in a search application to sift out the incorrect ones.

■ CONCLUSIONS

The presented test cases outline two possible uses for the open search pipeline. In the first, the pipeline can be used to validate cross-linking chemistries that are presumably well understood. Even for these seemingly clear cases, unexpected chemistries can be detected with important implications for the ways in which the reagents can be used. In the second, the pipeline can be applied to reagents for which the cross-linking chemistry is largely unknown. As shown in the case of glutaraldehyde, the inferred linker masses are instructive toward a better understanding of the reaction. In any case, we highly recommend that any new cross-linking reagent is tested with the open search pipeline (see Code Availability). The run times are measured in minutes and therefore the user can obtain important insights into the chemistry with very little effort.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.0c03292.

Modifications on linear peptides from structured proteins cross-linked by glutaraldehyde (Figure S1) (PDF)

List of cross-links identified from EMCS cross-linking, searching 193.0739 Da as the linker mass (Table S1) (XLSX)

List of cross-links identified from EMCS cross-linking, searching 540.1600 Da as the linker mass (Table S2) (XLSX)

List of cross-links identified in glutaraldehyde cross-linking, searching 64.0313 Da as the linker mass (Table S3) (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Nir Kalisman − *Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem 9190401, Israel;* orcid.org/0000-0003-1615-7136; Email: nirka@mail.huji.ac.il

Authors

Moriya Slavin − *Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem 9190401, Israel*

Tamar Tayri-Wilk − *Institute of Life Sciences and Institute of Chemistry, The Hebrew University of Jerusalem, Jerusalem 9190401, Israel*

Hala Milhem − *Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem 9190401, Israel*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.0c03292

Author Contributions

Conceptualization, M.S. and N.K.; methodology, M.S., T.T.-W., and N.K.; investigation, M.S., T.T.-W., and H.M.; software, N.K.; writing, M.S. and N.K.; visualization, M.S.; supervision, N.K.; and funding acquisition, N.K.

Funding

This work was supported by the Israel Science Foundation grant number 1768/15.

Notes

The authors declare no competing financial interest.

The mass spectrometry data have been deposited to the ProteomeXchange Consortium via the PRIDE[28] partner repository with the data set identifier PXD020704.

■ REFERENCES

(1) Olsen, J. V.; Mann, M. *Mol. Cell. Proteomics* **2013**, 3444−3452.

(2) Chen, Y.; Chen, W.; Cobb, M. H.; Zhao, Y. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 761−766.

(3) Tanner, S.; Shu, H.; Frank, A.; Wang, L. C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. *Anal. Chem.* **2005**, *77*, 4626−4639.

(4) Hansen, B. T.; Davey, S. W.; Ham, A. J. L.; Liebler, D. C. *J. Proteome Res.* **2005**, *4*, 358−368.

(5) Frank, A.; Pevzner, P. *Anal. Chem.* **2005**, *77*, 964−973.

(6) Kim, S.; Na, S.; Sim, J. W.; Park, H.; Jeong, J.; Kim, H.; Seo, Y.; Seo, J.; Lee, K. J.; Paek, E. *Nucleic Acids Res.* **2006**, *34*, W258−W263.

(7) Na, S.; Bandeira, N.; Paek, E. *Mol. Cell. Proteomics* **2012**, *11*, No. M111.010199.

(8) Tyanova, S.; Temu, T.; Cox, J. *Nat. Protoc.* **2016**, *11*, 2301−2319.

(9) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. *Nat. Methods* **2017**, *14*, 513−520.

(10) Li, Q.; Shortreed, M. R.; Wenger, C. D.; Frey, B. L.; Schaffer, L. V.; Scalf, M.; Smith, L. M. *J. Proteome Res.* **2017**, *16*, 1383−1390.

(11) Chi, H.; Liu, C.; Yang, H.; Zeng, W. F.; Wu, L.; Zhou, W. J.; Wang, R. M.; Niu, X. N.; Ding, Y. H.; Zhang, Y.; Wang, Z. W.; Chen, Z. L.; Sun, R. X.; Liu, T.; Tan, G. M.; Dong, M. Q.; Xu, P.; Zhang, P. H.; He, S. M. *Nat. Biotechnol.* **2018**, *36*, 1059−1066.

(12) Herzog, F.; Kahraman, A.; Boehringer, D.; Mak, R.; Bracher, A.; Walzthoeni, T.; Leitner, A.; Beck, M.; Hartl, F. U.; Ban, N.; Malmström, L.; Aebersold, R. *Science* **2012**, *337*, 1348−1352.

(13) Rappsilber, J. *J. Struct. Biol.* **2011**, *173*, 530−540.

(14) Chen, Z. A.; Jawhari, A.; Fischer, L.; Buchen, C.; Tahir, S.; Kamenski, T.; Rasmussen, M.; Lariviere, L.; Bukowski-Wills, J. C.; Nilges, M.; Cramer, P.; Rappsilber, J. *EMBO J.* **2010**, *29*, 717−726.

(15) Müller, M. Q.; Dreiocker, F.; Ihling, C. H.; Schäfer, M.; Sinz, A. *Anal. Chem.* **2010**, *82*, 6958−6968.

(16) Yang, L.; Tang, X.; Weisbrod, C. R.; Munske, G. R.; Eng, J. K.; Von Haller, P. D.; Kaiser, N. K.; Bruce, J. E. *Anal. Chem.* **2010**, *82*, 3556−3566.

(17) Kao, A.; Chiu, C.; Vellucci, D.; Yang, Y.; Patel, V. R.; Guan, S.; Randall, A.; Baldi, P.; Rychnovsky, S. D.; Huang, L. *Mol. Cell. Proteomics* **2011**, *10*, No. M110.002212.

(18) Steigenberger, B.; Pieters, R. J.; Heck, A. J. R.; Scheltema, R. A. *ACS Cent. Sci.* **2019**, *5*, 1514−1522.

(19) Kalisman, N.; Adams, C. M.; Levitt, M. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 2884−2889.

(20) Lenz, S.; Giese, S. H.; Fischer, L.; Rappsilber, J. *J. Proteome Res.* **2018**, *17*, 3923−3931.

(21) Slavin, M.; Kalisman, N. *Methods Mol. Biol.* **2018**, *1764*, 173−183.

(22) Iacobucci, C.; Götze, M.; Ihling, C. H.; Piotrowski, C.; Arlt, C.; Schäfer, M.; Hage, C.; Schmidt, R.; Sinz, A. *Nat. Protoc.* **2018**, *13*, 2864−2889.

(23) Tayri-Wilk, T.; Slavin, M.; Zamel, J.; Blass, A.; Cohen, S.; Motzik, A.; Sun, X.; Shalev, D. E.; Ram, O.; Kalisman, N. *Nat. Commun.* **2020**, *11*, No. 3128.

(24) Iacobucci, C.; Sinz, A. *Anal. Chem.* **2017**, *89*, 7832−7835.

(25) Patiny, L.; Borel, A. *J. Chem. Inf. Model.* **2013**, *53*, 1223−1228.

(26) Boyatzis, A. E.; Bringans, S. D.; Piggott, M. J.; Duong, M. N.; Lipscombe, R. J.; Arthur, P. G. *J. Proteome Res.* **2017**, 2004−2015.

(27) Migneault, I.; Dartiguenave, C.; Bertrand, M. J.; Waldron, K. C. *BioTechniques* **2004**, 790−802.

(28) Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Pérez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yilmaz, Ş.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaíno, J. A. *Nucleic Acids Res.* **2019**, *47*, D442−D450.