Mimicked synthetic ribosomal protein complex for benchmarking crosslinking mass spectrometry workflows

Manuel Matzinger[1,*,§], Adrian Vasiu[1,*], Mathias Madalinski[1], Florian Stanek[1], Karl Mechtler[1,2,§]

[1]Institute of Molecular Pathology (IMP), Vienna BioCenter (VBC), Vienna, Austria

[2]Institute of Molecular Biotechnology, Austrian Academy of Sciences, Vienna BioCenter (VBC), Vienna, Austria

* These authors contributed equally

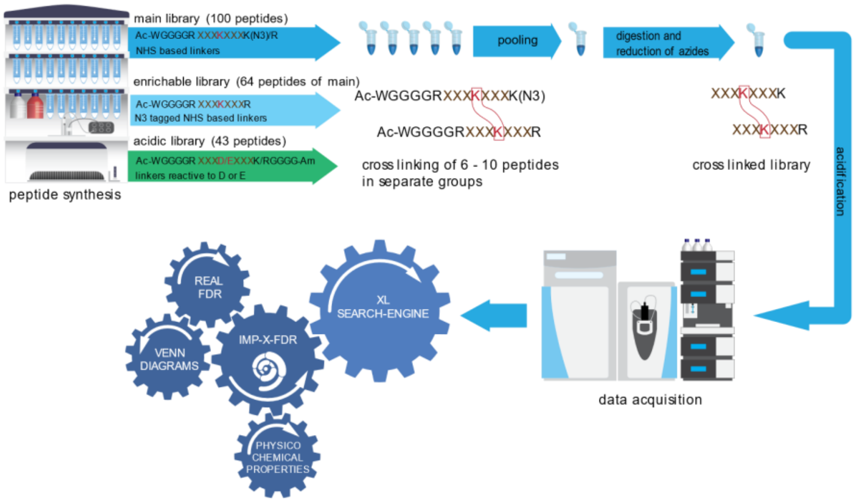
§Correspondence to: manuel.matzinger@imp.ac.at, karl.mechtler@imp.ac.at

ABSTRACT:

The field of cross-linking mass spectrometry has matured to a frequently used tool for the investigation of protein structures as well as interactome studies up to a system wide level. The growing community generated a broad spectrum of applications, linker types, acquisition strategies and specialized data analysis tools, which makes it challenging, especially for newcomers, to decide for an appropriate analysis workflow. Therefore, we here present a large and flexible synthetic peptide library as reliable instrument to benchmark crosslinkers with different reactive sites as well as acquisition techniques and data analysis algorithms. Additionally, we provide a tool, IMP-X-FDR, that calculates the real FDR, compares results across search engine platforms and analyses crosslink properties in an automated manner. The library was used with the reagents DSSO, DSBU, CDI, ADH, DHSO and azide-a-DSBSO and data were analysed using the algorithms MeroX, MS Annika, XlinkX, pLink and MaxLynx. We thereby show that the correct algorithm and search setting choice is highly important to improve ID rate and FDR in combination with software and sample-complexity specific score cut-offs. When analysing DSSO data with MS Annika, we reach high identification rates of up to ~70 % of the theoretical maximum (i.e. 700 unique lysine-lysine cross-links) while maintaining a low real FDR of < 3 % at cross-link level and with extraordinary high reproducibility, representatively showing that our test system delivers valuable and statistically solid results.

Keywords: Crosslinking, mass spectrometry, synthetic peptide library, FDR control

Graphical abstract:


main library (100 peptides)
Ac-WGGGGR XXXKXXXK(N3)/R
NHS based linkers
enrichable library (64 peptides of main)
Ac-WGGGGR XXXKXXXR
N3 tagged NHS based linkers
acidic library (43 peptides)
Ac-WGGGGR XXXD/EXXXK/RGGGG-Am
linkers reactive to D or E
peptide synthesis
pooling
digestion and reduction of azides
Ac-WGGGGRXXXKXXXK(N3)
Ac-WGGGGRXXXKXXXR
cross linking of 6 - 10 peptides in separate groups
XXXKXXXK
XXXKXXXR
cross linked library
acidification
data acquisition
XL SEARCH-ENGINE
REAL FDR
IMP-X-FDR
VENN DIAGRAMS
PHYSICO CHEMICAL PROPERTIES

INTRODUCTION

The field of cross-linking mass spectrometry has matured and now represents a frequently used technique for the investigation of protein structures as well as to freeze (transient) protein-protein interactions and uncover whole interactomes on a system wide level. Numerous reviews already summarized successful applications but also limitations of this technique.[1–4] The growing community also participated in the generation of a wide variety of cross-linker reagents bearing chemical reactivities mainly towards lysine (e.g. via N-Hydroxysuccinimide esters[5,6]) but also towards acidic amino acids (e.g. by amide formation[7] or hydrazines[8]), cysteine (e.g. via maleimides[9,10]) or even without any specificity (e.g. via diazirine groups[11]). With a focus on proteome wide studies and *in vivo* cross linking, MS-cleavable linkers, like DSSO[12] or DSBU[13], are facilitating data analysis by generating characteristic doublet ions and became commonly used. Aiming to dig deeper in the interactome of complex samples, reagents bearing an affinity tag for selective enrichment of cross-linked peptides were further developed.[14–16] The optimization of cross-linker specific acquisition strategies[17] and most recently the implementation of ion-mobility[18,19] or FAIMS filtering[20] as additional separation technique further boosted the number of possible crosslink (XL) identifications.

The broad spectrum of applications, linker types and acquisition strategies[4] led to the development of lots of specialized data analysis tools[21] which makes it challenging, especially for newcomers, to decide for an appropriate analysis workflow.

Therefore, a synthetic peptide library as previously published by our group[22] is a valuable tool for standardization and can be used as a basis to decide for the optimal analysis tool in dependency of the used crosslinker and acquisition strategy. This peptide library was based on 95 synthetic peptides of the protein Cas9.

In this study we present a significantly improved and extended peptide library that now contains a total of 141 peptides from 38 different proteins of the *E. coli* ribosomal complex. This enables finding inter- and intra-protein cross-links in our results. Furthermore, the number of theoretical correct cross-link combinations is increased from 426 in the previously published version to up to 1018 in this library. In conclusion a more reliable and, if supported by the data analysis tool, separate inter/intra false discovery rate (FDR) calculation can be performed. In contrast to our previously published library system of Cas9, the peptides were now combined to 3 different libraries designed to be compatible not only with lysine but also with aspartic- and glutamic-acid reactive cross-linkers as well as for crosslinkers bearing an azide as affinity tag, respectively.

With the here reported peptide library, we mimic a real protein complex and a system that is appropriate to find optimal settings for real biological samples as well as to benchmark different crosslinker types and data analysis tools.

To increase the usability of that library, we additionally created a tool, IMP-X-FDR, that is capable to calculate the real FDR, scale the number of cross-link IDs to the same real FDR of 1 or 5% by applying a score-cutoff as well as to compare the results obtained from several search engines or cross-linkers in Venn diagrams. IMP-X-FDR completes this task in an automated manner and includes an easy-to-use graphical user interface, which broadens the potential user group. IMP-X-FDR is free to use and can be downloaded from Github (https://github.com/fstanek/imp-x-fdr).

RESULTS

We synthesized 141 peptides based on sequences from 38 proteins of the *E. coli* ribosomal complex (Supplementary Table 1). They are designed to contain exactly one crosslink-able position. Peptides are grouped to 6 -10 peptides and XLed groupwise. After that all groups are pooled to obtain the XLed library were links between peptides of different groups or to not synthesized peptides are known false positives. The main library consists of 100 peptides containing exactly one XL-able lysin residue. All peptides start with the sequence WGGGGR- and their n-termini are protected by an acetate group to hinder any XL reaction at this position. Tryptophan thereby facilitates photometric quantitation of peptides after synthesis. C-terminal lysin residues are modified to an azide (instead of an amine) to again block XL reaction. During sample processing the protected n-terminal sequence part is removed by tryptic digestion and azide modified lysines are reduced to amines yielding ordinary tryptic peptides with a known XL position for MS/MS analysis. We additionally compiled a library not containing any azide protected lysin residue but instead exclusively those 64 peptides of the main library ending with arginine. This "enrichable library" is compatible with azide-based affinity enrichment as done with the reagent azide-tagged acid-cleavable disuccinimidylbissulfoxide, (DSBSO). Finally, a third library, made from 43 peptides, is designed to contain exactly one reactive aspartic-acid or glutamic-acid for use with crosslinker reagents reactive to carboxylic acids. In this "acidic library" the c-terminal peptide part is amide protected and all sequences end as GGGG after a K or R which will again release ordinary tryptic peptides after digestion.

Benchmarking XL search engines with linkers targeting lysin.

To benchmark commonly used XL search-algorithms we applied the MS cleavable linker reagents disuccinimidyl sulfoxide (DSSO), ureido-4,4′-dibutyric acid bis(hydroxysuccinimide) ester (DSBU) and 1,1'-carbonyldiimidazole (CDI) to the main library (Supplemental Table 2). As representatively shown on the data generated with DSSO the benchmarked search engines all output higher real FDRs than the estimated 1% on XL level (Figure 1A). For this dataset MS Annika[23] and MaxLynx[24] perform best, both by means of correct FDR estimation as well as by means of unique ID numbers. We additionally applied post-score-cutoffs to correct the real FDR to ≤ 1%. The obtained results are in line with minimal scores recommended by the software developers (i.e. scores >100 are considered good for MeroX[25], 50 is default and 75 seems reliable from our data; 40 is default for XlinkX[26], 41 seems reliable from our data). Although using a score-cutoff is an effective strategy to correct for acceptable FDR, our data also shows, that built in (usually target-decoy based) FDR estimations are not sufficient yet. Especially when using pLink[27], we had the impression that (score-based) separation of correct and incorrect IDs does not work properly meaning that the majority of XL IDs is lost upon applying our FDR correction. Of note, pLink was initially designed to work with non-MS-cleavable linker reagents and is not optimized for HCD data, which might explain its weak performance in this dataset compared to all other tested algorithms.

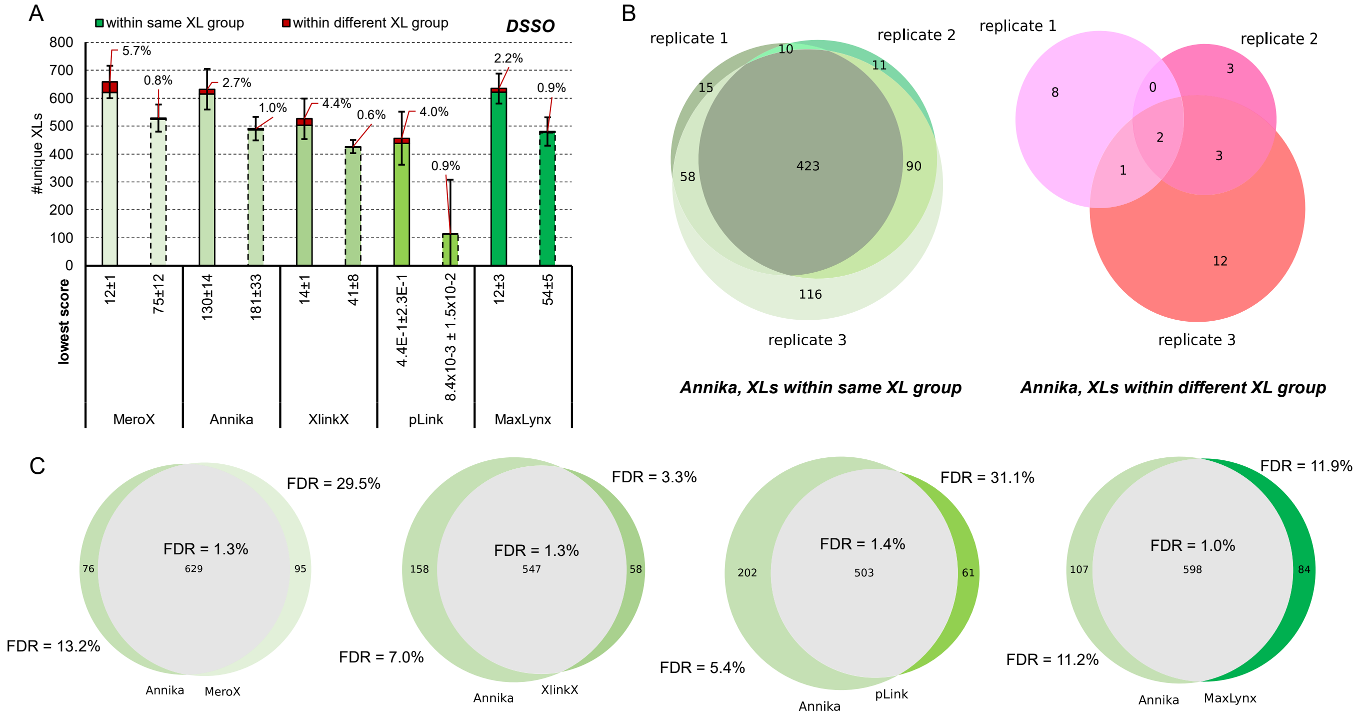


Figure 1: ***Benchmarking of data analysis tools on the example of DSSO.*** *(**A**) Average XL ID numbers using DSSO applied to the main library and the algorithms MeroX*[25]*, Annika*[23]*, XlinkX*[26]*, pLink2*[27] *or MaxLynx*[24] *for analysis. All results were obtained at 1% estimated FDR and corrected by applying a post-score cutoff to reach a real FDR<=1. The real FDR is shown as callout, error-bars indicate standard deviations, n=3 (**B**) Overlap of XLs within same XL group (considered correct) or within different groups (considered incorrect) identified in each replicate using Annika (**C**) Overlap of cross links identified in replicate 3 after analysis using Annika or an alternative algorithm as given. Real FDRs for commonly found and exclusively found links are given.*

Instead of using score cutoff values, the comparison of identified crosslinks with more than one search engine can significantly contribute to improve the confidence in results (Figure 1C). Using our in house-developed tool IMP-X-FDR we visualized the overlap of search results obtained from MS Annika and a second search engine and calculated the FDR in an automated manner (Examples of output Figures automatically created by IMP-X-FDR are shown in Supplemental Figure 4 and 5). The fraction of commonly identified unique crosslinks contains up to 629 entries (Annika + MeroX) and within this fraction the real FDR is ≤ 1.4 % in all cases and therefore very close to the accepted 1%. On the contrary those crosslinks exclusively identified by only one search engine contain most false positives yielding to FDR rates of up to 31%. A similar effect is also observed for replicate measurements (Figure 1B). Of 425 unique crosslinks commonly found in three replicates only 2 (0.5 %) were incorrect. The overlap of incorrect crosslinks is very low, showing that accumulating IDs from several replicate measurements to boost XL numbers is prone to also accumulate wrong hits, and should therefore be avoided. We further investigated those two XLs that were incorrectly assigned in all three replicates using Annika (Figure 1B): The first one is a homeotypic link of the peptide MAKLTK that does not exist in the library (but in date database used to search the files). A peptide with the sequence MAKTIK of the same mass is however part of our library and was therefore very likely generating the wrongly annotated spectra. The second one connects two existing peptides (LSYDTEASIAKAK- VAVIKAVR) that are however within different groups.

In a next step we benchmarked the reagents DSSO, DSBU and CDI on the main library (Figure 2 A and B). Expectedly, the performance of DSSO and DSBU is on a similar level, since both have comparable spacer lengths of 10.1 and 12.5 Å respectively and the same reactive site. The two linkers bear different liable groups for MS based fragmentation which might lead to the assumption that differences in spectra quality explain the slight difference in unique XL numbers. Notably this effect is software specific. Annika performs very well with DSSO and scores DSSO crosslinks better than DSBU links (average score 279 for all DSSO links vs 269 for all DSBU links from our main library). In contrast MeroX performs very well with DSBU and scores those links slightly better (average score 131 for all DSSO links vs 133 for all DSBU links from our main library). In conclusion, when comparing MeroX results, DSBU (767 XLs on average) outperforms DSSO (658 unique XLs on average) in terms of XL numbers (data shown in Supplementary Table 2).

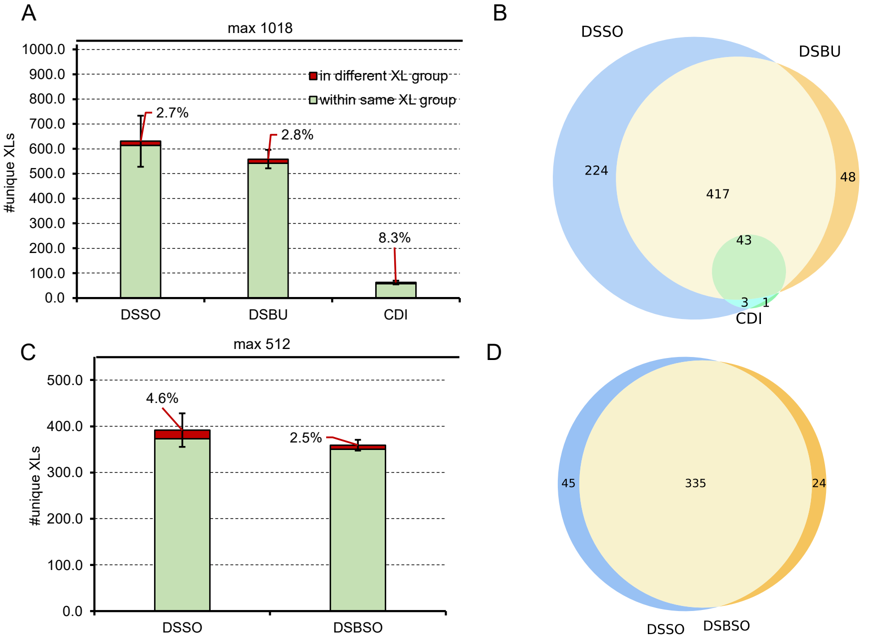


*Figure 2: **Benchmarking the linker reagents DSSO, DSBU, CDI and DSBSO**. (**A&C**) Average number of unique XL IDs and maximal theoretical number of true XL combinations when applying the indicated crosslinker to the main library (**A**) or the enrichable library (**B**) and data analysis using Annika at 1% estimated FDR. The real FDR is shown as callout, error-bars indicate standard deviations, n=3. (**B&D**) Overlap of identified XLs from one representative replicate of A (in **B**) or C (in **D**) respectively.*

The zero length crosslinker CDI yielded in ≤ 80 unique XLs identified with all tested algorithms. This low numbers might be reasoned by no “real” interaction sites within the peptide library that relies on crosslink connections formed between freely moving peptides in solution. Therefore, the likelihood of two peptides being connected by a crosslinker with a very short spacer is lowered compared to those linkers with a 10 – 12 Å spacer. A full list of XL numbers and real FDRs using all five tested algorithms can be found in Supplemental Table 2.

Next, we compared detectability of DSSO vs DSBSO using the enrichable peptide library (Figure 2 C and D). In this artificial system any potential steric hindrance of the azide tag of DSBSO can be neglected, hence we assume that differences in observed crosslinks are reasoned mainly by the ionizability of the resulting connected peptides. As illustrated in the Venn diagram in Figure 2 D, the overlap of identified crosslinks is indeed very high and could not be distinguished to an overlap of replicate measurements from the same linker (compare to Figure 1 B). Furthermore, DSSO only slightly outperforms DSBSO by means of crosslink numbers indicating similar reactivity and ionizability.

Benchmarking XL search engines with carboxylic acid reactive linkers

Next, we investigated two reagents targeting acidic amino acids: The non-cleavable adipic acid dihydrazide (ADH) and dihydrazide sulfoxide (DHSO) (Figure 3). These linkers were applied to a smaller peptide library with a reduced number of only 280 theoretically possible crosslinks formed, however, less than 40 % of this number was identified in all cases. This indicates a lowered reaction efficiency compared to the more established NHS ester-based linkers, where more than 60 % of the theoretical crosslink number was reached (Figures 1 and 2).

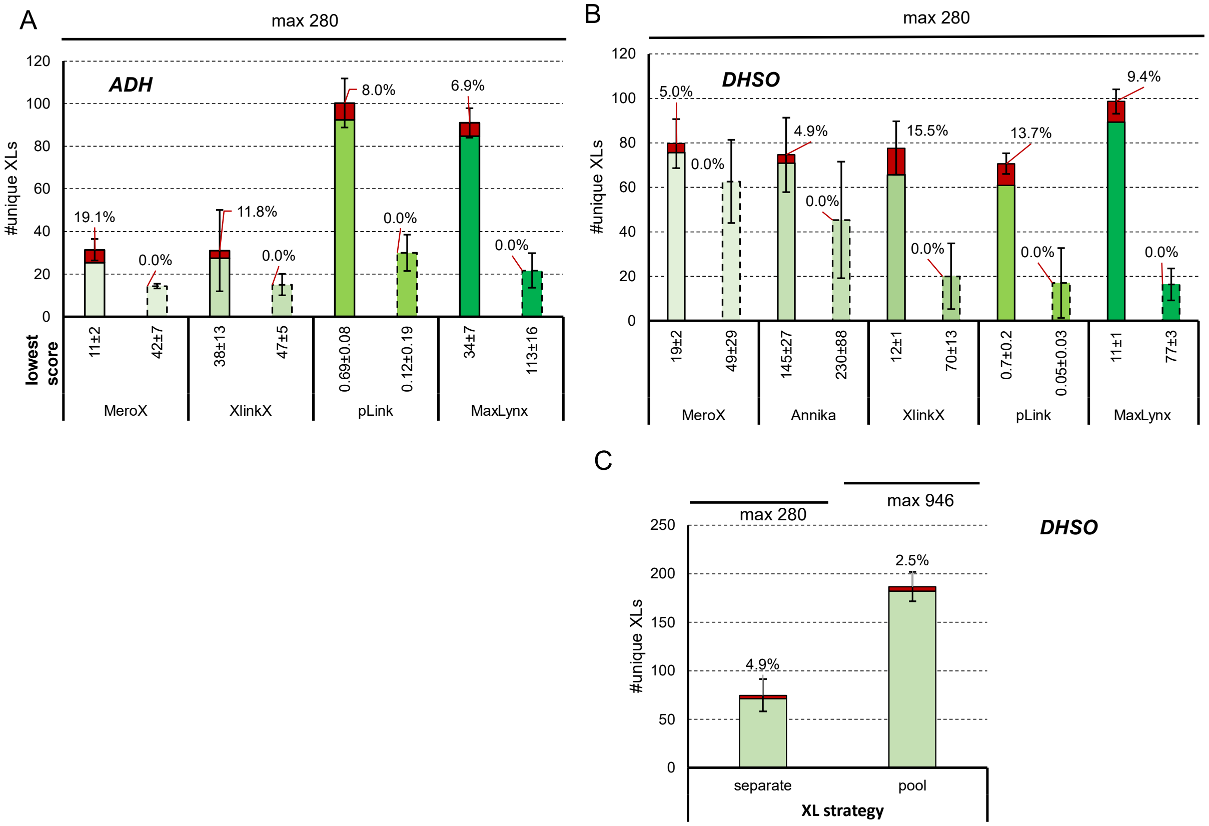


Figure 3: ***Benchmarking linker reagents linker reagents reactive to acidic amino acids****. (****A&B****) Average number of unique XL IDs and maximal theoretical number of true XL combinations when using ADH (****A****) or DHSO (****B****) to crosslink the acidic library. Data analysis was performed using the indicated algorithm at 1% estimated FDR and corrected by applying a post-score cutoff to reach a real FDR<=1. The real FDR is shown as callout, error-bars indicate standard deviations, n=3. (****C****)*

For the non-cleavable ADH linker, pLink and MaxLynX seem to perform significantly better, both by means of reliability and ID numbers, than their competitors. However, calculated real FDR values seem extraordinarily high for both reagents and every software tested on the acidic library. A proper FDR calculation might be hindered by the relatively small number of crosslinks available in this system.

Of note, 4-methylmorpholinium chloride (DMTMM), that was used as coupling reagent for ADH and DHSO, could form zero-length connections between amines and carboxylic acids. However, only two synthetic peptides of the acidic library contain any lysine residue except for those that are terminal after tryptic cleavage. We investigated the presence of DMTMM crosslinked (undigested) peptides and found no evidence for such a side reaction. The low number of crosslink identifications in the acidic library might be reasoned by a slow reaction kinetics and the fact that two steps (activation of carboxylic acids by DMTMM followed by nucleophilic attack of the hydrazine group) are required instead of only one as is the case for NHS based reagents. To boost the number of crosslink IDs we further tested DHSO on a pool of all 41 peptides of the acidic library. This increases the number to possible XL combinations from 280 to 946 and therefore close to the value of the main library. The number of identified crosslinks maintained low at ~20% of the theoretical maximum when using Annika (Figure 3 C).

To better understand the reaction chemistry of these hydrazine-based linkers, we analyzed the results obtained for DHSO using our in house developed tool IMP-X-FDR to investigate the distribution of amino acids in detected crosslink-sequence-matches (CSMs) (Supplementary Figure 3 A and B). We thereby compared the average frequency of specific amino acids in proximity to the crosslinked aspartic- or glutamic-acid in identified CSMs to the theoretically expected distribution. The theoretical distribution was calculated from all, in silico generated, crosslinks that can exist within the acidic library (either crosslinked in separate groups or within one pooled group) under the assumption that every crosslink combination led to exactly one CSM. By that, missing or predominant combinations can be visualized. For both DHSO based datasets (pooled and separate, as shown in Figure 3 C) similar dependencies popped up: Histidine, isoleucine, phenylalanine, tryptophan, tyrosine, and glutamine seem to reproducibly hinder the formation or identification of a crosslink from the acidic library. The frequency of amino acids within identified linear peptide sequence matches of the (non-crosslinked) acidic library was compared to the theoretical amino acid distribution under the assumption of equimolar peptide quantities (Supplemental Figure 3 C) in an additional experiment. Thereby the same MS method as for crosslink samples was used, meaning ions with a charge ≥3 are selected for fragmentation. We indeed found less peptides containing isoleucine than expected, but clearly more peptides than expected containing histidine. All other amino acids that seemed to have a negative impact on crosslink formation were found in relative frequencies as expected. Except for isoleucine this data strengthens the hypothesis that those amino acids negatively influence the reactivity of DHSO. Especially the basic histidine might cross-react with the activated carboxylic acid to form an intrapeptidal link, therefore impeding the reaction to DHSO.

Testing the influence of separate FDR calculation and minimal peptide length

Apart from MeroX, all tested search engines allow their users to decide on performing a separate inter-/intra-crosslink FDR calculation. MeroX calculates FDR of intra- and inter-protein XLs as well as dead-end-links in separate groups by default. A separated target-decoy based FDR calculation is considered useful as the group of interprotein connections is much larger compared to the theoretical intra-protein crosslinks that can form. Especially when using large databases, most wrong identifications were reported in the group of interprotein connections[28,29], and was estimated in the range of 20 – 25 % false positives within a dataset of 2 % overall FDR.[28] In contrast to our previously published peptide library[22] consisting of peptides from only one protein, the main library of this study contains 842 theoretical inter-protein crosslinks, 100 intra-protein crosslinks and 100 homeotypic crosslinks, which is why this system nicely represents a real protein mix sample (i.e. *E.coli* ribosome). In conclusion we were wondering if the FDR calculation in separate groups does also influence our results using the peptide library (Figure 4 A). Using pLink or MaxLynx the number of crosslinks but also FDR slightly increases when disabling separate FDR calculation. XlinkX predominantly adds wrong crosslinks to its result file upon enabling separate FDR calculation. Enabling or disabling this option does however not influence the result when using Annika. We representatively investigated the real FDR for inter- and intra-crosslinks on one replicate, yielding to the unexpected result of a higher FDR for intra-XLs: 50 intra-crosslinks of which 6.0 % are wrong and 615 inter-crosslinks of which 2.4 % are wrong were obtained. In contrast to our expectations, the separate FDR calculation did not significantly improve overall FDR or ID numbers independently of the search tool used. This might still be reasoned by the nature of our artificial library system that was searched against a database of 171 ribosomal proteins. Hence, peptides of 133 proteins contained in the database are not existent in the sample, leading to a disproportional large number of theoretical vs existing inter-protein crosslinks.

Next, we tested the influence of the peptide length of the shorter peptide within a linked pair on result quality. Figure 4 B clearly illustrates that shorter peptides are more prone for wrongly annotated spectra. This fits our expectations as (too) short peptides will generate fewer fragments and therefore yield in less confident identifications. In a large database the chance of sequences from different protein overlapping by chance is furthermore increased with decreased peptide length potentially leading to ambiguous identifications. Based on our data a minimal peptide length of 6 or even 7 seems beneficial, although >100 unique crosslinks are lost when excluding results containing peptides of 6 amino acids length. Of note, our library contains no peptide that has a sequence length of only 5, which is why that group contains exclusively wrong hits.

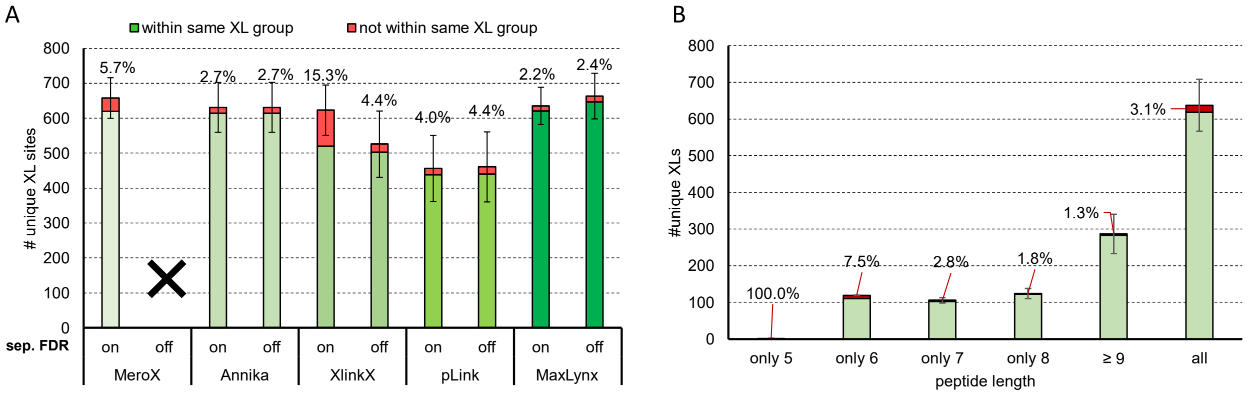


Figure 4: ***Effect of separate inter/intra FDR calculation and minimal peptide length on FDR****. (****A****) Average number of unique XL sites identified from the DSSO crosslinked main library with separate FDR calculation for inter- and intra-crosslinked peptides set on or off. Error bars indicate standard deviations, real FDRs are shown as callout, n=3 (****B****) Average number of crosslinks from the DSSO linked main library identified with Annika at 1 % estimated FDR when filtering for crosslinked peptides of the given length.*

Influence of increased sample complexity and XL enrichment.

To mimic more realistic conditions – where non-crosslinked linear peptides are way more abundant – we spiked the main library into a tryptic digest of linear HEK peptides at a mass ratio of 1:5. The resulting mixture was analyzed by means of LC-MS/MS and crosslink searches were performed against databases of different sizes, starting with the ribosomal database (171 proteins) that was also used for all other searches and ending with proteome wide searches (Figure 5 A). Annika, pLink and MaxLynx maintain FDR at levels below 10 % but loose up to 50 % or more of their identifications upon increasing the database size to a proteome wide search. In contrast, MeroX and XlinkX maintain their identification numbers at a high level at cost of data reliability, leading to very high FDR values of up to 26 %. The database size furthermore influences the minimal score (maximal for pLink) accepted as more decoy hits can be found (e.g. Annika increases its accepted minimal score from 120 to 214). When enriching the spiked library by size exclusion chromatography (SEC) we were able to (re-)boost identifications to the level seen without spiking (Figure 5 B) and obtaining the same trends with regards to FDR. We additionally applied post-score cutoffs to the results using the largest database based on the scores that yielded in 1 % real FDRs in our initial non-spiked measurements (shown in Figure 1A) and that are more stringent than those cutoff values recommended by the authors of the respective search engines. This improves the real FDR, that is however still ranging from 2.1 % for pLink to 10.5 % for MeroX in the spiked samples (Supplementary Figure 1). Our results suggest that the choice of a properly sized database is of high importance for the reliability of the results as well as that post-score cutoffs to minimize effects of improper FDR estimation need to be empirically determined in dependence of used software and complexity of the sample.

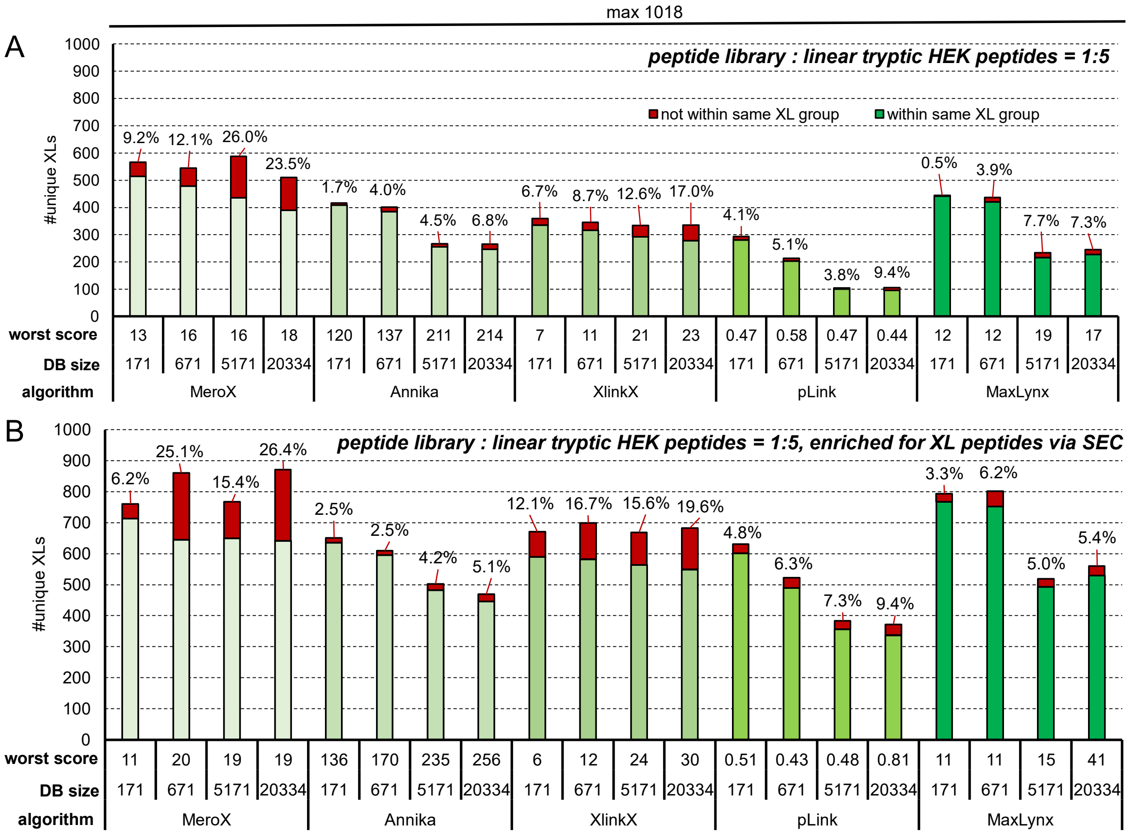


Figure 5: ***Performance benchmarking in a mimicked complex environment and upon increased database size****. The DSSO linked main library was mixed with linear tryptic HEK peptides (1:5 w/w). Bars indicate the number of unique XL sites identified using the indicated algorithm at 1% estimated FDR when using databases containing exclusively 171 E. coli ribosomal proteins, or additional 500, 5000 or 20163 human proteins. (**A**) direct measurement (**B**) measurement after enrichment for crosslinked peptides by size exclusion chromatography.*

To check for the performance of affinity-enrichment using the azide tagged linker DSBSO we also spiked the enrichable library, containing no azide-protected lysines, with linear tryptic HEK peptides in mass ratios of 1:10 or 1:100 (Figure 6). Enrichment was performed by clicking crosslinked peptides to beads functionalized with dibenzocyclooctyne (DBCO) as previously described.[30]

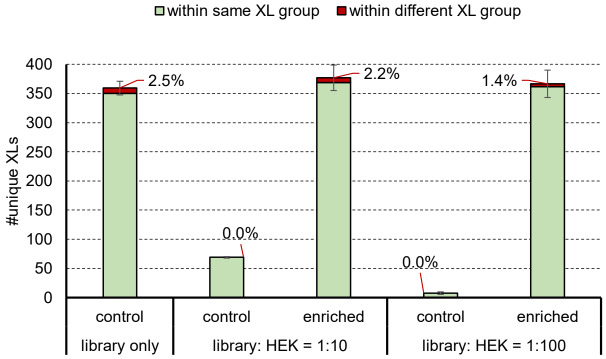


Figure 6: ***Affinity enrichment of DSBSO crosslinked synthetic peptides from a complex environment****. Average number of unique XL sites identified in the DSBSO crosslinked enrichable library with or without spiking using linear tryptic HEK peptides as indicated. Controls were directly used for measurement; enriched samples were subjected to affinity enrichment.*

The total amount of peptides subjected to MS analysis was kept constant at 1 µg for all injections as this seemed maximal for our LC-MS setting. This means that the 1:100 control sample contains 10 ng crosslinked peptides. For enrichment 20 µg crosslinked peptides were spiked with 200 µg or 2 mg HEK peptides resulting in 1.3 µg total peptides on average, and independent of the spike ratio, in the enriched fraction. Although quite some input material was lost during enrichment, the theoretical input can be freely upscaled to compensate. The resulting enriched samples were of high purity, enabling the injection of close to 1 ug cross-linked material even in samples with high amounts of background (instead of only 10 ng, as in the 1:100 spiked control) and therefore maintaining constantly high crosslink numbers and low FDR values independently of the sample complexity prior to enrichment.

Influence of additional FAIMS separation on resulting XL identification numbers and properties

High-field asymmetric-waveform ion-mobility spectrometry (FAIMS) adds another separation dimension and therefore decreases spectrum complexity and reduces noise. Both effects were reported beneficial for the identification of crosslinked peptides.[20] We probed the effect of FAIMS on the here presented synthetic peptide system using DSSO (Supplementary Figure 2). In line with the results of Schnirch and coworkers[20], we observe a maximum number of XL IDs when using compensation voltages (CV) in the range of -50 to -60V. Furthermore, we observed very high reproducibility when using FAIMS and a trend to lower FDR values upon lowered CV. The combination of 3 CVs within one run boosted our overall identification number to 700 which is a 10% increase compared to our measurements without FAIMS (numbers from analysis using Annika at 1 % estimated FDR).

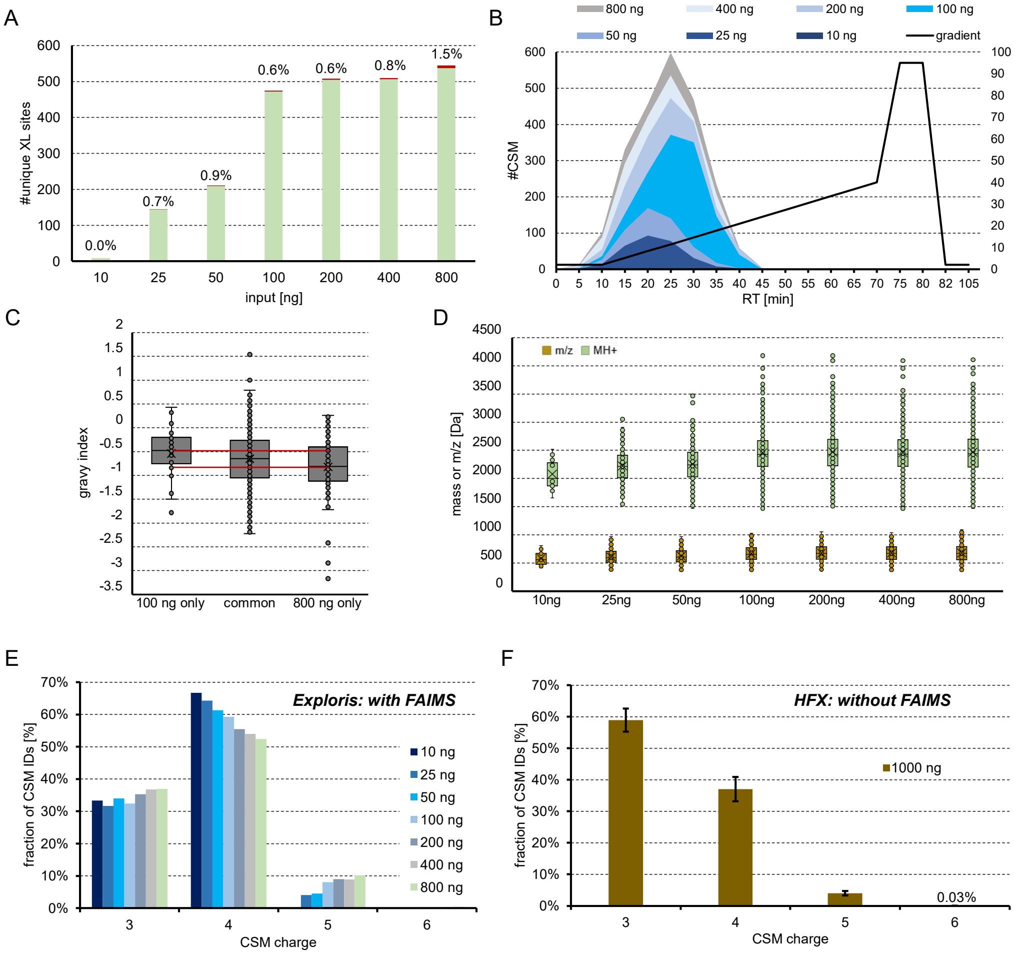


*Figure 7: **Variation of input amount and physicochemical properties of XLs.** Lowered input amounts, as indicated, of the DSSO cross-linked main library were measured on an Orbitrap Exploris 480. Data was analyzed using Annika at 1% FDR. **(A)** Unique XL sites and real FDR given above bars. **(B)** Distribution of spectral matches over retention time and used gradient. **(C)** Distribution of hydrophobicity of CSMs identified exclusively in the 100 or 800 ng sample or in both samples. **(D)** Distribution of identified cross-linked peptide masses of CSMs (M+H) and m/z in dependence of used input amount. **(E)** Charge distribution obtained with FAIMS using CVs -50, -55 and -60 for acquisition and stepwise changing the input amount from 10 – 800 ng or **(D)** without FAIMS.*

Next, we investigated if lower injection amounts are sufficient for identification of crosslinks thanks to the improved signal to noise ratios when using FAIMS. Without FAIMS 1000 ng were injected for all samples, which was stepwise lowered to down to 10 ng with FAIMS. Judged from ID numbers, 100 ng yielded only 1/8 of the maximal amount used, was sufficient to still identify close to 500 unique XLs which is ~90% of the identified XLs using 800 ng input (Figure 7 A). Further lowered amounts lead to a drastic decrease of XL IDs. When comparing where within the gradient crosslinked peptides were identified, we observed a slight shift towards higher retention times when higher peptide amounts were used (Figure 7 B). This effect is not reasoned by an overloaded column as the retention time of individual crosslink-sequence-matches (CSMs) did not change, but rather by identified additional CSMs. When looking into the physicochemical properties of the additionally identified CSMs, a shift in hydrophobicity (Figure 7 C) as well as an increase of molecular weights and m/z (Figure 7 D) can be observed, which might explain shifted retention times upon increasing input amounts. Furthermore, the relative charge distribution (Figure 7 E) depended on the input amount. The fraction of high charged z = 5 ions is increased, while the fraction of z = 4 charged ions is lowered with higher input. This observation is in line with the seen dependency of molecular weight to input amount as larger peptides are more likely highly charged in an acidic environment. When comparing this charge distribution to those observed without FAIMS (Figure 7 F) a clear shift from predominant z = 3 charged ions without FAIMS to predominant z = 4 charged ions with FAIMS can be observed, which is beneficial for the detection of predominantly higher charged crosslinked ions over linear peptides.

DISCUSSION

The here presented peptide libraries represent a new and highly flexible standard to benchmark crosslinker reagents of different chemical reactivity as well as for comparison of search engines or acquisition strategies. Thanks to peptide sequences originating from 38 different proteins, the library represents a realistic digest from the *E. coli* ribosomal protein complex, allowing for in-depth analysis of search-engine specific FDR calculation. Our in house developed tool IMP-X-FDR comes with an easy-to-use user interface and allows FDR calculation, comparison of crosslink results across software platforms and investigation of crosslink-properties in an automated manner also for non-bioinformaticians.

Our results suggests that additional and empirical score cutoffs are a valuable instrument to correct the actual FDR. The height of this cutoff value is not only software specific but also dependent on the size of the database used and the sample complexity. Our library helps to find such specific score-cut-off values but also showed that built-in target-decoy based FDR estimation overestimates correct results in case of all tested search engines. Especially when FDR calculation is performed on spectrum level, false hits were shown to propagate during grouping to the final unique crosslinked sites leading to higher than expected final real FDR[31]. In conclusion, FDR calculation at crosslink level, machine learning approaches as well as the inclusion of additional parameters like retention time or ion mobility to (re-)score identifications might improve crosslink numbers and validity of FDR in future algorithms.

The most challenging part of data analysis seems to find a good compromise in the trade between high identification numbers and low FDR. In our data, MS Annika and MaxLynx seem to find this optimum best for cleavable crosslinkers, while pLink seems to perform very well for the tested non-cleavable linker.

All tested search engines allow to define new crosslinker reagents, however in some cases we observed lacking possibilities to define all properties of linker reagents properly. To give an example, sulfoxy based linkers as DSSO or DSBSO were reported to generate characteristic doublet peaks of two different delta masses upon MS fragmentation thanks to water elimination[12,32]. Upon the tested algorithms, MaxLynx and pLink do not allow a definition of more than two fragments and their results might be further improved upon implementation. Another example is Thermos Proteome Discoverer 2.5, were the definition of a fragment mass 0 is impossible but needed for the zero-length linker CDI. As a workaround a very low mass ≥$1E^{-5}$ must be defined. This affects search engines as XlinkX or Annika when running as a node within Proteome Discoverer.

To conclude, we believe that no gold standard for data-analysis in the field of XL-MS has accomplished yet and future software updates, or entirely novel algorithms will stepwise improve the reliability of results while increasing coverage of crosslink identifications. Our data will therefore provide valuable input to benchmark new or updated search engines. The freely available IMP-X-FDR can be easily adopted for automated FDR calculation with any novel XL search engine thanks to the open-source code. Furthermore, improvements in crosslinker reagents, MS instrumentation or chromatography can be validated using the physical library where the exact number of theoretically reachable crosslinks is well defined.

METHODS

Peptide synthesis

Solid phase peptide synthesis was done using Fmoc chemistry on a SYRO with Tip Synthesis Module (MultiSynTech GmbH). Each coupling step was performed as double coupling using HATU/DIEA for carboxylic acid activation. Lysine residues at the C-terminus bore an azide group instead of an amine to hamper any cross-linking at this position. N-termini were designed as acetyl protected WGGGGR sequence tag and C-termini were designed as amide protected RGGGG sequence tag (for peptides to be used with linkers reactive to acids, see Supplementary Table 1 for all sequences). For this Fmoc-L-Arg(Pbf)-TCP (# PC-01-0126), Fmoc-Rink-Amide-(aminomethyl) (#PC-01-0501) or Fmoc-L-Lys(N3)-TCP (custom synthesized) resins were used respectively (all: INTAVIS Peptide Services GmbH & Co. KG). Purification was performed using a C18 kinetex column (5 µm) and a 30 min gradient. All peptides were analyzed using a 4800 MALDI TOF/TOF (Applied Biosystems) for quality control purposes. Lyophilized peptides were solubilized in water and their concentration was estimated by measuring their absorption via a nanodrop (DeNovix DS-11 FX+) at 280 nm and calculating the sequence specific extinction coefficient using the ProtParam tool[33]. Peptide solutions were dried under reduced pressure, resolubilized in 50 mM HEPES pH 7.5 at a concentration of 5 mM and mixed to groups for cross-linking (Supplementary Table 1).

Sample preparation

For lysine reactive cross-linker reagents (DSSO, DSBSO, DSBU, CDI) 9.3 mM cross-linker reagent stock solutions were freshly prepared in dry DMSO. 0.5 µL of stock solution was added to 1 µL of each peptide group in separate vials. Additional stock solution was added 4x every 30 min adding up to a total of 2.5 µL cross-linker stock solution. The resulting 3.5 µL reaction mix were quenched using 31.5 µL 100 mM ammoniumbicarbonate (ABC) buffer for 30 min and pooled to a single tube. The resulting mix was digested by addition of 5 ng trypsin/group over night at 37 °C. Azide protection groups were finally reduced to the respective amines by incubation to 50 mM (final concentration) tris(2-carboxyethyl)phosphine (TECEP) for 30 min at room temperature. Reduced peptides were pooled to a single vial, aliquoted and stored at -70 °C upon further usage.

For aspartic acid and glutamic acid reactive cross-linker reagents (DHSO, ADH) 300 mM cross-linker reagent and 1.2 M (4-(4,6-dimethoxy-1,3,5-triazin-2-yl)-4-methyl-morpholinium chloride) (DMTMM) stock solutions were prepared in 25 mM HEPES pH 7.5. 0.25 µL of cross-linker and DMTMM stock-solution were added 5x every 30 min to 1 µL of each peptide group. The reaction was quenched by adding trifluoracetic acid (TFA) to a final concentration of 4 % (w/v) for 20 min followed by re-neutralization by addition of 50 µL 1M Tris pH 7.5 buffer. Peptides were pooled and digested as described above.

Enrichment strategies

To mimic complex mixtures, cross-linked and digested peptide pools were mixed with a 5-100x excess (by mass) of tryptic HEK peptides. The resulting spiked samples were enriched either by size exclusion chromatography (SEC) or via affinity enrichment.

For SEC, ~10 µg of cross-linked peptide-library + typtic HEK peptides were fractionated on a TSKgel SuperSW2000 column (300 mm × 4.5 mm × 4 µm, Tosoh Bioscience), which was operated at 200 µl/min in 30 % ACN, 0.1 % TFA. Fractions were collected every minute, ACN was removed under reduced pressure to obtain a concentrated sample for LC-MS/MS

DSBSO cross-linked peptides (+ linear HEK peptides in varying mass ratios) were affinity enriched using dibenzylcyclooctyne (DBCO) immobilized to beads as described elsewhere[30].

Tryptic HEK peptides were generated as follows: HEK cells were lysed in 10 M urea in 100 mM Tris by ultrasonication. The cleared lysate was reduced at a final concentration of 10 mM dithiothreitol in the presence of benzonase for 1 h at 37 °C. This was followed by alkylation at a final concentration of 20 mM iodoacetamide for 30 min at room temperature in the dark. Digestion was performed using LysC (1:200 w/w) for 2 h at 37 °C in 6 M urea followed by addition of trypsin (1:200 w/w) for 16 h t 37 °C in 2.5 M urea.

Chromatographic separation and mass spectrometry

Samples were separated using a Dionex UltiMate 3000 HPLC RSLC nano-system coupled to an Q Exactive^{TM} HF-X Orbitrap mass spectrometer or to an Orbitrap Exploris™ 480 mass spectrometer equipped with a FAIMS pro interface (all: Thermo Fisher Scientific). Samples were loaded onto a trap column (Thermo Fisher Scientific, PepMap C18, 5 mm × 300 µm ID, 5 µm particles, 100 Å pore size) at a flow rate of 25 µL min-1 using 0.1 % TFA as mobile phase. After 10 min, the trap column was switched in line with the analytical column (Thermo Fisher Scientific, PepMap C18, 500 mm × 75 µm ID, 2 µm, 100 Å). Peptides were eluted using a flow rate of 230 nl min^{-1}, with the following gradient: 0 -10 min 2 % buffer B, followed by an increasing concentration of buffer B up to 40 % until min 130. This is followed by a 5 min gradient from reaching 95 % B, washing for 5 min with 95% B, followed by re-equilibration of the column in buffer A at 30°C (buffer B: 80 % ACN, 19.92 % H_2O and 0.08 % TFA, buffer A: 99.9% H_2O, 0.1% TFA).

The mass spectrometer was operated in a data-dependent mode, using a full scan (m/z range 375-1500, nominal resolution of 120.000, target value 1E6). MS/MS spectra were acquired by stepped HCD using an NCE (normalized collision energy) of 27±6 for sulfoxy group linkers (DSSO, DSBSO, DHSO), 30±3 for urea-based linkers (DSBU, CDI) and 28±4 for non-cleavable linkers (ADH). An isolation width of 1.0 m/z, a resolution of 30.000 and a target value of 5E4 (on HF-X) and 1E5 (on Exploris) was set. Precursor ions selected for fragmentation (± 10 ppm, including exclusively charge states 3-8) were put on a dynamic exclusion list for 30 s. Measurements including FAIMS were performed on the Orbitrap Exploris under alteration of used compensation voltages as given for each result.

Data Analysis and post processing

Data analysis was performed against a custom shotgun database containing 171 *E. coli* ribosomal proteins at 1 % FDR level. For analyses using Annika or XlinkX, Thermo raw files were loaded to Thermos Proteome Discoverer 2.5 that and both search engines were used as node within that software. MaxLynx was used as part of MaxQuant v 2.0.2.0 by direct usage of Thermo raw files as well. For MeroX, raw files were converted to mzML and for pLink files were converted to mgf using MSConvertGUI v3.0.21084. The result files are available for download in the PRIDE repository[34] using the identifier PXD029252. The software specific settings are furthermore summarized in Supplementary Table 3.

Post processing was done using the graphical user interface of our in house developed tool IMP-X-FDR (Supplemental Figure 4 E). It enables to calculate the real FDR based on crosslinks only allowed as correct in case they are formed within the same XL group (see Supplementary Table 1 for allocation of peptides to groups). We call this functionality “FDR recalculation” and adopted the code for each XL search engine, due to differences in their output format. For a correct FDR recalculation, a support file containing all group-allocated peptides of all used (sub) peptide libraries is provided with the software. The tool outputs a csv file containing a list of al XLs within the same or different group as well as informative graphs showing the number of IDs and the score vs real FDR or number of crosslinks (Supplemental Figure 4 A-C). The functionality “Venn diagrams” of IMP-X-FDR was used to visualize the overlap of replicates of searches from different search algorithms (example output shown in Supplemental Figure 4 D). This functionality uses the output of “FDR recalculation” as input, which ensures a uniform format and compares peptide sequences, their originating protein, and the position of the peptide in that protein.

The third function of IMP-X-FDR is to investigate physicochemical properties of crosslinks. To do so the freely available tools from Biopython[35], specifically from the Bio package, Bio.SeqUtils subpackage and Bio.SeqUtils.ProtParam module, were used. Crosslinked peptides were represented in a linearized form to ensure compatibility with the used packages originally designed for linear peptides. IMP-X-FDR outputs a csv file containing calculated crosslink properties, which includes the isoelectrical point, fraction of aromatic amino acids, molecular mass, gravy value and amino acid distribution. The obtained data is automatically compared to the respective properties of all (in silico generated) theoretically formed crosslinks within the library. Thereby we assume the identification of exactly one CSM for each theoretical crosslink. The unnormalized output graphics are constructed on the crosslink level and histograms constructed on CSM level are normalized to a total area of 1. Finally IMP-X-FDR investigates amnio acid motives using the module seqlogo 5.29.8[36] to create position probability matrices. Thereby the closest three neighboring amino acids of the linker’s binding site are investigated for frequent or rare amino acids and can be compared to the (theoretically expected) crosslinks within the library. Representative output graphs are illustrated in Supplementary Figure 5. A user’s manual, containing a detailed explanation of each output file and used functions is delivered with IMP-X-FDR. The code is freely available (https://github.com/fstanek/imp-x-fdr) and can be used on command line basis or via a graphical user interface.

ASSOCIATED CONTENT

Supporting Information

Supplementary Table 1: List of all synthesized peptides and their annotation to groups for crosslinking

Supplementary Table 2: List of all crosslink IDs at 1% estimated FDR from crosslinked library samples measured without FAIMS and real FDRs from main, enrichable and acidic library.

Supplementary Table 3: Search settings used form MeroX, Annika, XlinkX, pLink or MaxLynx analyses.

Figure S 1: Performance benchmarking in a mimicked complex environment searching against the full proteome.

Figure S 2: DSSO linked peptides on a FAIMS equipped device

Figure S 3: Influence of specific amino acids in proximity to the XL-site influencing the formation of a crosslink

Figure S 4: Exemplary output figures of IMP-X-FDR.

Figure S 5: Exemplary output figures of the physicochemical cross-link properties functionality of IMP-X-FDR

AUTHOR INFORMATION

Author Contributions

The study was designed by MM and KM. Experiments were performed by ADV. IMP-X-FDR was created by ADV and wrapped into a user interface by FS. M. Madalinski performed peptide synthesis. Experiments were performed and the manuscript was written by MM. All authors have given approval to the final version of the manuscript.

ACKNOWLEDGMENT

This work supported by the EPIC-XS, Project Number 823839, funded by the Horizon 2020 Program of the European Union and the by the ERA-CAPS I 3686 project of the Austrian Science Fund. We thank the IMP for general funding and access to infrastructure and especially the technicians of the protein chemistry facility for continuous laboratory support. Our gratitude further goes to Dr. Elisabeth Roitinger for fruitful discussions and for her valuable inputs to the manuscript.

REFERENCES

1. Iacobucci, C., Götze, M. & Sinz, A. Cross-linking/mass spectrometry to get a closer view on protein interaction networks. *Curr. Opin. Biotechnol.* **63**, 48–53 (2020).
2. Piersimoni, L. & Sinz, A. Cross-linking/mass spectrometry at the crossroads. *Anal. Bioanal. Chem.* (2020) doi:10.1007/s00216-020-02700-x.
3. Belsom, A. & Rappsilber, J. Anatomy of a crosslinker. *Curr. Opin. Chem. Biol.* **60**, 39–46 (2021).
4. Matzinger, M. & Mechtler, K. Cleavable Cross-Linkers and Mass Spectrometry for the Ultimate Task of Profiling Protein-Protein Interaction Networks in Vivo. *J. Proteome Res.* **20**, 78–93 (2021).
5. Pilch, P. F. & Czech, M. P. Interaction of cross-linking agents with the insulin effector system of isolated fat cells. Covalent linkage of 125I-insulin to a plasma membrane receptor protein of 140,000 daltons. *J. Biol. Chem.* **254**, 3375–3381 (1979).
6. Staros, J. V. N-hydroxysulfosuccinimide active esters: bis(N-hydroxysulfosuccinimide) esters of two dicarboxylic acids are hydrophilic, membrane-impermeant, protein cross-linkers. *Biochemistry* **21**, 3950–3955 (1982).
7. D'Este, M., Eglin, D. & Alini, M. A systematic analysis of DMTMM vs EDC/NHS for ligation of amines to hyaluronan in water. *Carbohydr. Polym.* **108**, 239–246 (2014).
8. Leitner, A. *et al.* Chemical cross-linking/mass spectrometry targeting acidic residues in proteins and protein complexes. *Proc. Natl. Acad. Sci.* **111**, 9455–9460 (2014).
9. Smyth, D., Blumenfeld, O. & Konigsberg, W. Reactions of N-ethylmaleimide with peptides and amino acids. *Biochem. J.* **91**, 589–595 (1964).
10. Partis, M. D., Griffiths, D. G., Roberts, G. C. & Beechey, R. B. Cross-linking of protein by ω-maleimido alkanoylN-hydroxysuccinimido esters. *J. Protein Chem.* **2**, 263–277 (1983).
11. Gomes, A. F. & Gozzo, F. C. Chemical cross-linking with a diazirine photoactivatable cross-linker investigated by MALDI- and ESI-MS/MS. *J. Mass Spectrom.* **45**, 892–899 (2010).
12. Kao, A. *et al.* Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. *Mol. Cell. Proteomics MCP* **10**, M110.002212 (2011).
13. Müller, M. Q., Dreiocker, F., Ihling, C. H., Schäfer, M. & Sinz, A. Cleavable cross-linker for protein structure analysis: reliable identification of cross-linking products by tandem MS. *Anal. Chem.* **82**, 6958–6968 (2010).

14. Burke, A. M. *et al.* Synthesis of two new enrichable and MS-cleavable cross-linkers to define protein–protein interactions by mass spectrometry. *Org Biomol Chem* **13**, 5030–5037 (2015).

15. Steigenberger, B., Pieters, R. J., Heck, A. J. R. & Scheltema, R. A. PhoX: An IMAC-Enrichable Cross-Linking Reagent. *ACS Cent. Sci.* **5**, 1514–1522 (2019).

16. Chavez, J. D., Weisbrod, C. R., Zheng, C., Eng, J. K. & Bruce, J. E. Protein Interactions, Post-translational Modifications and Topologies in Human Cells. *Mol. Cell. Proteomics* **12**, 1451–1467 (2013).

17. Stieger, C. E., Doppler, P. & Mechtler, K. Optimized Fragmentation Improves the Identification of Peptides Cross-Linked by MS-Cleavable Reagents. *J. Proteome Res.* **18**, 1363–1370 (2019).

18. Ihling, C. H., Piersimoni, L., Kipping, M. & Sinz, A. Cross-linking/Mass Spectrometry Combined with Ion Mobility on a timsTOF Pro Instrument for Structural Proteomics. (2021) doi:10.1101/2021.03.26.437136.

19. Steigenberger, B. *et al.* Benefits of Collisional Cross Section Assisted Precursor Selection (caps-PASEF) for Cross-linking Mass Spectrometry. *Mol. Cell. Proteomics MCP* (2020) doi:10.1074/mcp.RA120.002094.

20. Schnirch, L. *et al.* Expanding the depth and sensitivity of cross-link identification by differential ion mobility using FAIMS. *Anal. Chem.* (2020) doi:10.1021/acs.analchem.0c01273.

21. Yılmaz, Ş. *et al.* Cross-linked peptide identification: A computational forest of algorithms. *Mass Spectrom. Rev.* **37**, 738–749 (2018).

22. 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.* **11**, 742 (2020).

23. Pirklbauer, G. J. *et al.* MS Annika: A New Cross-Linking Search Engine. *J. Proteome Res.* (2021) doi:10.1021/acs.jproteome.0c01000.

24. Yılmaz, Ş., Busch, F., Nagaraj, N. & Cox, J. *Accurate and automated high-coverage identification of chemically cross-linked peptides with MaxLynx*. 2021.08.26.457759 https://www.biorxiv.org/content/10.1101/2021.08.26.457759v1 (2021) doi:10.1101/2021.08.26.457759.

25. Iacobucci, C. *et al.* A cross-linking/mass spectrometry workflow based on MS-cleavable cross-linkers and the MeroX software for studying protein structures and protein–protein interactions. *Nat. Protoc.* **13**, 2864 (2018).

26. Liu, F., Lössl, P., Scheltema, R., Viner, R. & Heck, A. J. R. Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. *Nat. Commun.* **8**, 15473 (2017).

27. Chen, Z.-L. *et al.* A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nat. Commun.* **10**, 3404 (2019).

28. de Jong, L., Roseboom, W. & Kramer, G. A composite filter for low FDR of protein-protein interactions detected by in vivo cross-linking. *J. Proteomics* **230**, 103987 (2021).

29. de Jong, L. *et al.* In-Culture Cross-Linking of Bacterial Cells Reveals Large-Scale Dynamic Protein–Protein Interactions at the Peptide Level. *J. Proteome Res.* **16**, 2457–2471 (2017).

30. Matzinger, M., Kandioller, W., Doppler, P., Heiss, E. H. & Mechtler, K. Fast and Highly Efficient Affinity Enrichment of Azide-A-DSBSO Cross-Linked Peptides. *J. Proteome Res.* **19**, 2071–2079 (2020).

31. Fischer, L. & Rappsilber, J. Quirks of Error Estimation in Cross-Linking/Mass Spectrometry. *Anal. Chem.* **89**, 3829–3833 (2017).

32. Kaake, R. M. *et al.* A New *in Vivo* Cross-linking Mass Spectrometry Platform to Define Protein–Protein Interactions in Living Cells. *Mol. Cell. Proteomics* **13**, 3533–3543 (2014).

33. Gasteiger, E. *et al.* Protein Identification and Analysis Tools on the ExPASy Server. in *The Proteomics Protocols Handbook* (ed. Walker, J. M.) 571–607 (Humana Press, 2005). doi:10.1385/1-59259-890-0:571.

34. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).

35. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

36. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).