In-Search Selection of Monoisotopic Peaks Improves the Identification of Cross-Linked Peptides

Swantje Lenz[†], Sven H. Giese[†], Lutz Fischer[‡], Juri Rappsilber[*,†,‡]

[†]Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany

[‡]Wellcome Centre for Cell Biology, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom

ABSTRACT: Cross-linking/mass spectrometry (CLMS) has gone through a maturation process similar to standard proteomics by adapting key methods such as false discovery rate control and quantification. However, methods for preprocessing mass spectrometric data of cross-linked peptides are currently underexplored. Preprocessing only requires computational efforts and can significantly improve data quality and search engine results. Our analysis shows that monoisotopic peak selection is a major weakness of current data handling approaches. High precursor masses paired with low intensities, typical of cross-linked peptides, are the main causes of the frequent misassignment of their monoisotopic peaks. We address this by 'in-search selection of the monoisotopic peak' in Xi (Xi-MPS). We compare and evaluate the performance of MaxQuant-Xi, OpenMS-Xi and Xi-MPS on three publicly available datasets. Xi-MPS always delivered the highest number of identifications with ~2 to 4-fold increase of PSMs without compromising identification accuracy as determined by FDR estimation and comparison to crystallographic models.

Several approaches have been utilized to increase the numbers of identified cross-links, e.g. enriching for cross-linked peptides[1-3], using different proteases[1,4] or optimizing fragmentation methods[5,6]. In parallel with experimental developments, data analysis has also progressed. Search software has been designed for the identification of cross-linked peptides, for example Kojak[7], xQuest[8], pLink[9], XlinkX[10] or Xi[11]. In addition, cross-linking workflows can make use of preprocessing methods to improve data quality and reduce file sizes, post-processing to filter out false identifications[7,12] and custom-tailored false discovery rate (FDR) estimation[13-15].

Preprocessing steps that improve the data quality for peptide identification can be split into those dealing with each of the two levels of acquisition: correction of the MS1 precursor ion m/z and simplifying MS2 fragment spectra. MS2 preprocessing, e.g. by deisotoping or removing less intense peaks, is search algorithm dependent due to distinct scoring methods on the fragment spectra[16]. Correction of the MS1 precursor ion can include correction of the charge state or m/z value. However, cross-linked peptides have characteristics that may make preprocessing methods used for linear peptide identification workflows nonoptimal: high-charge states, large masses, and low abundances.

Several cross-link search engines include preprocessing steps in their pipeline: In pLink, spectral quality filtering and preprocessing of the MS2 spectra is implemented by removing noise peaks. The group also published a tool for correction of monoisotopic peaks[17], and while it was not specifically designed for cross-link detection, it is implemented in their workflow[18]. The search engine Kojak averages precursor ion signals of neighboring scans to create a composite spectrum and infer the true monoisotopic mass of the precursor. Secondly, Kojak includes processing of MS2 spectra recommended for high-resolution spectra by collapsing isotope distributions and filtering for the most intense peaks in the fragment spectrum. Previously for searches with Xi, MaxQuant was used to perform the preprocessing. While not solely a preprocessing tool it performs a variety of preprocessing steps[19,20]. For example, it corrects the precursor m/z by an intensity-weighted average if a suitable peptide feature is found, reassigns the monoisotopic peak and contains options for intensity filtering of MS2 peaks. The precursor correction function within MaxQuant has been evaluated and shown to improve peptide mass accuracy[19]. However, we are not aware of a detailed evaluation of the impact of different preprocessing techniques on cross-link identification. Correction of the monoisotopic mass of the precursor, although acknowledged as an issue[7,21], awaits systematic evaluation.

In this study, we show that errors in assigning monoisotopic peaks during data acquisition adversely affect the identification of cross-links. Raw LC-MS data acquired from cross-linked peptide mixtures of varying complexities were preprocessed via MaxQuant and OpenMS[22,23]. We show that both software suites correct monoisotopic precursor masses with differing results. We then implement an option in Xi to consider multiple precursor masses to minimize the impact of

false monoisotopic precursor mass assignment on the identification of cross-links.

METHODS

Datasets. In this study we used three publicly available datasets (**Table 1**). The three datasets were chosen to reflect a range of applications of cross-linking mass spectrometry and a range of data complexity: the first dataset is Human Serum Albumin (HSA) cross-linked with succinimidyl 4,4-azipentanoate (SDA) and fragmented using five different methods (PXD003737)[24]. The second dataset is a pooled pseudo-complex sample with seven separately cross-linked proteins with bis(sulfosuccinimidyl) suberate (BS3) (PXD006131)[6]. This dataset includes data from four different fragmentation methods. The third dataset is the most complex sample, composed of 15 size exclusion chromatography fractions of *Chaetomium thermophilum* lysate cross-linked with BS3 and fragmented only with HCD (PXD006626)[25]. The first and last size exclusion fractions were used to optimize the search parameters for this dataset. All samples were analyzed on an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Fisher Scientific, San Jose, CA) using Xcalibur (version 2.0 and 2.1).

Table 1. Overview of datasets used.

Dataset	Sample	Database size[1]	Reference
1	HSA	1	Giese et al. 2016
2	Pseudo-complex	7	Kolbowski et al. 2017
3	*C. thermophilum*	198 to 400[2]	Kastritis et al. 2017

[1] Database size refers to the number of proteins in the database.
[2] Multiple size exclusion chromatography fractions (n = 15).

Preprocessing. Raw files were preprocessed independently using MaxQuant (1.5.5.30), OpenMS (2.0.1) and the ProteoWizard[26] tool msconvert (3.0.9576) for comparison. Scripts automating the preprocessing, search and evaluation were written in Python (2.7).

The essential steps during the preprocessing can be divided into two parts: (1) correction of the m/z or charge of the precursor peak for MS2 spectra and (2) denoising of MS2 spectra. MaxQuant and OpenMS both try to correct the precursor information via additional feature finding steps, i.e. identifying a peptide feature from the retention time, m/z and intensity domain of the LC-MS run. Additionally, denoising of the MS2 spectra is performed by simply filtering the most intense peaks in defined m/z windows. The preprocessing is by default enabled in MaxQuant and was run using the partial processing option (step 1 to 5) with default settings except for inactivated 'deisotoping' and 'top peaks per 100 Da' set to 20. The OpenMS preprocessing workflow includes centroiding, feature finding[27], precursor correction using the identified features and MS2 denoising as described above (**Figure S1**). Msconvert was used for converting the raw files to mgf files without any correction. These peak files were denoted as 'raw data' and used as our baseline to quantify the improvements in the subsequent database search. For the 'in-search selection of the monoisotopic peak' in Xi (Xi-MPS), we used msconvert to convert raw files to mgf files and included a MS2 peak filter for the 20 most intense peaks in a 100 m/z window.

Data Analysis. Peak files were searched separately in Xi (1.6.731) with the following settings: MS accuracy 3 ppm, MS/MS accuracy 10 ppm, oxidation of methionine as variable modification, tryptic digestion, 2 missed cleavages. For samples cross-linked with SDA, linkages sites were allowed on lysine, serine, tyrosine, threonine and protein n-terminus on one end and all amino acids on the other end of the cross-linker. Variable modifications were mono-link SDA (110.048 Da), SDA loop-links (82.0419 Da), SDA hydrolyzed (100.0524 Da), SDA oxidized (98.0368 Da)[24] as well as carbamidomethylation on cysteine. For searches with BS3, linkage sites were lysine, serine, threonine, tyrosine and the protein n-terminus. Carbamidomethylation on cysteine was set as fixed modification. Allowed variable modifications of the cross-linker were aminated BS3 (155.0946 Da), hydrolyzed BS3 (156.0786 Da) and loop-linked BS3 (138.0681 Da). For collision-induced dissociation (CID) and beam-type CID, also referred to as higher-energy C-trap dissociation (HCD), b- and y-ions were searched, whereas for electron transfer dissociation (ETD) c- and z-ions were allowed. For ETciD and EThcD, b-, c-, z- and y-ions were allowed. The HSA and pseudo-complex datasets were searched against the known proteins in the sample. For each protein fraction of the *C. thermophilum* dataset, the databases of the original publication were used, where for each fraction a database was created by taking the most abundant proteins (iBAQ value above 10^6). Datasets 1 and 2 were searched with a reversed decoy database, while dataset 3 was searched with a shuffled decoy database due to palindromic sequences.

For cross-linking, there are different information levels: PSMs, peptide pairs, residue pairs (links) and protein pairs. The false discovery rate (FDR) can be calculated on each one of these levels and should be reported for the level for which the information is given[13]. The FDR was calculated with xiFDR (1.0.14.34) and a 5% PSM level cutoff imposed. The setting 'uniquePSMs' was enabled and the FDR calculated separately on self and between links. Minimal peptide length was set to 6. In dataset 2, identified cross-linked residues were mapped to the crystal structure of the respective protein and the Euclidian distance between the alpha-carbons was calculated. Structures were downloaded from the PDB (IDs: 1AO6, 5GKN, 2CRK, 3NBS, 1OVT, 2FRJ).

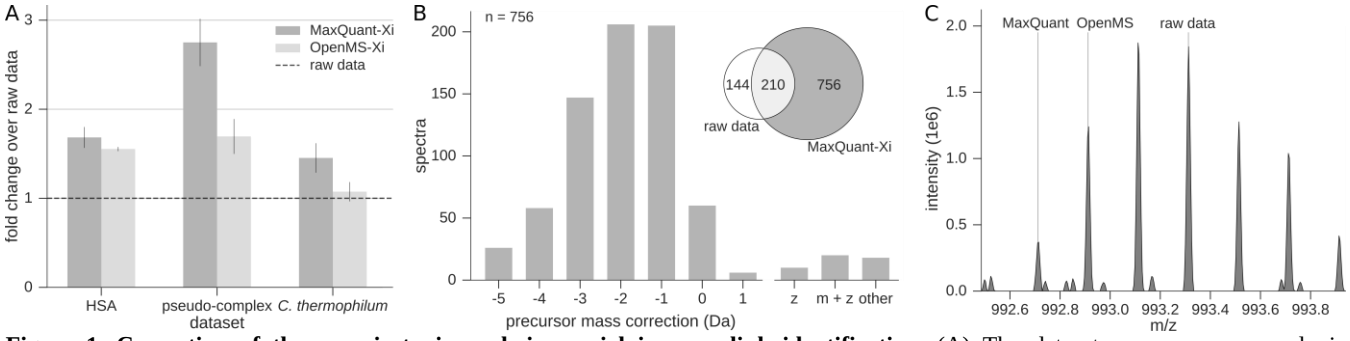

Figure 1. Correction of the monoisotopic peak is crucial in cross-link identification. (A) The datasets were preprocessed via MaxQuant and OpenMS, leading to more identified PSMs in all cases. Fold changes from raw data (msconvert conversion of Xcalibur data) were calculated for each file separately and the mean plotted. Error bars represent the standard error of the mean between different acquisitions (HSA: n = 3, pseudo-complex: n = 3, *C. thermophilum*: n = 8) (B) The majority of additional identifications after preprocessing are due to correction of the precursor mass to lighter monoisotopic masses. Spectra unique to MaxQuant preprocessed searches of HCD acquisitions of dataset 2 were evaluated regarding the correction of the precursor. The main proportion of the gain was corrected to lighter masses up to -3 Da, while charge state correction or correction to heavier masses did rarely occur. (C) Isotope cluster of a corrected precursor of m/z 992.71 (z = 5, m = 4958.6 Da) solely identified in MaxQuant preprocessed results. In OpenMS preprocessed and raw data, the wrong monoisotopic mass was selected for unknown reasons.

RESULTS AND DISCUSSION

We evaluated the impact on cross-link identification in Xi of changing the precursor monoisotopic mass initially assigned during data acquisition ('raw data'). In this analysis, MaxQuant and OpenMS were used as preprocessing tools. We made use of three different datasets that differ in complexity and fragmentation regimes. To measure the improvements from the preprocessing tools, a simple conversion from raw files to mgf format was done with msconvert and used as a baseline. Processed data were searched separately in Xi and evaluated on PSM or (unique residue pair) link level, with a 5% FDR. Finally, the newly implemented in-search selection of monoisotopic peaks in Xi was compared to the elaborate preprocessing pipelines in OpenMS and MaxQuant.

Preprocessing increases the number of cross-link PSMs by finding the correct monoisotopic peak. The datasets were preprocessed in MaxQuant and OpenMS and numbers of identified PSMs were compared to those obtained by using raw data. Dataset 1 and 2 were acquired with different acquisition methods so, for comparability to dataset 3, we will focus on the HCD acquired data in the following sections. Cross-links between proteins were excluded, either because they were experimentally not possible (dataset 2) or observed in too low numbers for reliable FDR calculation (dataset 3).

For raw data, 672, 354, and 2157 cross-link PSMs resulted for the HSA dataset, pseudo-complex dataset, and first and last fractions of *C. thermophilum* respectively. Both preprocessing approaches improved numbers of identified PSMs for all datasets: Preprocessing in MaxQuant led to 1127 (68% increase), 966 (173% increase) and 2966 (38% increase), while for OpenMS, 1044 (55% increase), 598 (69% increase) and 2394 (11% increase) PSMs were identified (**Figure 1A**).

We assessed the gains in identified PSMs of preprocessed data compared to raw data (dataset 2) regarding three forms of precursor correction: (1) correction of the monoisotopic mass, (2) charge state correction and (3) small corrections of the m/z value based on averaging the m/z values across the peptide feature (**Figure 1B**). Precursor mass and charge state of spectra identified solely in MaxQuant-Xi were compared to their counterparts when searching raw data in Xi. Of the 756 newly identified spectra, 686 (91%) had a different monoisotopic precursor mass. Precursors were primarily corrected to lighter masses by MaxQuant, i.e. the monoisotopic peak correction by -1 (208 spectra), -2 (215 spectra), -3 (149 spectra) and -4 Da (62 spectra). Greater shifts (-5 to -7 Da) only occurred 30 times, and corrections to heavier masses were observed 22 times. Only 30 spectra (4%) were corrected in their charge state. For the 60 spectra (8%) without correction in charge state or monoisotopic peak we identified only nine spectra had a higher error than 3 ppm before preprocessing indicating a small correction of the initial precursor m/z (by averaging of peptide feature peaks). The main proportion of these identifications is likely a product of noise removal in MS2 spectra or small changes in the score distribution. Similarly, for OpenMS-Xi, the correction of the monoisotopic peak had the greatest impact: Of the 314 spectra that OpenMS added over raw data, 139 were precursor corrected by -1 Da and 108 to -2 Da. In contrast to MaxQuant, corrections to -3 or lighter were not observed, which might explain the higher number of identifications obtained with MaxQuant-Xi.

In summary, preprocessing, especially monoisotopic peak correction, leads to a notable increase in identifications. Using the 3-dimensional peptide feature is advantageous compared to the on-the-fly detection of the monoisotopic peak. If the preceding MS1 spectrum was acquired during the beginning (or end) of the elution profile of a peptide the intensity will be low. Thus, the monoisotopic peak might not even be detectable at the time of fragmentation. For large (cross-linked) peptides this effect might be exasperated by the monoisotopic peak usually being less intense than other

isotope peaks. Therefore, using the additional information from the retention time domain will be beneficial. The same feature information can also be used to determine or validate the assigned charge state of the precursor. However, the instrument software almost always assigned the same charge state as MaxQuant or OpenMS. Thus, the main advantage for the identification of cross-linked peptides arises from monoisotopic peak correction.

Interestingly, OpenMS and MaxQuant did not always agree on or find the same monoisotopic peak (**Figure 1C**). Of the total MaxQuant corrected spectra with a different monoisotopic mass, 81% were not corrected and 6% differently corrected by OpenMS. Vice versa, 15% of the monoisotopic peaks corrected by OpenMS were not corrected by MaxQuant and 25% were corrected differently. Both MaxQuant and OpenMS have their own implementations for precursor correction - therefore there might be instances where MaxQuant is able to find a corresponding peptide feature where OpenMS is not and vice versa. Although OpenMS did not lead to the same improvements in the number of identifications as MaxQuant, it did correct some precursors which the latter did not. We therefore suspect that there are also precursors with a false monoisotopic peak assigned that were corrected with neither algorithm. Furthermore, the 3-dimensional detection of peptide features is challenging to do for low intense peptides. In conclusion, there likely remain falsely assigned monoisotopic peaks in the data, ultimately leading to missed or false identifications.

In-search monoisotopic peak selection increases the number of identifications. We observed multiple cases where MaxQuant and OpenMS disagreed in their monoisotopic peak choice indicating that the problem of monoisotopic peak selection (MPS) cannot be solved easily at MS1 level. Indeed, we found instances where the monoisotopic peak is simply not distinguishable from noise, so a feature-based correction would not be feasible. Nevertheless, the associated MS2 spectra could be matched to a cross-linked peptide when considering multiple different monoisotopic masses during search. This shows that the extra information of obtaining a peptide-spectrum match assists MPS over considering MS1 information alone. Therefore, we implemented a monoisotopic peak selection in Xi: for each MS2 spectrum, multiple precursor masses are considered during a single search and the highest scoring peptide-pair assigns the precursor mass. Note that this is different from simply searching with a wide mass error for MS1. The mass accuracy of MS1 is minimally compromised as multiple candidates for the monoisotopic mass are taken and considered with the original mass accuracy of the measurement.

To find a good trade-off between increased search space and sensitivity we tested different mass range settings on the datasets. For dataset 2 (HCD subset), the number of PSMs increased using ranges of up to -5 Da on the considered monoisotopic masses **(Figure 2A)**. However, the increase in identifications from -4 to -5 Da is only 3% and considering the

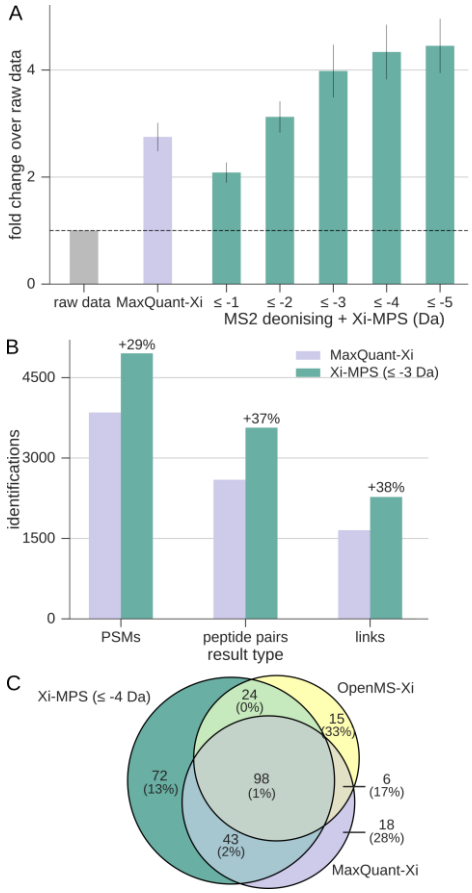

Figure 2. In-search monoisotopic peak selection outperforms preprocessing. (A) Performance of Xi-MPS on dataset 2. HCD data of the pseudo-complex dataset were searched assuming different ranges of missing monoisotopic peaks. With increased ranges, the number of identified PSMs also increases. (B) Performance of Xi-MPS on the complete *C. thermophilum* dataset. All 15 fractions were searched with the original preprocessed data as well as with Xi-MPS. (C) Overlap of identified residue pairs of MaxQuant-Xi and OpenMS-Xi to residue pairs gained from Xi-MPS (dataset 2). Numbers in brackets are the proportion of decoys in the respective regions.

increase in search time, we continued with a maximal correction to -4 Da as the optimal setting for this dataset. Xi-MPS yielded 1508 PSMs, which is a 326% increase compared to searching raw data and a 56% increase compared to MaxQuant-Xi. Similar improvements are observed for the other fragmentation methods in this dataset (**Figure S2**). Additionally, we tested up to -7 Da to test if a large increase in search space increases random spectra matches as measured by the target-decoy approach. The number of identifications at 5% FDR decreased only slightly compared to -5 Da (-1%), but still leads to more identifications than up to -4 Da (3%). In the HSA dataset, Xi-MPS with up to -4 Da increased the number of identified PSMs by 170% compared to raw data (**Figure S3**).

As a final evaluation of in-search monoisotopic peak selection, we searched the complete dataset of *C. thermophilum*. We used 0 to -3 Da as the range of Xi-MPS,

since an initial analysis of the first and last fraction of the *C. thermophilum* dataset returned a similar number of identifications when running Xi-MPS up to -4 Da or -3 Da (**Figure S4**). As a comparison we took the original peak files obtained from PRIDE. The FDR was calculated separately on self and between links, with a minimum of 3 fragments per peptide and a minimal delta score of 0.5. For the original peak files, which were preprocessed in MaxQuant, we identified 3848 PSMs, 2594 peptide pairs and 1653 cross-links, with a 5% FDR on each respective level **(Figure 2B)**. Xi-MPS resulted in 4952 PSMs (29% increase), 3566 peptide pairs (37% increase) and 2273 cross-links (38% increase).

We looked next into the complementarity of search results by the different approaches, using dataset 2 at 5% link-FDR. Preprocessing via MaxQuant and OpenMS led to 172 and 158 links, respectively, while Xi-MPS resulted in 243 links. While the overlap between links of OpenMS-Xi and MaxQuant-Xi is only 50%, Xi-MPS identifications cover 76% of both searches (Figure 2C). 19 and 23 links are uniquely found in MaxQuant and OpenMS preprocessed data respectively. However, there are 5 decoy links as well in each unique set (resulting in a link-FDR of 26% and 22%). For Xi-MPS, there are 75 unique target links with 12% link-FDR.

Identification based correction employed by Xi-MPS results in more identifications than the feature-based correction algorithms of OpenMS and MaxQuant. Neither OpenMS nor MaxQuant correct all precursor masses that are incorrectly assigned during data acquisition. In Xi-MPS, spectra are searched with multiple monoisotopic masses, thereby less relying on the MS1 information. The quality of the precursor isotope cluster does not play a role in the decision of the monoisotopic mass and spectra for which correction failed will be identifiable now. One could hypothesize that increasing the search space by considering multiple masses will lead to more false positives, thereby reducing the number of true identifications. This is not the case, as we match substantially more PSMs at constant FDR by considering alternative monoisotopic masses. As a second plausible caveat, this approach increases the search time. However, the use of relatively cheap computational time appears balanced by the notable increase in identified cross-links. The optimal range of additional monoisotopic peaks to search will however be dependent on complexity and quality of MS1 acquisition and plausibly also on the version of the instrument software. To reduce the mass range considered in Xi-MPS, we developed a MS1 level-based approach. For each precursor, we search lighter isotope peaks in MS1 and use this to narrow the search space (explained in detail in Supporting Information). This led to an average of 24% less values to be considered, while only reducing the number of identifications by 3%. We hope that our observation of the monoisotopic peak detection challenge in cross-linking together with our publicly available datasets will lead to further improvements in monoisotopic peak-assignment algorithms in the future, possibly tailored to cross-link data.

Note that the cross-link search engine Kojak incorporates monoisotopic peak prediction in its preprocessing workflow for which they reported 49% of the spectra to be reassigned with another mass, mentioning up to -2 Da but no evaluation metric[7]. We showed that considering a shift of the precursor mass of up to -4 Da during database search can substantially increase the number of identifications.

In-search monoisotopic peak selection does not compromise search accuracy. Changing the search could lead to several problems. We excluded already that the increased search space leads to high-scoring decoy matches that in turn reduce the number of identifications at a given FDR cut-off. However, it is possible that the target-decoy approach itself fails for unknown reasons. We therefore assessed our results against known PDB structures using the HCD data of the pseudo-complex dataset (dataset 2), at 5% link-FDR. Assuming a crystal structure is correct, a cross-link can be overly long either because the link is false or because of in-solution structural dynamics. If, however, the proportion of long-distance links in two result files is identical, at least the two files have equal quality.

We first tested the results of all three approaches against crystal structures. Residue pairs were mapped to PDB structures and the distance between the two alpha-carbons was calculated (see methods). 30 Å was set as the maximal distance for BS3, links with a greater distance were classified as long-distance. In this evaluation we excluded the protein C3B because its flexible regions make it unsuitable for this analysis. For MaxQuant and OpenMS preprocessed results, 11.8% and 6.1% long-distance cross-links were identified, respectively. In Xi-MPS, 8.1% long distance cross-links were identified (**Figure 3A**). Of the uniquely identified links through Xi-MPS, only 5.3% were long distance links. Therefore, Xi-MPS as such does not lead to an enrichment in long-distance cross-links. However, it could be that mass corrected precursors tend to have a higher proportion of long-distance links. We therefore split the Xi-MPS results into five groups corresponding to the monoisotopic mass change (0, -1, -2, -3, -4 Da) and looked at their match to crystal structures. If a link originated from PSMs with different mass corrections, all of those were considered. We conducted a 'nonparametric ANOVA' (Kruskal–Wallis test) to detect any significant changes in the distance distributions of Xi-MPS identifications with different shifts and the decoy distribution. However, we fail to reject the null hypothesis at the predetermined significance level of $\alpha = 0.05$ (p-value: 0.13), indicating that the distance distributions for all subsets are similar. This matches the visual inspection of the distance distributions (**Figure 3B**). Furthermore, all individual distance distributions were significantly smaller than the derived reference distribution (one-sided Wilcoxon test, see **Table S5**). In conclusion, we fail to see any evidence of in-search monoisotopic mass selection leading to increased conflicts with crystal structures. We evaluated next the effect of in-search monoisotopic mass selection on PSM quality as

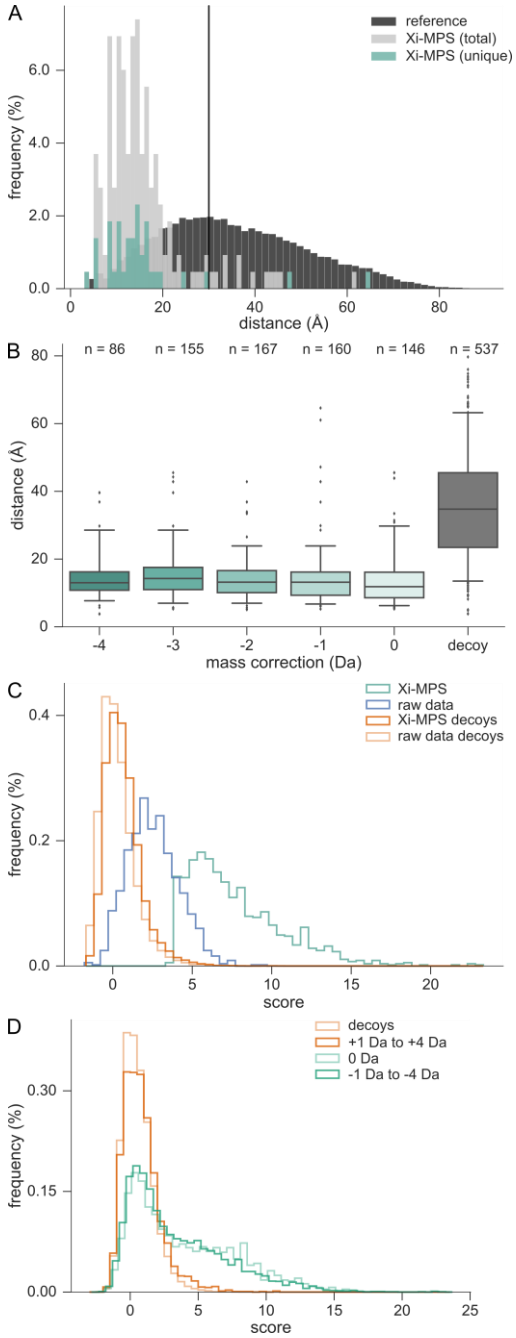

Figure 3. Matches with and without in-search mass shift show similar quality metrics. (A) Evaluation of Xi-MPS derived links on crystal structures (dataset 2). Distances between alpha carbon atoms of identified cross-linked residues in the crystal structure of the proteins are shown in light grey while a reference distribution of all possible pairwise C-alpha distances of cross-linkable residues is shown in dark grey. 30 Å is set as a limit, above which links are defined as long distance. (B) Distance distribution of identifications with different mass corrections. No significant difference could be calculated between the different mass shifts, while all had a significant difference to the decoy distribution. (C) PSM scores of spectra identified with a mass shift are significantly higher than the corresponding score in raw data. Shown are the score distributions of raw and Xi-MPS results, as well as the corresponding distribution of the decoys. (D) Score distribution of PSM matches of the 'decoy mass search'. Identifications with a positive mass shift follow the decoy distribution while identifications with a negative shift resemble unshifted identifications. The scores of negative shifted PSMs are significantly higher than those of positive shifted PSMs (one-sided Wilcoxon test, p-value < 2.2e-16).

assessed by the search score. First, we compared the scores of PSMs with a mass shift (Xi-MPS identifications) with the scores of the same spectrum without a mass shift (raw data). While scores with shifted mass have a median of 6.7, the median score is 2.3 when using the raw masses (**Figure 3C**). As one would expect from an increased search space, also the scores of decoy hits improve, albeit only marginally. We find that the score difference of target PSMs is significantly larger than of decoy PSMs (one-sided Wilcoxon test, p-value: < 2.2e-16). We then turned to a 'decoy mass search' for which we not only searched the range from 0 Da to -4 Da, but also +1 Da to +4 Da. Assuming the monoisotopic peak in the raw data is rarely lighter than the true monoisotopic peak, the new identifications should score like decoy identifications. Indeed, the resulting score distributions for targets with a positive mass shift follow the decoy distribution (**Figure 3D**). In contrast, the identifications with a negative shift are distributed like those identifications without mass shift. In conclusion, in-search monoisotopic mass change leads to significantly improved scores with a distribution that resembles that of precursors that did not see a mass change (0 Da). Importantly, these improvements are not random since an equally large search space increase (+1 Da to +4 Da) results in a completely different score distribution that resembles the decoy distribution but not the distribution of identifications without a mass shift.

Heavy and low intense peptides are corrected more frequently. One would expect shifted mass assignment especially for peptides of high mass and those of low abundance. Possibly these criteria would even allow to further narrow the search space. For large peptides (approximately >2000 Da), the monoisotopic peak will not be the most intense peak in the isotope cluster. If the peptide is low abundant, the monoisotopic peak may be too low intense to be detected. We therefore analyzed the monoisotopic peak selection in Xi-MPS regarding the precursor mass and intensity. Indeed, precursors with higher masses are more often corrected to lighter monoisotopic peaks (**Figure 4A**). While the median precursor mass for uncorrected matches is 2952 Da, for matches corrected by -2 Da it is already 4062 Da and for -4 Da 4684 Da. Of the identifications with a mass higher than 3000 Da, 88% were identified with a lighter mass. For precursors lighter than 3000 Da, the proportion is 42%. Like mass dependency, there is a trend to larger correction ranges for lower intense peptides (**Figure S5**).

When evaluating the newly matched precursors of Xi-MPS, the advantage of not relying on MS1 identification is evident. Matches not made through any of the preprocessing methods

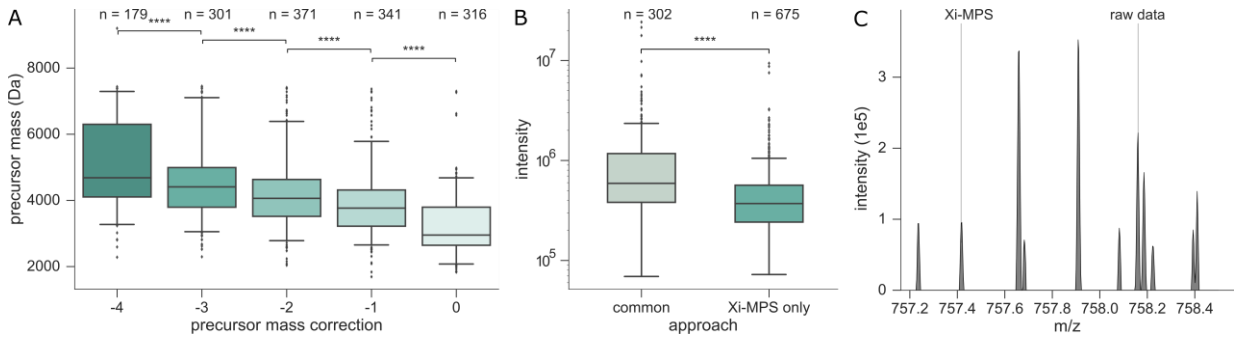

Figure 4. Correction is dependent of precursor mass and intensity. (A) Box plot of the precursor masses and their monoisotopic mass correction of identified PSMs after Xi-MPS. Higher masses more often require monoisotopic mass correction to lighter masses. Whiskers show the 5 and 95% quantiles of the data. Asterisks denote the significance calculated by a one tailed t-test (****: p-value < 0.0001). (B) Precursors of cross-links identified in all three approaches, MaxQuant-Xi, OpenMS-Xi, and Xi-MPS ('common'), are more intense than precursors of cross-links identified only in Xi-MPS. In other words, successful correction happens more often for abundant precursors, while Xi-MPS identifies less intense precursors. (C) MS1 isotope cluster of a cross-linked peptide. The monoisotopic peak of m/z 758.16 (z = 4, m = 3028.6 Da) was falsely assigned during acquisition and not corrected in any preprocessing approach. Xi-MPS identifies a PSM for a precursor with a mass 3 Da lighter.

are generally much less intense (**Figure 4B**) and larger (**Figure S6**) than matches in common for all approaches. Attempts to narrow the search space using precursor mass and intensity failed in our hands, unfortunately (data not shown).

Manual analysis of isotope clusters of corrected precursors of dataset 2 revealed many cases where the monoisotopic peak is present in the MS1 spectrum but was not recognized during acquisition. For some, this might be due to the peak being of low intensity and discarded as noise, or because of other interfering peaks (**Figure 4C**). However, there are also cases where the cluster is well resolved (**Figure 1C**). Without details of how the instrument software determines the monoisotopic peak a full evaluation is difficult. For a complete list of precursor m/z of Xi-MPS identifications and corresponding m/z of raw, MaxQuant and OpenMS data see **Table S1-S3**.

Note that in many acquisition methods, the machine only fragments peaks where it can successfully identify a full isotope cluster. Therefore, there might be instances of cross-linked peptides not being fragmented because of insufficient isotopic cluster quality, leading to lost identifications.

CONCLUSION

The size and low abundance of cross-linked peptides leads to frequent misassignment of the monoisotopic mass by instrument software, but also sophisticated correction approaches employed by MaxQuant and OpenMS. Considering multiple monoisotopic masses during search increases the number of cross-link PSMs 1.8 to 4.2-fold. The problem of wrongly assigned monoisotopic peaks will have an impact on most cross-link search engines since all rely in some part on the precursor mass. Even with improved acquisition or correction software, there will be instances where the monoisotopic peak cannot be determined correctly before search due to low intensity. Our search-assisted monoisotopic peak selection provides a general solution to this problem.

ASSOCIATED CONTENT

Supporting Information

Precursor m/z of different processing methods (XLS).
MS1 based mass range reduction, OpenMS preprocessing workflow, performance of Xi-MPS on additional data, dependency of correction on precursor intensity (PDF).

AUTHOR INFORMATION

Corresponding Author

*** E-mail: juri.rappsilber@tu-berlin.de**

Notes

The Authors declare no competing financial interest.

ACKNOWLEDGMENT

We thank Dr. Francis O'Reilly for comments and helpful discussions. This work was supported by the Einstein Foundation, the DFG [RA 2365/4-1], and the Wellcome Trust through a Senior Research Fellowship to JR [103139] and a multi-user equipment grant [108504]. The Wellcome Centre for Cell Biology is supported by core funding from the Wellcome Trust [203149].

REFERENCES

(1) Leitner, A.; Reischl, R.; Walzthoeni, T.; Herzog, F.; Bohn, S.; Förster, F.; Aebersold, R. *Mol. Cell. Proteomics* **2012**, *11*, M111.014126.

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

(3) Tan, D.; Li, Q.; Zhang, M.-J.; Liu, C.; Ma, C.; Zhang, P.; Ding, Y.-H.; Fan, S.-B.; Tao, L.; Yang, B.; Li, X.; Ma, S.; Liu, J.; Feng, B.; Liu, X.; Wang, H.-W.; He, S.-M.; Gao, N.; Ye, K.; Dong, M.-Q.; Lei, X. *Elife* **2016**, *5*.

(4) Belsom, A.; Schneider, M.; Fischer, L.; Mabrouk, M.; Stahl, K.; Brock, O.; Rappsilber, J. Wellcome Open Research 2016, 24.

(5) Liu, F.; Lössl, P.; Scheltema, R.; Viner, R.; Heck, A. J. R. *Nat. Commun.* **2017**, *8*, 15473.

(6) Kolbowski, L.; Mendes, M. L.; Rappsilber, J. *Anal. Chem.* **2017**, *89*, 5311–5318.

(7) Hoopmann, M. R.; Zelter, A.; Johnson, R. S.; Riffle, M.; MacCoss, M. J.; Davis, T. N.; Moritz, R. L. *J. Proteome Res.* **2015**, *14*, 2190–2198.

(8) Leitner, A.; Walzthoeni, T.; Aebersold, R. *Nat. Protoc.* **2013**, *9*, 120–137.

(9) 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. *Nat. Methods* **2012**, *9*, 904–906.

(10) Liu, F.; Rijkers, D. T. S.; Post, H.; Heck, A. J. R. *Nat. Methods* **2015**, *12*, 1179–1184.

(11) Fischer, L. *https://github.com/Rappsilber-Laboratory/XiSearch*, Accessed 23.09.17.

(12) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. *Nat. Methods* **2007**, *4*, 923–925.

(13) Fischer, L.; Rappsilber, J. *Anal. Chem.* **2017**, *89*, 3829–3833.

(14) Walzthoeni, T.; Claassen, M.; Leitner, A.; Herzog, F.; Bohn, S.; Förster, F.; Beck, M.; Aebersold, R. *Nat. Methods* **2012**, 9, 901–903.

(15) Maiolica, A.; Cittaro, D.; Borsotti, D.; Sennels, L.; Ciferri, C.; Tarricone, C.; Musacchio, A.; Rappsilber, J. *Mol. Cell. Proteomics* **2007**, 6, 2200–2211.

(16) Renard, B. Y.; Kirchner, M.; Monigatti, F.; Ivanov, A. R.; Rappsilber, J.; Winter, D.; Steen, J. A. J.; Hamprecht, F. A.; Steen, H. *Proteomics* **2009**, *9* (21), 4978–4984.

(17) Yuan, Z.-F.; Liu, C.; Wang, H.-P.; Sun, R.-X.; Fu, Y.; Zhang, J.-F.; Wang, L.-H.; Chi, H.; Li, Y.; Xiu, L.-Y.; Wang, W.-P.; He, S.-M. *Proteomics* **2012**, *12* (2), 226–235.

(18) Fan, S.-B.; Meng, J.-M.; Lu, S.; Zhang, K.; Yang, H.; Chi, H.; Sun, R.-X.; Dong, M.-Q.; He, S.-M. *Curr. Protoc. Bioinformatics* **2015**, *49*, 8.21.1–19.

(19) Cox, J.; Mann, M. *Nat. Biotechnol.* **2008**, *26*, 1367–1372.

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

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

(22) Sturm, M.; Bertsch, A.; Gröpl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; Kohlbacher, O. *BMC Bioinformatics* **2008**, *9*, 163.

(23) Röst, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmström, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O. *Nat. Methods* **2016**, *13*, 741–748.

(24) Giese, S. H.; Belsom, A.; Rappsilber, J. *Anal. Chem.* **2016**, *88*, 8239–8247.

(25) Kastritis, P. L.; O'Reilly, F. J.; Bock, T.; Li, Y.; Rogon, M. Z.; Buczak, K.; Romanov, N.; Betts, M. J.; Bui, K. H.; Hagen, W. J.; Hennrich, M. L.; Mackmull, M.-T.; Rappsilber, J.; Russell, R. B.; Bork, P.; Beck, M.; Gavin, A.-C. *Mol. Syst. Biol.* **2017**, *13*, 936.

(26) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. *Bioinformatics* **2008**, *24*, 2534–2536.

(27) Weisser, H.; Nahnsen, S.; Grossmann, J.; Nilse, L.; Quandt, A.; *Brauer*, H.; Sturm, M.; Kenar, E.; Kohlbacher, O.; Aebersold, R.; Malmström, L. *J. Proteome Res.* **2013**, *12*, 1628–1644.
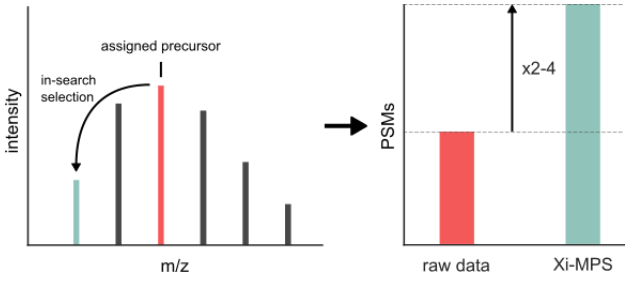
Table of Contents graphic