universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Benchmark of crosslinking-mass spectrometry workflows using a peptide library“

verfasst von / submitted by

Adrian-Daniel Vasiu, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2023 / Vienna 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 875
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Bioinformatik
Betreut von / Supervisor:	Univ.-Prof. Dr. Bojan Zagrovic, AB
Mitbetreut von / Co-Supervisor:	

Acknowledgements

Firstly, I would like to thank Dr. Manuel Matzinger, who supported me continuously and assisted me along the way. You are an incredibly good researcher and a great mentor.

Furthermore, I am extremely grateful to Karl Mechtler for choosing me to be part of his group and giving me the opportunity to do cutting edge research and the freedom to test new ideas.

I gratefully thank Univ.-Prof. Dr. Bojan Zagrovic for giving me his support and expertise along the way and supervising my thesis.

Additionally, I would like to say a big thank to Florian Stanek and Dr. Gerhard Dürnberger because they have always given me positive impulses and sometimes saved me from problems that seemed impossible to solve at first. Programming is a beautiful world, but sometimes incomprehensible at first sight.

Finally, I like to say how thankful I am to my parents and Andreas who listened to me and were there when everything seemed hopeless.

Contributions

All the Python analysis scripts in IMP-X-FDR were developed by Adrian-Daniel Vasiu. The Github User's manual was written by Adrian-Daniel Vasiu and edited by Manuel Matzinger. The Graphical User Interface of IMP-X-FDR was developed by Florian Stanek.

Abstract

In nearly every biochemical research field, bioinformatics became almost indispensable in the last decades. This also stands true for crosslinking mass spectrometry (XL-MS).

This master thesis focuses on an extensive benchmark and quantitative comparison between several existing XL-MS data analysis tools and the development of novel Python scripts for maximizing XL counts and minimizing false discovery rate (FDR) values. Firstly, a description of common bioinformatics tools in proteomics is given, together with a brief introduction to crosslinking theory. Various algorithms and useful platforms for both crosslink search and visualization are presented, with a more detailed analysis of how the chosen benchmarked algorithms work. Using 3 synthetic peptide libraries, cleverly designed for specific crosslinker reactivities, the data analysis tools MeroX, MS Annika, XlinkX and pLink2 are benchmarked. They are challenged and pushed to their limits with regard to several parameters (such as sample complexity and user-settings). The focus is though on maximizing XL count and minimizing FDR values.

Software settings were harmonized to allow a fair comparison, and user-parameter influences were mapped. Diverse strategies, such as multiple analysis of the same sample with different algorithms for FDR mitigation, were proposed and found useful. Certain interesting incorrect annotations were discussed in more detail, and factors that a bioinformatician would need to consider in an *in vivo* experiment were mentioned, such as peptide length or database size.

Finally, the physicochemical properties of the identified links were examined, and certain patterns in the sequence were found to inhibit the reaction. Apart from this, trends were noticed indicating that the mass and GRAVY value of a peptide pair may play an important role in the identification of a crosslink. All of this was possible thanks to a newly developed tool in the bioinformatics arsenal of XL-MS research called IMP-X-FDR, which was written in Python and comes with a tailored user interface.

Zusammenfassung

In jedem biochemischen Forschungsfeld ist in den letzten Jahrzehnten die Bioninformatik fast unentbehrlich geworden. Dies gilt auch für die Crosslinking Massenspektrometrie.

Diese Masterarbeit konzentriert sich auf einen umfassenden Benchmark und einen quantitativen Vergleich zwischen verschiedenen bestehenden XL-MS-Datenanalysentools und die Entwicklung neuartiger Python-Skripte zur Maximierung der XL-Zahlen und Minimierung der Falschentdeckungsrate (False discovery rate or FDR). Im Laufe dieser Masterarbeit wird zunächst eine Beschreibung gängiger bioinformatischer Tools in der Proteomik gegeben, zusammen mit einer kurzen Einführung in die Crosslinking Theorie. Es werden verschiedene Algorithmen und nützliche Plattformen für sowohl die Crosslink-Suche als auch die Visualisierung präsentiert, wobei die Funktionsweise der ausgewählten gebenchmarkten Algorithmen genauer analysiert wird. Anhand mehrerer geschickt synthetisierter Peptidbibliotheken werden MeroX, MS Annika, XlinkX und pLink2 bezüglich mehrerer Parameter herausgefordert und an ihre Grenzen gebracht. Der Fokus liegt dabei auf der Maximierung von XL-Anzahl und der Minimierung von FDR-Werten.

Die Einstellungen der Software wurden harmonisiert, um einen fairen Vergleich zu ermöglichen, und die Einflüsse der Nutzerparameter wurden abgebildet. Diverse Strategien, wie beispielsweise eine multiple Analyse derselben Probe mit verschiedenen Algorithmen zur FDR-Minderung, wurden vorgeschlagen und als nützlich erachtet. Bestimmte interessante falsche Annotationen wurden genauer besprochen und Faktoren, die ein Bioinformatiker bei einem in vivo Experiment berücksichtigen müsste, wurden erwähnt, wie zum Beispiel die Peptidlänge oder die Größe der Datenbank.

Schließlich wurden die physikochemischen Eigenschaften der identifizierten Links untersucht, wobei sich bestimmte Muster in der Sequenz als hemmend für die Reaktion erwiesen haben. Abgesehen davon wurden Trends bemerkt, die darauf hinweisen, dass die Masse und der GRAVY-Index eines Peptidpaares eine wichtige Rolle bei der Identifizierung eines Crosslinks spielen können. All dies war möglich dank eines neuen Tools im bioinformatischen Arsenal der XL-MS-Forschung namens IMP-X-FDR, das in Python geschrieben wurde und über eine maßgeschneiderte Benutzeroberfläche verfügt.

Table of Contents

1. Theory

1.1. Crosslinking Mass Spectrometry

Crosslinking Mass Spectrometry (XL-MS) can be described as a vast scientific domain. This field, which is relatively new in the world of science, has once again united chemists, structural biologists, mathematicians, and bioinformaticians. XL-MS workflows start with a chemical reaction between linker (usually bifunctional) and specific amino acids of one or more proteins. The linker can be classified in many ways depending on:

- its spacer arm length (from zero-length[1] to CDI[2] to BS3[3]). The linkers with a shorter spacer arm would theoretically provide higher resolution XLs and a lower number of links created by a random interaction. On the other side, the quantity of the data is also reduced because a strong restriction is present: two "right" residues must be close to each other, otherwise the interaction cannot be observed.

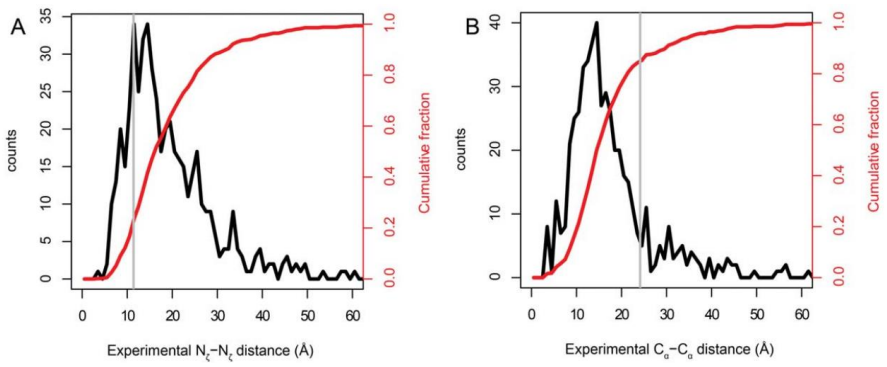


Figure 1: In subfigure A there are the counts of observed links depending on the N_ζ–N_ζ distances. In subfigure B one can observe the number of links relative to the C_α-C_α distances. The black lines denote the absolute frequency from the data set. The red line shows the cumulative distribution function. The vertical grey line represents the cut-off distance by taking in consideration the length of the linker and the lysine chain respectively. Table adapted from [3]

The data comes from XL experiments[4] performed on proteins with BS3 and DSS. They both have a spacer arm length of 11.4 Å. The fact that only approximately 20% of the found crosslinks are under this cut-off is not surprising taking into consideration the flexibility of the side chain (in this case two lysine residues). When the lysine side chains are included in the distance constraints, a cut-off of 24 Å is obtained for the C_α-C_α maximal distance. 84% of the intramolecular crosslinks observed do fit into this constraint suggesting that accounting for the high flexibility of side chains is a necessary measure. The rest of 16% (percentage may be varying in other studies, but it is continuously reported) suggest that not all of the crosslinks are coming from aggregates, but the protein backbone possesses a certain flexibility.[3] In the figure below, one can more clearly visualize what is meant by the C_α-C_α (of 24 Å) and N_ζ–N_ζ (of 11.4 Å) distances.

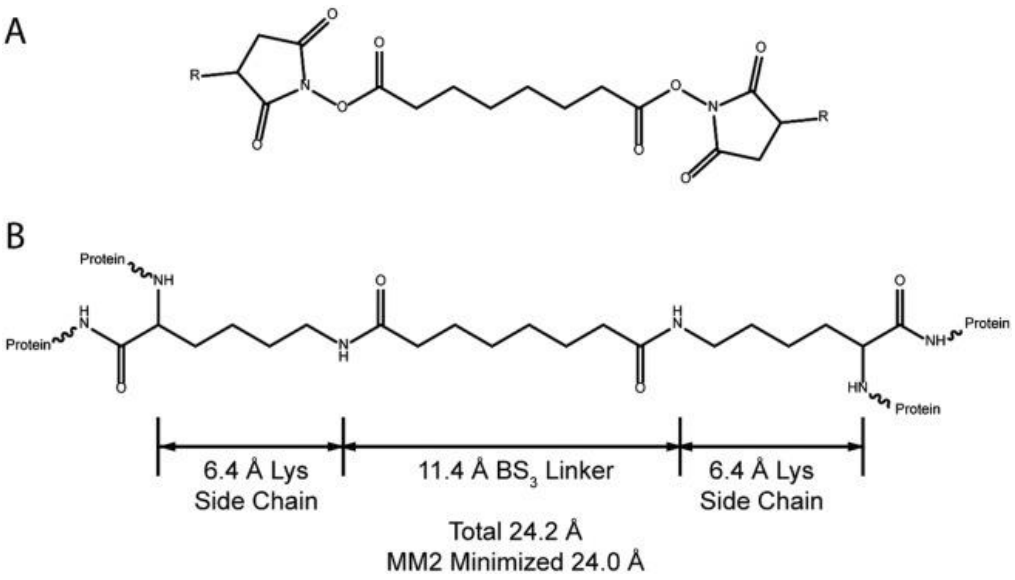


Figure 2:Chemical structure of a NHS ester- linker reagent in two contexts: A) with the NHS moiety attached and thus before the reaction B) after reacting on both ends with lysine residues. Figure adapted from "Distance restraints from crosslinking mass spectrometry: mining a molecular dynamics simulation database to evaluate lysine-lysine distances" [3]

- Enrichable/not enrichable. Most of the reagents used in XL-MS are non-enrichable linkers. The enrichable ones are trifunctional. The third handle is not designed for linking to a protein, but to be affinity- or reactive-enriched. Already in 2005, Protein Interaction Reporter (PIR) was designed and synthetized.[5] Besides its numerous functional moieties described in the figure below, it was proven that PIR is capable of entering cells and showing valuable PPI networks[6,7]. Its long spacer arm constitutes a problem for the distance constraints, but proved efficient in providing a reporter ion signal which is signalling the presence of an XL.[5]

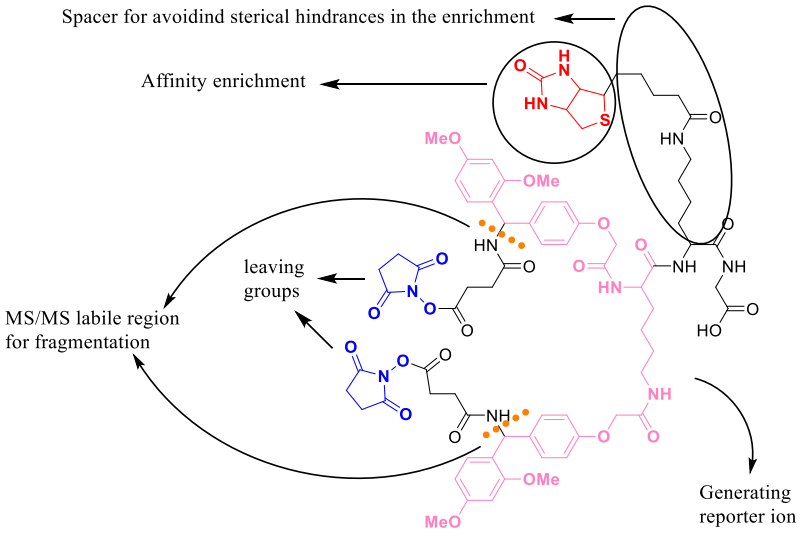


Figure 3: Description of the PIR[5]

- Chemistry type of the reactive groups. The first linkers introduced were N-hydroxysuccinimide (NHS) esters[8] and up to this day are this the preferred linkers because they are amine reactive and proved reliable yields. There are many advantages when lysine (which are primary amines) targeting linkers are employed such as high abundancy of lysine residues throughout the protein sequences, steric accessibility and favourable reaction kinetics. The disadvantage that

one must consider, when targeting lysines is that they do not present rigid side chains and therefore the distance constraints are relatively limited. NHS esters can also be equipped with sodium sulfonate functional group for enhanced solubility.[9] Thiol-reactive linkers are targeting the free cysteines in a protein. The cysteines involved in disulfide bonds cannot be targeted, because reducing the disulfide bond prior to crosslinking might affect the entire structure of the protein. 1,3- diallylurea (DAU) represents an example of such a linker which is photoreactive and MS-cleavable as well, but cysteines can also be targeted with maleic acid imides (maleimides).[10–12] Another option is the photoreactive crosslinkers with either a diazirine or a benzophenone[13,14]. Diazirine reacts through a highly reactive carbene with all 20 amino acids, but do prefer the acidic amino acids.[15,16] The exclusively carboxyl-reactive crosslinkers such as 4-(4,6-dimethoxy-1,3,5-triazin-2-yl)-4-methylmorpholinium chloride (DMTMM) accompanied by a hydrazide have gained attention in the recent years because glutamic and aspartic acid are very often met in the protein composition.[17,18] Even though the data-interpretation is very tedious due to its high reactivity and secondary reactions, formaldehyde could also be used as linker.[19]

- Cleavable/Not Cleavable. In the case on uncleavable crosslinkers, the search-space of the possible peptide-peptide combinations increases quadratically with an increasing database. Not only that, but the chances of a random hit also become higher conducting eventually to a decreased confidence of the obtained results. Proteome wide studies also represent an impediment when the search-space increases quadratically. This so called n^2 can be successfully tackled by using cleavable crosslinks. This is one of the cases where the chemistry had to be adjusted in order to remove a bioinformatical bottleneck. An uncleavable linker would produce theoretically a single peak after which 2 peptides would have to be recognised. On the other hand, a cleavable crosslink would normally generate 4 peaks on the MS2 level, making the identification of the peptides pair algorithmically easier and with a better confidence.[20,21] In the figure below, there is an insightful exemplification of the signals produced in different types of fragmentation techniques normally employed in MS.

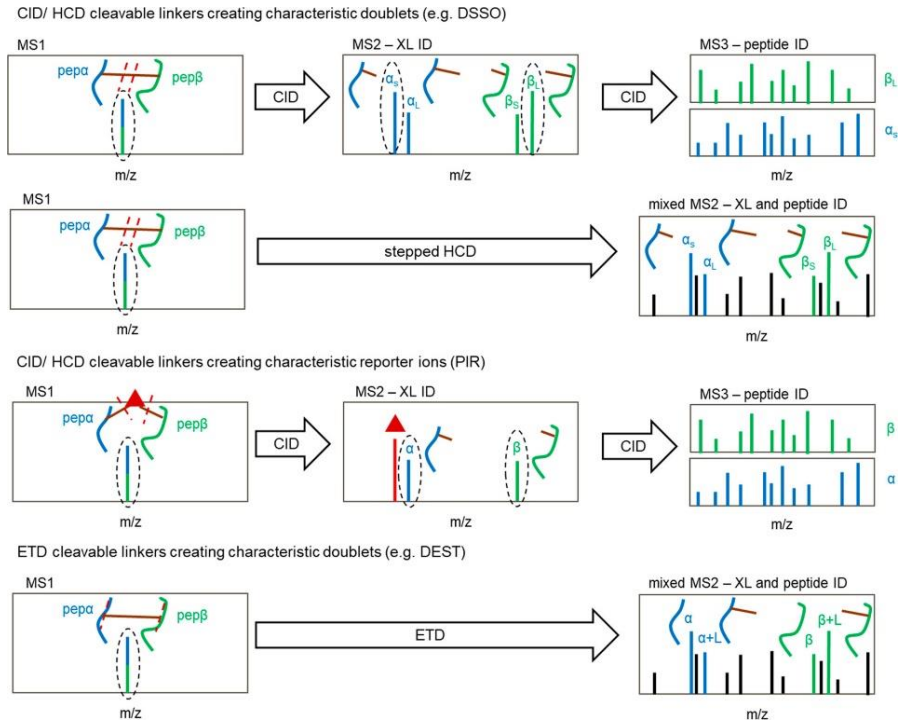


Figure 4: Fragmentation patterns of the cleavable crosslinkers upon different fragmentation techniques at different MS levels. In the case stepped HCD and ETD, MS3 is skipped, and peptides are annotated based on the 4 diagnostic ions present and the other peptide fragments present in the MS2 spectrum. For the case of PIR, a reporter ion of a known mass is

produced in the MS2, which points out to the existence of a linked peptide pair. The two peptide sequences are then selected and fragmented further for elucidating the succession of the amino acids. In the case of a linker such as DSSO, there is a specific difference in the signals pair which indicates a crosslink. Table adapted from [20]

The cleaving sites of the cleavable crosslinker that are normally positioned central in the spacer arm should have a lower fragmentation energy than the ones necessary to break the backbones of the peptides and proteins.

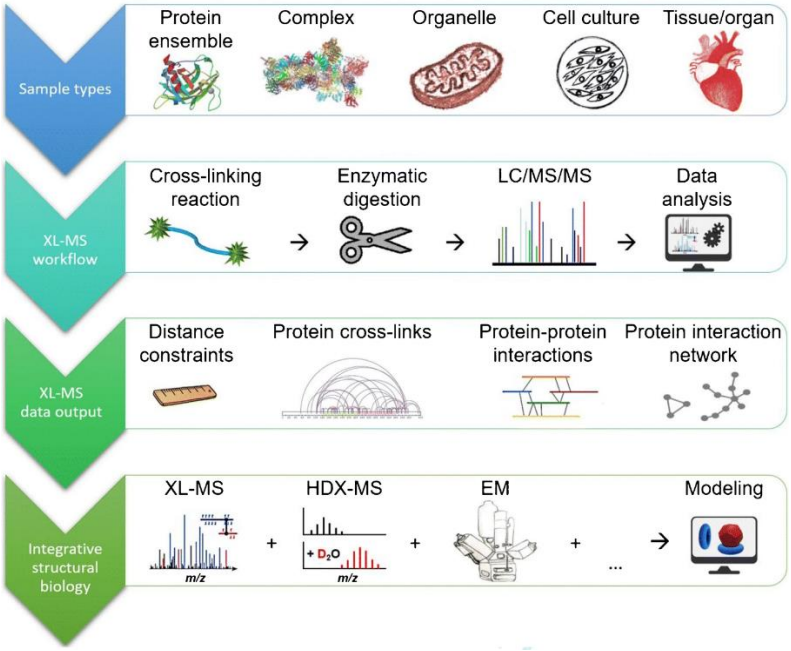


Figure 5: Descriptive Scheme of the XL-MS universe. Sample types comprise protein ensembles, complexes, organelles, cell cultures and tissues from different organs. The workflow is usually composed of the chemical linking reaction, digestion of proteins, MS-measurement, and data interpretation. The data output can be also very multifaceted (from distance constraints until interactions networks). XL-MS can be integrated with other techniques as well offering complementary information. Figure adapted from " Cross-linking/mass spectrometry at the crossroads" [22]

1.2. History of Protein Structure prediction

The first proteins structures were determined in the 1950s with the help of X-ray crystallography. It starts with a diffraction pattern obtained from the crystals. From the diffraction pattern, an electron density map can be obtained, which again constitutes the base for fitting an atomic model of the protein. The X-ray crystallography has the disadvantage that at typical resolutions hydrogen atoms cannot be identified, but just the heavy atoms of the proteins. This means that sometimes it may be difficult to identify all the hydrogen bonds present in the proteins, which contribute to the structure stability. Also, the so-called NQ flip (one letter code for asparagine and glutamine) can constitute a problem, where the 2 different conformation possibilities of the -$CONH_2$ cannot be differentiated. This happens because the two possibilities present similar electron density. Many pioneers of these field who readapted these technologies from the physics field were awarded the Nobel prize for their achievements. Max Perutz and John Kendrew received it in 1962 for determining the 3D structure of globular proteins (firstly myoglobin[23–25] followed by haemoglobin) with the help of X-ray crystallography.[26,27] Kurth Wüthrich obtained a shared Nobel Prize in 2002 for his research in structure and functions of proteins with the help of nuclear magnetic resonance.[28,29] This was a milestone not only because it offered a complementary method of the already established X-ray, but it also allowed

to solve the protein structure in its natural milieu, namely in solution. Last but not least, the Nobel for Cryo Electron Microscopy was awarded to Jacques Dubochet, Joachim Frank and Richard Henderson in 2017[30].

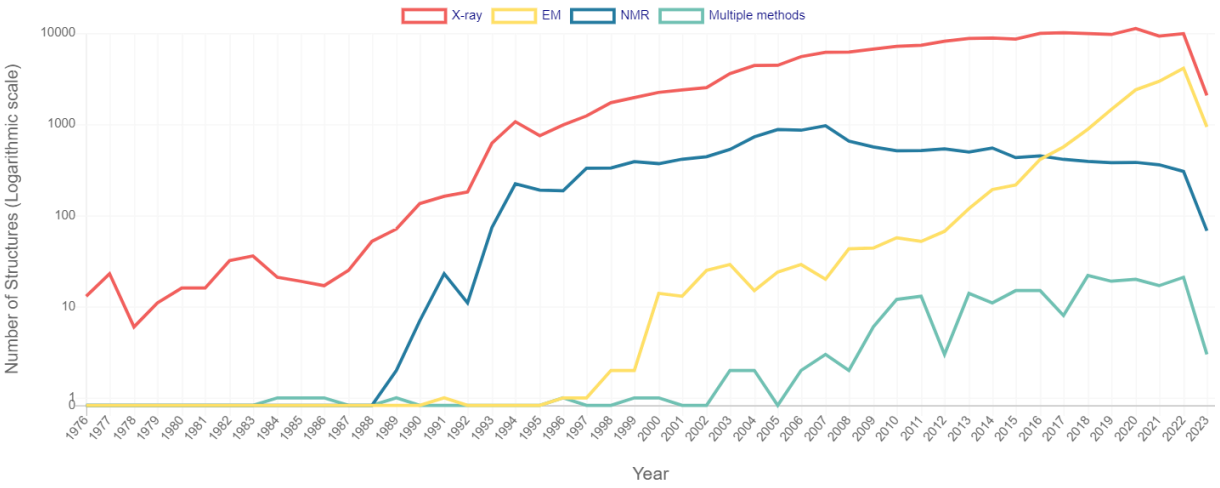


Figure 6: Number of structures on a log 10 scale acquired with different techniques per year. X-ray includes X-ray diffraction, fibre diffraction and powder diffraction. NMR comprises solution NMR and solid-state NMR. EM refers to electron microscopy, electron crystallography and electron tomograph. The term "Multiple methods" refers to more than one experimental method. If a structure was determined with both neutron diffraction and X-ray diffraction, it will be counted only once at "Multiple methods" section. Of note, for the year 2023 the number of structures is lower, because the statistic is based on the data acquired until the end of March and there is no reason to expect the number of new structures to decrease. Information and table adapted from [31]

X-ray remains up until 2022 the technique which produces the most PDB entries. It is followed by EM method which raises the number of structures produced each year. NMR has a slow decline, partly because it is mostly compatible with small structures. There is a lot of signal overlap, and the peak assessment can be a very time-consuming process. Also, NMRs can represent a big financial burden due to their high acquisition and maintaining costs.

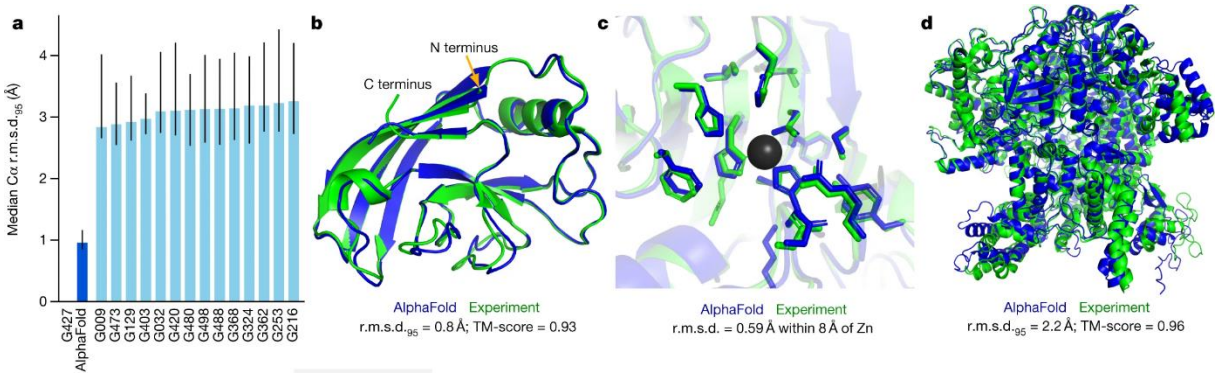


Figure 7: a) The performance of AlphaFold on the CASP14 dataset (n = 87 protein domains) relative to the top-15 entries (out of 146 entries), group numbers correspond to the numbers assigned to entrants by CASP. Data are median and the 95% confidence interval of the median, estimated from 10,000 bootstrap samples. b) AlphaFold prediction of CASP14 target T1049 (PDB 6Y4F, blue) compared with the true (experimentally determined) structure (green). Four residues in the C terminus of the crystal structure are B-factor outliers and are not depicted. c) CASP14 target T1056 (PDB 6YJ1). An example of a well-predicted zinc-binding site. d) CASP target T1044 (PDB 6VR4)—a 2,180-residue single chain—was predicted with correct domain packing (the prediction was made after CASP using AlphaFold without intervention). Figure adapted from [32]

CASP is the abbreviation for **C**ritical **A**ssessment of protein **S**tructure **P**rediction. It was first held in 1994 and it represents an important event world-wide in the research community from the field of structural biology. The event proceeds as follows: researchers send solved protein structures which

are still unpublished, but experimentally determined. Research groups then try to computationally predict the protein structure with different methods and algorithms. This "race" is a way to assess the weaknesses and strengths of each method by evaluating the predicted structure with respect to a series of parameters like RMSD, overall model quality, etc.[33,34]

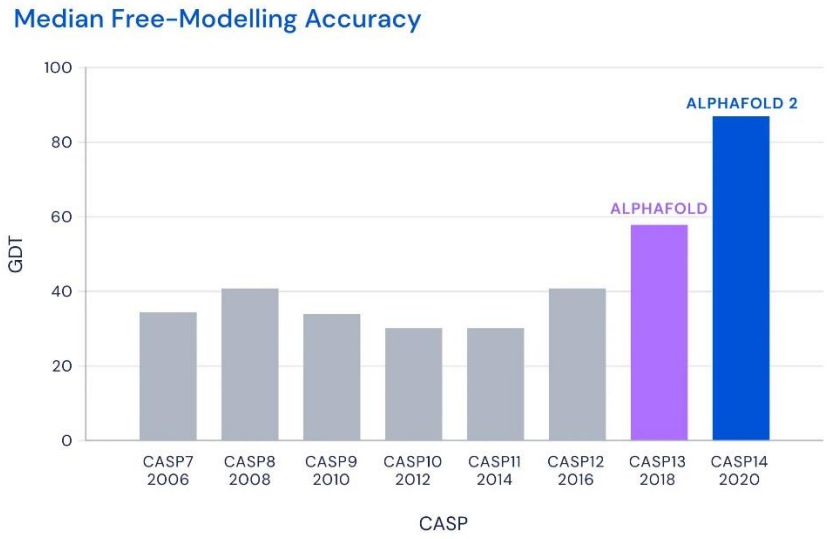


Figure 8:On the X-axis there are different CASP events represented, while on the Y-axis there is the median value for free-modelling of the best groups at each event where the GDT was measured. GDT (Global Distance Test) is a conventional measure for the accuracy of a predicted protein structure in relationship to the experimentally solved structure. GDT is a good orientation point of the overall matching instead of highlighting local errors (which can lead to a high RMSD). Let us assume that the following cut-offs are being used (1Å, 2Å, 4Å, 8Å). The predicted structure is then optimally superimposed so that as many residues as possible can be matched between the predicted structure and target structure. By "matched", one understands that the deviation of the predicted residue is within the cut-off value. Then, GDT is being calculated by averaging the percentages of matched residues for different cut-offs. Free modelling category represents the most challenging category because one cannot detect any structural templates by using the sequence. Ab initio models must be utilized.[35] *Table adapted from* [36]

Root-Mean-Square Deviation (RMSD) is a common measure in the field of bioinformatics to assess the average distance between the superimposed structures of a molecule and its reference. RMSD does not necessarily have to be used on a protein; it can also be used on any kind of organic molecule. It is calculated as follows:

$$RMSD\left(x, x^{ref}\right) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|x_i - x_i^{ref}\right|^2}$$

where x and x^{ref} are the coordinate arrays of the two structures, after optimal roto-translational superposition, and n the number of matches considered. Normally, not all the atoms are included in the analysis: the analysis is frequently restricted to the backbone atoms in a protein or only to the alpha carbon atoms. The contribution of a given atom to the RMSD can also be weighted by its respective mass.[37–39] In this case, the equation takes the following form:

$$RMSD(x, x^{ref}) = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n} w_i \left| x_i - x_i^{ref} \right|^2}{\sum_{i=1}^{n} w_i}}$$

The TM-score parameter is considered more robust than the RMSD because it also takes into account the topology of the protein and it is defined as follows.[40,41]

$$TM - score = max\left[\frac{1}{L_N}\sum_{i=1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right]$$

L_N is the number of residues in the native structure, L_T is the number of the aligned residues to the reference structure, d_i represents the distance of the ith match of the aligned residues, and d_0 is utilised to normalise the difference in distance of the pair. The TM-score can take values between 0 and 1.

There are free-to-use servers like Phyre2 (from 2015, or Phyre already present from 2009)[42,43], where the user can simply introduce the primary sequence of a protein, and a three-dimensional structure with annotated confidence intervals per region is being predicted with the help of advanced homology detection methods. The revolutionary solution proposed by Alphafold2 may also be considered, as it has proved immensely accurate and almost reaching the quality of experimental determination.

Now, that the three main experimental techniques for solving the proteins structures and the very potent computational tools were presented, a very valid question arises: why do we still need XL-MS?

System wide XL-MS studies are also possible. Until few years ago, protein-protein interactions (PPI) analysis at a large-scale analysis was considered rather a niche field. This is not the case anymore with the last evolutions of mass spectrometers and development of highly potent linkers. Being able to follow the PPIs, especially in their living cells, is highlighting the advantages of XL-MS. With this technique, weak non-covalent interactions can be frozen, as well as the contact area where the interaction is taking place, helping to elucidate the interaction mechanism and the dynamic changes during the transient interaction.[44,45]

System-wide studies were already performed on three different levels:

- cell lysate (from human cell lines[45,46])
- tissues (mouse brain[47] and heart[48])
- organelles (such as murine mitochondria[49], nucleus of HEK293[50])
- living cells (HEK293 cells[51])

Intrinsically disordered proteins (IDP) and intrinsically disordered regions (IDR), but also multi-conformational proteins represent maybe the aspect where the conventional methods are outperformed by the XL-MS, even though it provides only low-resolution information.[52] Through crosslinking, the protein is not forced to take a specific conformation in order to be analysed as in X-ray, but it "solidifies" the conformation the protein has at the moment. A representative example of an IDP is p53, the known tumor-supressor protein (so called "guardian of the genom")[53]. There is a study conducted on the p53 tetramer where apart from the valuable structural information, also

insights regarding the impact of the crosslink modification on the conformation of the IDP was gained. P53 maintained its DNA binding function by using DSBU, but had a complete function loss in the case of DMTMM.[54]

1.3. Crosslinking Mass Spectrometry - Data analysis

Recently, a multitude of crosslink search software were developed. They each address different aspects of the crosslinking matter. Some are specialized in uncleavable/cleavable links, need different data formats as input (e.g. mzML or raw), or utilize various search-algorithms. Apart from the XL-engines benchmarked in this work (pLink2, MS Annika, MeroX, XlinkX2), which are discussed in detail, some of the relevant XL-identifying software are mentioned in the table below.

XL Search Engine	Website
CRIMP[55]	https://www.msstudio.ca/downloads/general/
Kojak[56]	https://kojak-ms.systemsbiology.net/
MaxLynx[57]	https://maxquant.org/
XiSearch/xiFDR[21,58,59]	https://www.rappsilberlab.org/software/xisearch/
Formaldehyde XL Analyzer[60]	https://github.com/Kalisman-Lab/Search_Formaldehyde_Cross-links
MassAI (CrossWork)[61]	http://massai.dk/
SIM-XL[62]	http://patternlabforproteomics.org/sim-xl/

1.3.1. MS Annika

MS Annika cannot be utilised as a standalone software yet, but it is fully integrated in Thermo Proteome Discoverer (versions 2.3, 2.4 and 2.5), manufactured by Thermo Fischer Scientific. The software is able to identify crosslinked peptides from MS2 spectra and represents a viable choice in the experiments where cleavable crosslinking reagents are being used. Proteome wide analyses are also less time consuming as a consequence of fast and parallelized processing. The software architecture can be described as three interdependent nodes connected in series and an optional format adaptation node to allow visualization of protein interactions in a three-dimensional space.

The first node in the series is MS Annika Detector which has the task of differentiating between the spectra with crosslinked peptides from those without them.[63]

The cleavable crosslinks are in most cases symmetric and present two fragmentation sites. By design, the fragmentation sites are displaced in such a way that, after fragmentation, two chains with unequal lengths result (a light chain and a heavy chain). As a result, one peptide is bounded to the heavy chain and the other to the lighter chain. Since fragmentation not always occurrs at the same side of the linker, in theory, one should see in a spectrum the same peptide attached to both the heavy and the light chain of the linker. This case, where two doublets are present in the spectrum, is seen as ideal, though not always the case. MS Annika Detector is equipped with 3 modii capabilities. In the evidence mode, at least one doublet and one ion must be detected in order to categorise the spectrum as crosslink-containing. That means that both peptides must be present. On the other hand, the indication mode is less demanding. The finding of one doublet is sufficient and the other peptide's mass will be deduced from the precursor mass and the already identified doublet. The third option would be to use the evidence and indication mode simultaneously in a so-called combined mode. As a result, spectra with complete and incomplete doublets will be identified. Another implemented feature is the search for diagnostic ions. They are a strong supplementary piece of evidence that the respective spectra are in fact containing crosslinked peptides. The diagnostic ions result from broken

linkers that are not bounded to a peptide. This event can occur because of the specific design of the linker or because the fragmentation occurs where the peptide is attached instead of the intended fragmentation bond. MS Annika Search, the second node, determines the crosslinked peptides from the matched spectra. Here, an extension of MS Amanda is used to find the two individual peptides in the spectrum independent of each other. For each peptide mass, the top N (defined by the user) candidates are being chosen. The peptide pairs are formed by combining each of the N candidates from the first peptide mass with the N candidates from the second peptide mass. The peptide pairs possess a combined score, AnnikaScore, which is calculated as:

$$AnnikaScore(s, pep1, pep2) = \min\left(AmandaScore(s, pep1), AmandaScore(s, pep2)\right)$$

In the equation given above *s* represents the matched spectrum, *pep1* and *pep2* the two peptide candidates. In other words, AnnikaScore is the minimum of the individually calculated AmandaScore values for each peptide.

Intuitively, but important to be underlined, is the fact that only peptide pairs that fulfil the following relation are regarded:

$$abs(mass_p - \big(mass(pep1) + mass(pep2) + mass(XL)\big) \leq T_D$$

$mass_p$ is the mass of the uncharged precursor, *mass(XL)* represents the mass of the crosslinker, and T_D stands for detection tolerance.

Once the spectrum has undertaken the workflow procedure and a viable peptide pair was found, one can talk about a CSM (crosslink spectrum match). The highest scoring CSM is selected and saved for a succeeding validation step.

The last in series node is MS Annika Validator and its job is to validate the results at CSM and crosslink level. It must be considered that the same peptide pair can appear in more than one spectrum. Additionally, in some cases, one peptide can be a substring of a longer peptide. They are in fact two different peptides but denote the same position in the protein. MS Annika Validator eases this ambivalence by resuming these CSMs to one single crosslink. When merging multiple CSMs into a single crosslink, the score attributed to the crosslink is the highest found CSM score. A higher number of CSMs for a link is seen in the field as clue to a better confidence in the obtained result. For the FDR calculations, a target-decoy is used. It is possible to set the wanted confidence level not just on CSM and crosslink level, but also to differentiate between interlinks and Intralinks. Intralinks are peptide pairs coming from the same protein and, in the case of interlinks, they originate from two different proteins.[63]

1.3.2. pLink 2

pLink 2 represents a rapid crosslink search-engine and an adequate fit for proteome scale identifications. The architecture of the algorithm tackles a multitude of problems in crosslinking domain and tries to improve the running times and quality of results by applying the latest findings in this field and integrating them in the software code. For a given number of crosslink-containing spectra, M, the matching of peptide pairs form N possible peptides would lead to a $O(MN^2)$ in case of a complete exploration of the proteomic space. The software reduces the complexity to O(MN) by searching for peptides individually. Studies have shown that, in the case of two bounded peptides, one of them usually fragments better than other.[64,65] For the first peptide, the α-peptide, the top 5 best scored candidates are chosen. This peptide is usually larger and fragments better. In the case of the second peptide, the β-peptide, the search engine calculates the mass differences by subtracting the crosslinker mass and the mass of the α-peptide candidates from the precursor mass and lets only the

mass fitting candidates pass. These are generally few when the mass-spectrometers present high-accuracy.

Firstly, the general workflow of pLink2 should be explained before going into detail. pParse analyses the MS1 spectra and filters the precursor candidates. Additionally, in each corresponding MS2 spectrum, candidates for α-peptide are being retrieved. In the third step, the compatible β-peptide candidates are calculated. During the fourth step, the prospect peptides are paired and fine-scored with the spectrum. Finally, all PSM (peptide spectrum match) are ranked again and filtered.

Some more elaborated working mechanisms for the presented workflow are important to be underlined. In the case of α-peptide candidate retrieval, the peaks recognised in the MS2 spectrum are converted into b and y ions for the query of fragment index. In order for a peptide to be coarse scored, it has to present at least two matched ions. From the list of coarse scored peptides, five of them with the highest coarse scoring are selected for further investigations. Differently, the β-peptide candidates are searched by the peptide index. After the fine-scoring of the peptide pair, the software searches for the mono-linked, loop-linked and unreacted peptides as well. For a specific MS2 spectrum, it will then be decided if a modified/unmodified peptide or crosslink will be further selected based on their score. A triage is then realised into either interlinks, intralinks, monolinks, looplinks or linear peptides to guarantee a conform machine learning and a proper FDR-check.[66]

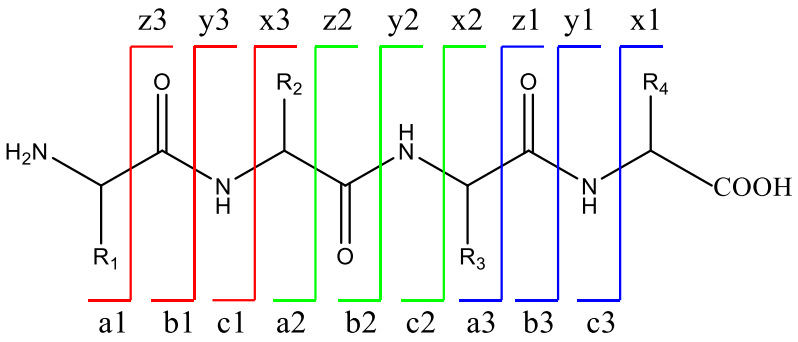


Figure 9: Standardized nomenclature for ions in peptide- mass spectrometry. Figure was drawn with ChemDraw 18.1

Already in 1984, a nomenclature for sequence ions of peptides in mass spectrometry was proposed. At that time, the necessity for a common nomenclature appeared due to peptide structure determination experiments. The nomenclature is independent of the ionization method applied.

If the charge is kept at the N-terminus fragment of the peptide, the three potential cleavage points of the peptide backbone are named *a*, *b*, *c*. In case the charge is retained on the side of C-terminus fragment, the cleavage points are called *x*, *y*, *z*. The numbering of the letters keeps track of which peptide bond is cleaving counting from N terminal-, respectively from the C terminal fragment. It also denotes the number of amino acids in the respective fragment.[67]

1.3.3. XlinkX v2.0

XlinkX 2.0, the successor of XlinkX, is another variant of crosslinked peptides search engine compatible and fully integrated in Proteome Discoverer software. It adopted an innovative algorithmic strategy with its own perspective in two decisive key points. As Orbitrap Fusion/Lumos instruments emerged

to the market, the presence of XlinkX v2.0 became helpful because it allows a crosslink search on both MS2 and MS3 levels.

Different than XlinkX, when XlinkX v2.0 executes the precursor mass determination, it combines Δm-based precursor mass determination approach with an intensity-based approach. In the first strategy, the precursor mass of both linked peptides is being deducted from the masses of the whole crosslink from MS1 and only one of the signature peaks (one ion of the two doublets from the CID-MS2 spectrum). Previously, the presence of all four signature peaks was needed (in other words, too much quantity was traded for a questionably higher quality of results). Intuitively, but worth mentioning, this is done by applying the following formulas:

$$\Delta m = m_L - m_S = m_{\alpha-L} - m_{\alpha-S} = m_{\beta-L} - m_{\beta-S}$$

$$m_{precursor} = m_{\alpha-L} + m_{\beta-S} = m_{\alpha-S} + m_{\beta-L}$$

$$m_\alpha = m_{\alpha-L} - m_L = m_{\alpha-S} - m_S$$

$$m_\beta = m_{\beta-L} - m_L = m_{\beta-S} - m_S$$

In the equations place above, α and β are the two crosslinked peptides, *S* and *L* stand for the shorter and longer linker chain of the linker after fragmentation.

In the complementary approach, the *n* most intense peaks in the CID-MS2 spectrum are chosen. *n* is a parameter defined by the user. Through this, the knowledge that crosslinker fragmentation will most probably generate product ions with high intensity is incorporated into the algorithm. Additionally, the masses of the peptides are deduced as:

$$m_\alpha = m_h - m_L \; or \; m_\alpha = m_h - m_S$$

$$m_\beta = m_{precursor} - m_\alpha - m_L - m_S$$

Where *h* is the one of the top *n* fragment ions in the MS2 spectrum. Of course, if the fragment ions chosen is not a signature peak, the above represented equations would deliver false results for both peptides, but the usage of a strict cut-off score for these fragment ions filters out these cases.

The other relevant feature that distinguishes XlinkX v2.0 from other peptide pairs search engines is the ability to work with data from a MS2-MS3 fragmentation strategy. During the matching step of the product ion, the theoretically generated fragment ions from the peptide candidates are firstly compared with the MS2 data (from both CID and ETD if accessible) and then with the mass difference-triggered MS3 spectra. For the confidence calculations of the peptide candidates, the software applies a probability scoring system. It is calculated as:

$$n_{score} = \left(1 - \sum_{i=0}^{n-1} \frac{e^{-xf}(xf^i)}{i!}\right) * N$$

In the equation above *n* represents the number of matching fragments, *x* the probability of an existent fragment ion to match a random theoretical fragment ion of peptide candidates, *f* stands for the number of all theoretical fragment ions of the crosslinked peptides and *N* the total number of proteins in the provided database.

The score utilizes the Poisson distribution in an intelligent manner: it was initially thought to determine the probability of a random protein match and thus the specificity of a search engine.[68] The random match probability summed over all matching fragments is subtracted from one (100%) and finally multiplied by total number of proteins to obtain the score. If the random match probability is low, the fragment ion match has higher chances to be true and therefore a higher score. The n-score for calculated

from the MS3 spectrum has a supplementary restriction: it needs to be higher than a user-defined cut-off score in order to be considered.

The final score for a peptide candidate is then obtained by multiplying the MS2 and MS3 scores:

$$n_{score}(\alpha) = n_{score}(MS2\alpha) * n_{score}(MS3\alpha)$$

$$n_{score}(\beta) = n_{score}(MS2\beta) * n_{score}(MS3\beta)$$

In the end, XlinkX v2.0 also employs a target decoy strategy for the FDR check.[69]

1.3.4. MeroX

MeroX is a standalone software, free to use and possesses a user-friendly graphical interface. This crosslink search engine is specialized in cleavable crosslinker and demonstrates high quality results. Its algorithmic design is beautifully and extensively explained by the authors and has a high versatility and modii operandi depending on the size, data, and expected quality of data.

Firstly, a fasta file with protein sequences needs to be provided. The amino acid sequences are then imported with regard to the static modifications an amino acid can suffer. For example, if cysteine is present in an alkylated form, MeroX will replace 'C' with 'B' in all strings of the proteins from the fasta file. For the generation of the decoy data base, all sequences are being shuffled, except for the positions of the residues where the cleavage takes place. For every protein the shuffling is seeded. This is crucial, as it keeps the variability of the results under control. Otherwise, the analysis of the same input data under the same conditions could lead to qualitative and quantitative differences. The peptides resulted from the artificially created decoy file are marked as wrong/decoy and appended to the peptide data base. An *in-silico*-proteolysis step is undertaken, and all possible peptides are generated from the protein sequences by cleaving the proteolysis site and taking into account the number of maximum missed cleavages. These are variables given by the user. Every created peptide sequence is saved and accompanied by information about the protein ID and location inside the primary sequence. In case of identical strings, only one string is saved together with positional information of two (or more) peptides. The maximum possible number of variable modifications and types of variable modifications considered are also user-defined. These modifications are implemented in all peptides and therefore enriching the peptide data base.

Because of its various capabilities, the vigorous discussion about MS data analysis will be restricted to the 'RISE' mode. There are four different modii operandi that deserve to be enunciated. For purified proteins or a protein mix of low complexity (up to 10 proteins), the 'quadratic mode' would be the recommended version. Because the algorithmic variant does not depend upon reporter ion in the fragment ion spectra, the 'quadratic mode' is compatible with non-MS-cleavable crosslinkers as well. 'RISE' and 'RISEUP' can sustain fasta files with up to 500 protein entries. RISEUP, the advanced form of RISE, is capable to deduce peptide mass from incomplete reporter ion patterns. The "proteome wide" mode is able to support many thousands of proteins in the data base and is conditioned just by the existence of reporter ion patterns for one of the two peptides at minimum.

To exemplify the workflow, the functioning steps of RISE method are described below.

The MS data, present in different format types (mgf, mzXML, mzML, pkl) is imported. Each spectrum is individually processed. The noise is eliminated by breaking the spectrum into maximally 10 bins of maximally 50 signals. For every bin, the noise is determined iteratively until the desired signal-to-noise ratio is obtained. Furthermore, the MS spectrum is deisotoped and the charge states are saved. This step is followed by deconvolution. Through deconvolution, a mathematical improvement of the resolution of the spectra is reached. A broad signal formed by overlapping several narrow signals can be resolved

by this processing function so that the positions of the individual signals can be determined.[70,71] If spectrum signal does not have a known charge state, all possible variants are being considered. The obtained mass list (in ascending order) is then analysed for corresponding crosslinker-fragment ion pairs. Subsequently, the discovered mass pairs are then searched in the peptide data base. The peptide pairs candidates that match the precursor mass in the user-defined precision interval are selected and scored.

During the scoring step, the theoretical fragment ions of a peptide pair candidate are generated and compared to the spectrum and to four other spectra, which have minimally shifted masses. The following characteristics are calculated:

- number of found signals per spectrum
- number of identified fragment ions per peptide pair candidate vs. its theoretical fragment number
- signal intensity of crosslinker fragment ions in the spectrum
- how many and how long the peptide backbone ion series are

In each of these characteristics, the difference between the result from the real spectrum $f_{n,0}$ and the result for the mean of the four shifted spectra $\bar{f}$ together with their standard deviation $\sigma_n$is computed.

$$x_n = f_{n,0} - (\bar{f} + \sigma_n)$$

x_n serves for calculating the probability for a specific characteristic to be a random match. The score is obtained after further mathematical processing and attributed to the candidates. The scoring step is done for all possible crosslinking sites as well (if this is the case). The crosslinking site with the highest probability is saved.

The FDR estimation is calculated as:

$$FDR = \frac{N_{decoy}}{3 * N_{target}}$$

N_{decoy} is number of decoy candidates above a certain score cut-off and N_{target} the number of valid candidates above a certain score cut-off.[72]

The software-infrastructure was also further expanded to allow visualization of the crosslinks at different complexity levels, protein interaction networks etc. Some of the computer programs available along with their field of application are shown below.

Visualisation Software/Website	Website	Utility
xiView[73]	https://xiview.org/xiNET_website/index.php	Main view, circular view, multiple proteins interactions, 3D NGL view, matrix, histograms, spectrum view
ProXL[74]	https://proxl-ms.org/	Circle plot view, protein bar view, 3D protein structure view, storing, viewing, analysing and sharing XL-MS data.
Jwalk/MNXL[75]	http://jwalk.ismb.lon.ac.uk/jwalk/	Choosing protein structure models constructed with different templates based on the crosslink restraints
TopoLink[76]	http://leandro.iqm.unicamp.br/topolink/home.shtml	Determination of the reliability of a structural model based on the experimental XL results

1.4. Elucidating FDR values

In a conventional run, prior to the analysis, the user defines the False Discovery Rate (FDR) of the results. The values that have prevailed and are seen as relevant are 1% (high confidence) and 5% (medium confidence). That means that the results reported by the software are statistically considered containing 1%, respectively 5% just random hits.

A widespread method for determining the FDR value is target-decoy approach[77]. The approach involves searching for crosslinks in two different databases. One is the true database, that is normally being provided by the user and the decoy database is made by either randomizing or shuffling the sequences from the target database. Of note, no sequence from the decoy data base exists in the target-database.

Throughout the search, each crosslink and decoy crosslink get scored in the same manner. After the scoring phase is finished, and both crosslinks and decoy crosslinks are ordered depending on their annotated score, the score-threshold is adapted such that only the wanted percentage of decoy-links is present. In the presentation of the results, they are of course eliminated as well, but it is assumed that the same percentage of random target hits is still present from the target database and was not removed.

The latest improvement in FDR-assessment was brought by Beveridge et al. In their study, a truly verified reality could be established by creating a peptide library which is capable of crosslinking, but only within groups and not because of actual interactions, but random interactions forced by the higher concentration. The workflow is elegant and non-discriminatory with respect to any peptide sequence.[78] Inspired by the ground-breaking research in the FDR- (re)calculation, it became evident that further expansions are necessary. The design of peptides, which includes a biotin moiety for proper elimination of unwanted peptide fragments, mimics the structure of a protein because of its -YGGGGR- sequence, which is afterwards subject to digestion. The tyrosine is built in here for quantification purposes. Also, extensive data analysis was performed, but it lacked any software support for streamlining a higher number of data sets. In the following chapter, it is described how this software void is being covered and which type of analyses are automated by the IMP-X-FDR. Different dependencies implying number of found links, the FDR values and the scores of the XLs are automatically calculated and displayed in descriptive graphs.

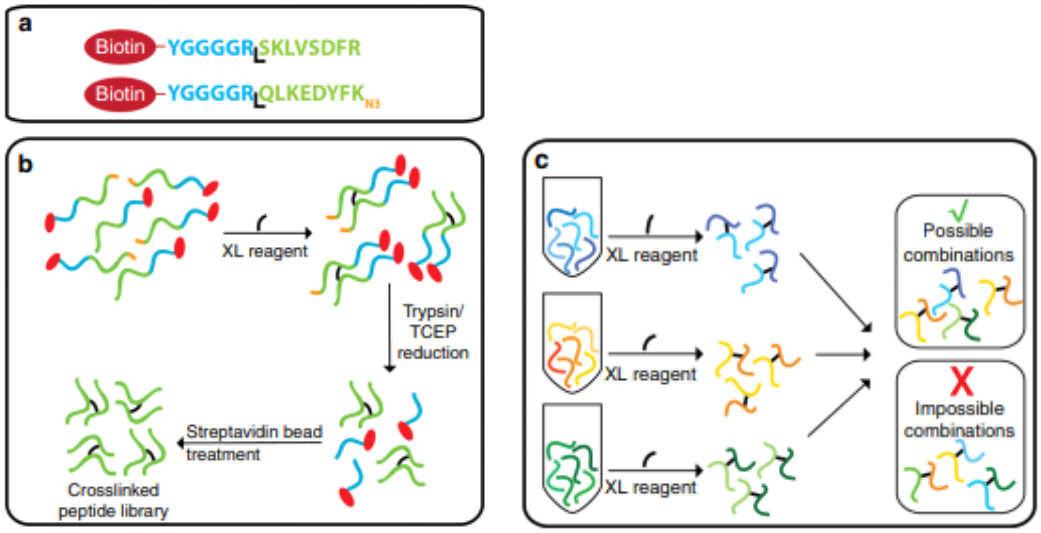


Figure 10: Subfigure a) shows the architecture of individual synthetic peptides; subfigure b) *presents the XL-workflow including the linking reaction, protein-like digestion with trypsin and azide group reduction followed by removal of biotin*

containing groups; subfigure c) *displays the logic of deciding which peptides correct and which are false. Table adapted from* [78]

2. Materials and Methods

All data analysis performed was done with the help of IMP-X-FDR software which was written in Python and tailored with a graphical user interface in C#. Throughout this chapter, the software architectural structure and technical details will be explained and exemplified.

2.1. Installation

In order for the application to function, the following software must be installed:

a) .NET 6 Desktop Runtime (or higher)
b) Python 3.9 (or higher)

The latest release of the software should be downloaded as it will likely be updated and improved over time. After unpacking the archive, IMP_X_FDR.exe should be run. In case of missing programs such as Python, the user will be automatically directed to the respective installation website. A suitable version for the user's computer should be additionally installed.

After opening IMP-X-FDR exe via a GUI (Graphical User Interface) as shown in the figure below, it is allowed to load (filtered) crosslink sequence match (CSM) or unique crosslink (XL) files as input. A corresponding library file that fits the used dataset should be set. Pre-set libraries can be found under resources/libraries in the installation folder. If CSM containing files are used as input, they can be grouped to XLs by IMP-X-FDR by ticking "Group crosslink spectrum matches to unique crosslinks". In case XL lists are used as input, this tick-box can be ignored as it does not have any effect.

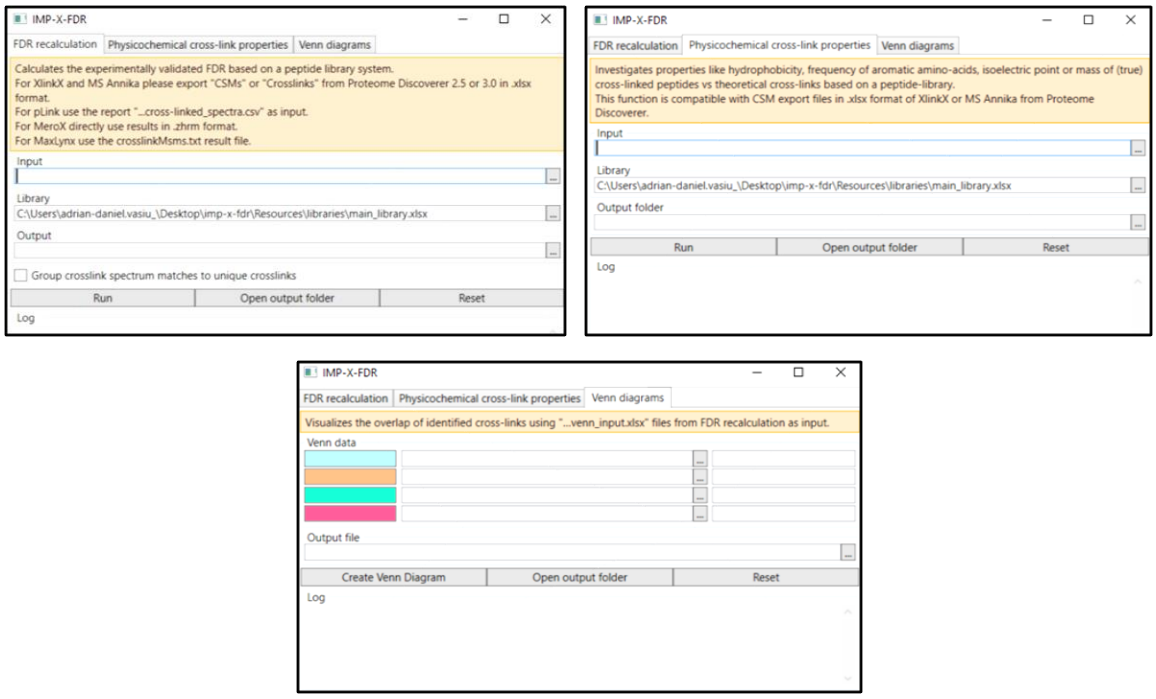


Figure 11:Detailed graphical user interface of IMP-X-FDR

2.2. Structure of the code

The whole code can be found in in the directory called “Resources” present in the installation folder.

The scripts utilized for FDR recalculations and their respective additional analyses can be found in the “search-engines” subdirectory, where a unique script was adapted for each XL-MS software to correspond to their format differences.

In the “libraries” subdirectory, a collection of support libraries utilized in the conducted experiments (in both digested and undigested form) and the Beveridge et al. Library of Cas9 peptides. The directory can be expanded as well with new support libraries.

“chemical-investigations-annika.py” and “venny.py” can be found directly in the “Resources” directory and are called in the case that “Physicochemical cross-link properties” or “Venn diagrams” options are chosen.

The file “requirements.txt” contains all the necessary Python libraries for the scripts to function properly that will be automatically installed.

2.3. FDR-recalculation

The FDR-recalculation function serves the purpose of reassigning the true FDR value of the measured peptide library system. This is possible because it possesses the information about the allocation of peptides to the corresponding groups (which the XL-search engines do not have at disposal).

Input-files

If “MS Annika” or “XlinkX” were chosen, the input file needed for the “FDR-recalculation” can be generated from Proteome Discoverer by exporting the (filtered) crosslink or CSM list to Microsoft Excel.

In case of using MeroX, a “.zhrm” result file will be used as input file. The software will temporarily convert it into a “.zip” data, extract the necessary information and then bring the file to its initial form. The script was created for version 2.0.1.4 of MeroX and it is recommended that input files come from this version of the software, otherwise the results may be incomplete or not analyzable. This script can also be run from the command prompt, where the user can give two optional arguments:

```
optional arguments:
  -h, --help            show this help message and exit
  -sintra SCORE_INTRALINK, --score_intralink SCORE_INTRALINK
                        introduce the intraprotein XL cut-off score for the selected FDR.
 It can be found in -Show decoy analysis-
  -sinter SCORE_INTERLINK, --score_interlink SCORE_INTERLINK
                        introduce the interprotein XL cut-off score for the selected FDR.
 It can be found in -Show decoy analysis-
```

Figure 12: Optional arguments in Merox data analysis refer to Decoy analysis performed by Merox to see scores for specific FDR

In some cases, the number of crosslinks delivered by the IMP-X-FDR does not match the number shown by the software (if it happens, it normally differs by 1-2 crosslinks and has a negligible impact on the overall analysis). This is because Merox performs an extra filtering step on the data displayed on its GUI, but does not impact the initial result file. The extra-filtering can also be performed by “merox_master_score.py” script when used via command prompt and adding the optional arguments.

The advantage of this feature is that the user is able to run several FDR-Check searches by just changing the intra- and interscore cut-off value without having to perform multiple Merox searches by changing the FDR value.

The figure below shows where the cut-off values of the MeroX search with a defined FDR of 1% can be found:

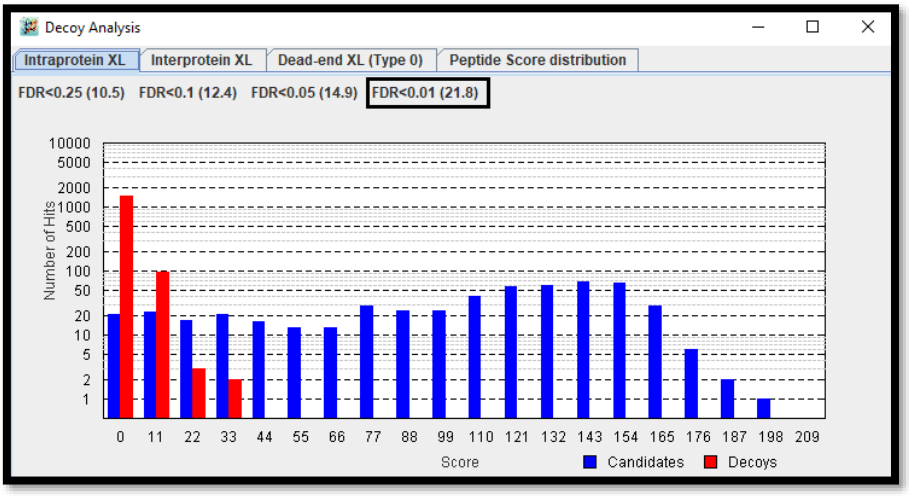


Figure 13: Cut-off scores for various FDR values, from Merox (Iacobucci, C et. al. Nat. Prot.., 2018)[72]

The needed file from pLink2 searches for the FDR-Check script is found in the pLink 2 output subdirectory "reports" with the default extension "filtered_crosslinked_spectra.csv". For the later incorporated "MaxLynx/MaxQuant" searches, the "crosslinkMsms.txt" output has been made compatible for IMP-X-FDR. Lastly, in the case of using XiSearch, the crosslink table from xiFDR can be exported as csv data and uploaded into the IMP-X-FDR.

Support libraries

The support library files for the published datasets coming from synthetic ribosomal peptides (peptide library 1-3)[79] and previously published Cas9 data[78] are provided within the resources/libraries folder. Additionally, their undigested form as well as the variants where the peptides are not divided into pools are also at disposal. Support files are necessary to guide IMP-X-FDR in assigning IDs as correct (within the same group) and incorrect crosslinks (connected peptides within different groups).

Other peptide library systems can be utilized as well. The structure and the extension of the document (.xlsx) must be kept consistent in case of developing another synthetic peptide library as indicated in Figure 13. An interesting use of IMP-X-FDR would be the conduct of the same type of FDR-check or physicochemical analysis even when using intact proteins or protein complexes. The proteins can be "*in silico* digested" and the same type of support library files can be generated. There are many options of simulating the digestion of a protein or a protein mix.[80–82]

	A	B
1	Group1	x
2	VALVAKIGENINIR	5
3	EIAEKMVEGR	4
4	KAGNVAADGVIK	0
5	EFIAKLQANPAK	4
6	MAALMKQR	5
7	ILKCGFR	2
8	VKDLPGVR	1
9	MAHIEKQAGELQEK	5
10	QAGELQEKLIAVNR	7
11	EVPAAIQKAMEK	7
12	Group2	x
13	VKGGFTVELNGIR	1
14	ITDVEVLKAQFEEER	7
15	TGAGMMDCKK	8
16	KAGFVTR	0
17	HKATLLGLGLR	1
18	MAKTIK	2
19	VKHPSEIVNVGDEITVK	1
20	ITLNMGVGEAIADKK	13
21	QCKANPWQQFAETHNK	2
22	ANPWQQFAETHNKGDR	12

Figure 14: Structure and outlook of the library support file

The peptide sequences contained in the first group are followed directly (without any empty cells) by the next groups. The numbers are only mandatory if a physicochemical investigation is performed and indicate the position of the crosslinker-binding amino acid in the peptide sequence minus 1 (because of list indexing in Python).

Output files

After the analysis of the IMP-X-FDR is complete, the following file will be created:

1.**"[name]_output.csv"**

The main output data is a csv document, which returns a detailed list of the crosslinks containing information about their sequences, theoretical origin-proteins, position in protein, their score and whether the crosslinks are formed within the same group (considered as true) or not (considered as false).

Sequence A	Sequence B	Accession A	Accession B	Position in protein A	Position in protein B	Score crosslink	Within same group
KSSAAR	TSGEKHLR	P0A7X3	P0A7N4	13	37	189.06	TRUE
EKLQER	KQQGHR	P0A6F5	P0AG48	364	85	184.14	TRUE
KDIHPK	KSSAAR	P0A7M9	P0A7X3	3	13	173.44	TRUE
KQQGHR	SKYGVK	P0AG48	P0A7S3	85	116	224.58	TRUE
KQQGHR	MAVQQNKPT	P0AG48	P0A7N4	85	7	193.84	TRUE
AMEKAR	TSGEKHLR	P0A7W1	P0A7N4	66	37	216.67	TRUE
KSSAAR	MAKTIK	P0A7X3	P0AG51	13	3	160.36	TRUE
HIGGGHKQA'	QPHAKGR	P60422	P25888	59	71	194.88	TRUE
AMEKAR	SGAIKAAK	P0A7W1	P0A6P1	66	48	204.26	TRUE
GEVKGK	HIGGGHKQAY	P0ABB0	P60422	318	59	206.21	TRUE
AMEKAR	ELKPHDR	P0A7W1	P0A9Q1	66	195	148.43	TRUE
TSGEKHLR	TSGEKHLR	P0A7N4	P0A7N4	37	37	168.81	TRUE

Figure 15: Structure of main "output" file

2.**"[name]_Number-XL.svg" (Fig. 16A)**

Case 1: Real FDR is above 5%:
First bar shows the type of crosslinks in the results delivered by the XL-search engines together with the real FDR. The second and third bars are obtained by filtering the results until the FDR reaches or is less than or equal to 5% and 1%.

Case 2: Real FDR is less than 5%, but greater than 1%:

First bar shows the type of crosslinks in the results delivered by the XL-search engines together with the real FDR. The third bar is obtained by filtering the results until the FDR reaches less than or equal to 1%.

Case 3: Real FDR is less than 1%:
In this case there will be only one bar displaying the real FDR without any filtering.

3.**"[name]_FDR-at-Score.svg" (Fig. 16B)**

Case 1: real FDR is above 5%:
First point: lowest accepted score, real FDR
Second point: the lowest score at which the real FDR is less or equal to 5%
Third point: the lowest score at which the real FDR is less or equal to 1%

Case 2: real FDR is under 5% and above 1%
First point: lowest accepted score, real FDR
Second point: the lowest score at which the real FDR is less or equal to 1%

Case 3: real FDR is under 1%
First and only point: lowest accepted score, real FDR

4.**"[name]_Score-vs-Number.svg" (Fig. 16C)** shows the distribution of the crosslinks depending on the score.

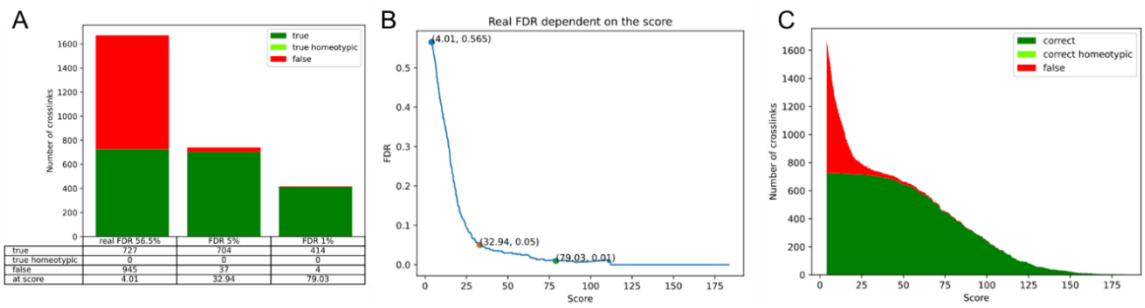


Figure 16: A) number of unique crosslink sites as given by the XL-search engine and of crosslink sites post-score cut-off in order to achieve a real FDR of 5% and/or 1% respectively B) real FDR-value development depending on the cut-off score C) XL sites and their nature (correct or incorrect) depending on the cut-off score. This prime example was chosen to demonstrate that unsupervised XL-search engines can produce results of a very low quality by setting a too low cut-off score. The loss of true crosslinks is negligible in the post-filtering step from an FDR-value of 56.5% to 5%.

5.**"[name]_output_venn_input.xlsx"**

It is used as input for the Venn diagram option and overcomes the different XL output formats by distinct search engines.

2.4. Venn Diagrams

The function of "Venn Diagrams" facilitates the comparison of found XLs, correct XLs and false XLs (previously categorized by the "FDR-recalculation").

To generate a uniform data structure of identified cross links across different search engines, the "[name]_venn_input.xlsx" file has been implemented which is an output of "FDR-recalculation". The script selects common features of crosslinks that are present in each software (i.e. position of XL in protein A and B, sequence A and B and protein name A and B) and sorts them alphabetically. The score is not considered by the Venn script, as each software has different methods of determining the quality of a crosslink.

When comparing more than one result, Venn diagrams for all, "true" or "false" XLs together with more detailed information is saved in a ".xlsx". Weighted Venn diagrams are generated for a maximum of three result files, while when comparing 4 results an unweighted Venn diagram will be displayed. The colours can also be chosen by the user. If no colours are selected, the default ones will be displayed.

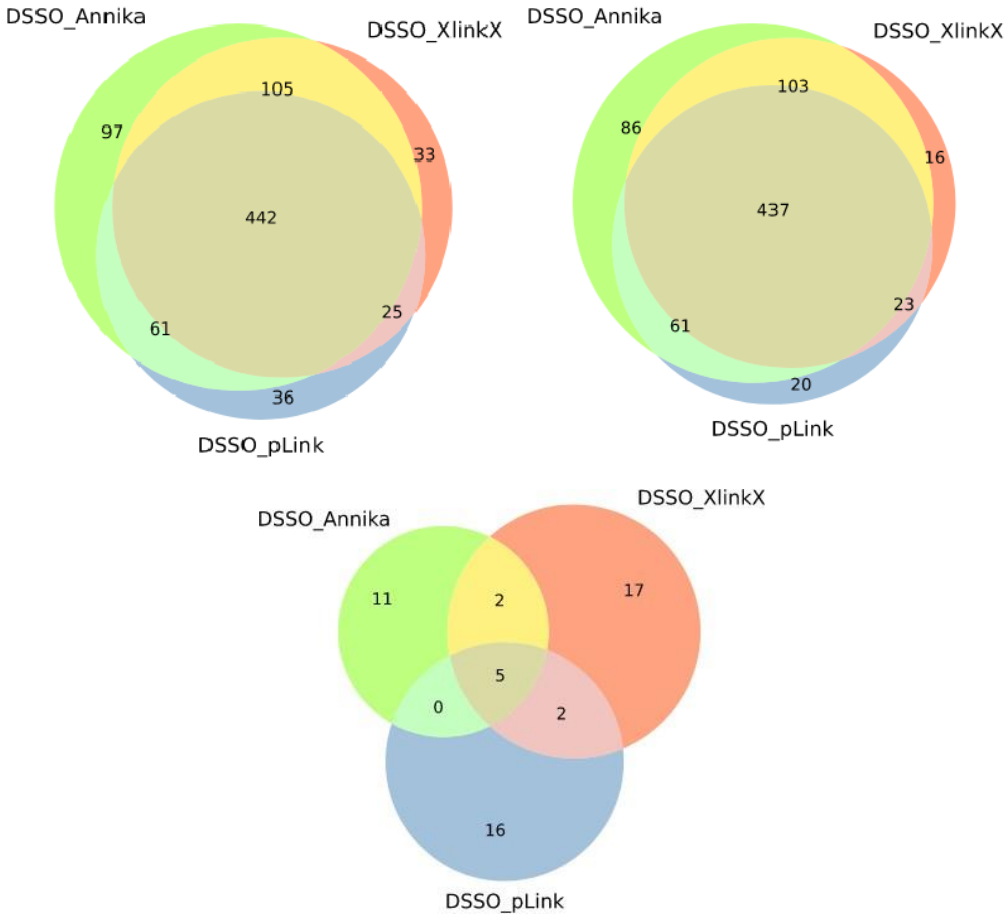


Figure 17: Venn diagrams showing the results from Annika, XlinkX 2 and pLink 2; on upper left side are all the crosslinks, upper right only the crosslinks within the same group and on the lower side the overlapping of the "false" crosslinks is represented.

Additionally, a document that includes three different sheets for crosslinks categorized as all, "true" and "false" will be produced. If the overlap of specific crosslinks should be investigated or, for example, common features of false crosslinks are identified, the user has access to the lists of the crosslinks present in each area within the mentioned document.

2.5. Physicochemical properties

All properties were calculated by using the Bio.SeqUtils package from Biopython 1.79[83](except for the frequency of the neighbors). The code was adapted for crosslinks, but, except for the mass determination, the crosslinker itself was not considered when calculating the properties. The frequency of the nearest neighbors to the binding amino acid was calculated by using the package "seqlogo".[84]

Although biopython features many functions, such as "charge_at_pH", "secondary_structure fraction", "molar_extinction_coefficient" or "instability_index", we decided to exclude those from our analysis package. A secondary structure analysis would be inaccurate in the case of the synthetic peptide library because they are mostly too short to converge to any secondary structure and remain linear. Nevertheless, it is worth mentioning that we initially incorporated the function “charge_at_pH” in our script, but later eliminated it, because it proved unreliable and necessitated a continuous fine tuning by manually trying to find the adequate pH value to fit the data. The function was also somewhat redundant because Proteome Discoverer would offer the charge of the discovered crosslinks without possessing any information about pH.

Input file

This tool is only compatible with Proteome Discoverer result files (i.e. “MS Annika” or “XlinkX”), which can be created by exporting the (filtered) CSM list to Microsoft Excel.

Support library

There are two types of support libraries: the digested form and the undigested form. In case the user wants to follow MS-detection related properties, when the peptides are already digested (by trypsin, for example), it is recommended to use the digested version of the library. In case the undigested version of the peptide may have a higher influence regarding the crosslinking process, the undigested version is recommended.

In any case, for the three existent libraries (the main, enrichable and “acidic” library) the correct results with regard to the “WGGGGR” sequence will be delivered in the function “most frequent neighbors of the binding amino acid”, even though the digested form of the peptide library was chosen. In case the user intends to utilize another library, an undigested form should be supplied for the function of the “most frequent neighbors of the binding amino acid” to work properly. For all the other properties and functions, the undigested semblance would be advantageous. Additionally, the ensembles of the peptide libraries, where peptides are in one single group, are also available (as mentioned before).

Normalization and comparison to theoretical crosslinks

We generate comparisons between “reality” and “theory”, which are calculated by assuming that each peptide from each group builds a crosslink (with only one CSM) with itself and with all the other peptides from the same group. Homeotypic duplicates are to be filtered out of the pool to eliminate the bias of them being present in two different groups.

Normalization is performed by setting the parameter “densitybool” as “True” when plotting the histogram in Python. When this Boolean parameter is changed from this default version (as “False”) to “True”, the script will return a valid probability density function. Every bin will display the bins raw count over the total number of counts and the bin width is adjusted so that the area under the histogram integrates to 1:

“(density = counts / (sum(counts) * np.diff(bins)))

(np.sum(density * np.diff(bins)) == 1)”[85]

The “stacked” parameter was also set to “True” for the sum of the histograms to be normalized to 1, as required by the documentation.[85]

For normalization, both “theory” and “reality” results are normalized to 1 (“density=True”). Because of that, some bins from the real results can be higher than the bins from the theoretically maximum reachable numbers of crosslinks (e.g., Peptide library 1 = 1018). The normalization is done on the CSM level, consequently a crosslink with 10 CSMs can influence the result 10 times more than the same crosslink with just one CSM.

In the unnormalized version, the area of all bins will not be equal, therefore neither the “reality”, nor the “theory” will be equal to one. In this case, the analysis is done on the crosslink level, hence duplicate CSMs will be eliminated, while two crosslinks have the same weight in the results, regardless of their CSM counts. The unnormalized variant is thought to detect rather qualitative deviation from the theoretical expected results. It reduces the dimension of CSM data to XL level. The theoretical results are always greater or equal to the experimental results because the crosslinking yields can not reach 100%.

Output Files

Isoelectric Point

The role of this function is to display the distribution of the correctly identified XL IDs depending on their pI value.

The pI value is calculated for each peptide of the identified crosslink and is weighted by the number of K, R, H, D, E present. This is, of course, just an approximation and might be subject to an intrinsic bias. Nevertheless, in the algorithm, C also has an influence on the theoretically calculated isoelectric point, but in our simulations, it has a negligible impact of 0.01. Histidine (H) has indeed a lower impact on the pI value, but not insignificant. In the conducted simulations on theoretical peptide sequences, it was observed that the algorithm differentiates between a basic residue at the N-terminus (a slightly higher pI) and a basic residue anywhere else. The same is true for acidic residues at the C-terminus and anywhere else. With that being said, no differences were observed when the relative position between two or more basic or acidic residues differed. The pI value was also impassive to changes of length in the peptide sequence as long as the new residues added did not have side chains with basic or acidic groups.

An example of how the pI value of a crosslink is calculated is presented below:

AHHKEGGR Peptide A
|
MTKEEGGR Peptide B

$$pI\ value(Crosslink) = \frac{pI\ value(Peptide\ A) * 5 + pI\ value(Peptide\ B) * 4}{5 + 4}$$

In the case of pI calculation, the lysines connected by the crosslinker are also included in the algorithm to focus on the isoelectric properties of the initial peptides influencing the formation of crosslinks.

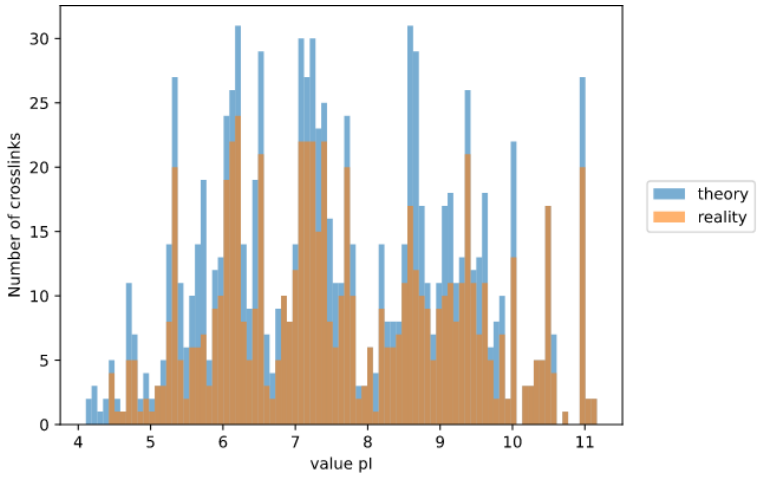


Figure 18: Isoelectric point distribution on the crosslink level where areas are not normalised; values are weighted to the frequency of K, R, H, D, E

Hydrophobicity

The role of this function is to display the distribution of the correctly identified XL IDs depending on their hydrophobicity on the GRAVY scale.

The GRAVY index of the crosslink is calculated based on Kyte and Doolittle scale[86]. Every amino acid possesses an empirically determined GRAVY value, which is used to calculate a weighted average considering the length of both peptides. Of note, even though the crosslinker itself is not included, the analysis allows for a comparative overview of the theoretical results.

$$Gravy\ value(Crosslink)$$

$$= \frac{[Gravy\ value(Peptide\ A) * Length(Peptide\ A)\ + Gravy\ value(Peptide\ B) * Length(Peptide\ B)]}{Length(Peptide\ A) + Length(Peptide\ B)}$$

R	K	N	D	Q	E	H	P	Y	W	S	T	G	A	M	C	F	L	V	I
-4.5	**-3.9**	**-3.5**	**-3.5**	**-3.5**	**-3.5**	**-3.2**	**-1.6**	**-1.3**	**-0.9**	**-0.8**	**-0.7**	**-0.4**	**1.8**	**1.9**	**2.5**	**2.8**	**3.8**	**4.2**	**4.5**

Figure 19: Hydrophobicity scale by Kyte and Doolittle- GRAVY value index

It is worth mentioning that the Kyte and Doolittle scale is not the only hydrophobicity scale existent, but the most utilised one, present on many bioinformatic platforms[87,88].

Another scale example would be the Hopp-Woods scale[89], which was developed to determine the potentially antigenic sites in proteins; predictions are most accurate when using a window size of 7 or the Cornette scale[90], which is suitable for predicting alpha helices regions in proteins. The GRAVY values may differ considerably from one scale to another.

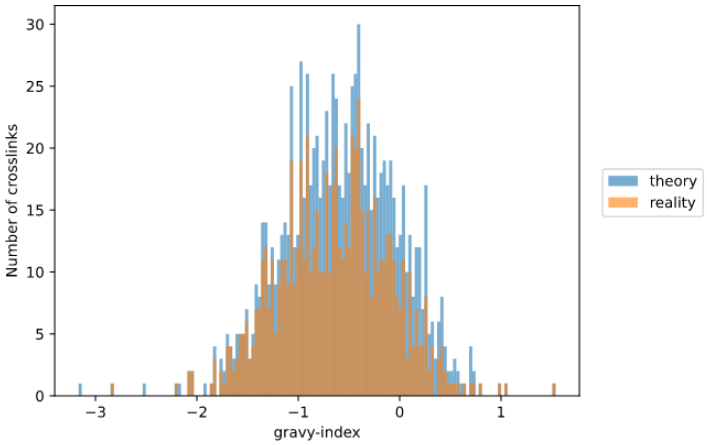


Figure 20: Hydrophobicity distribution on the crosslink level where areas are not normalised.

Frequency of amino acids in crosslinks

This function's role is to calculate and display the average frequencies of different types of amino acid residues across the entire data set of correctly determined links.

The image below shows the frequency (number of occurrences of a specific amino acid/ total number of amino acids per CSM) of each amino acid regardless of their position in the peptide across all CSMs. The crosslinker reactive amino acids are also included in the calculation of the frequency.

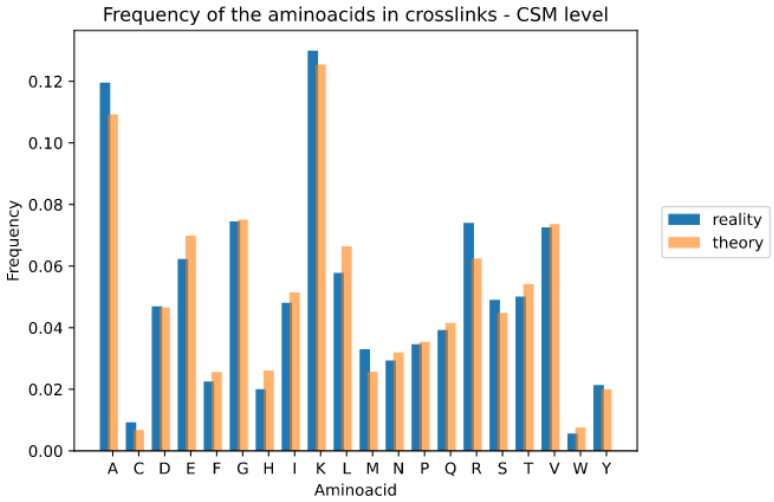


Figure 21: Frequency of the amino acids weighted on the CSM level of a result file.

The percentages themselves do not offer any new valuable information, but relative to the theoretical results, one can observe if the sole presence of some residue types (independent of the relative position to the crosslinker-reactive amino acid) can have a positive or negative influence on the formation of the crosslink.

Frequency of aromatic amino acids (F, W, Y)

The scope of the function "frequency of aromatic amino acids" is to calculate and present the distribution of frequency of aromatic amino acids through the XLs marked as correct.

The amino acids phenylalanine (F), tryptophan (W) and tyrosine (Y) are considered aromatic by the Biopython 1.79. Histidine is not being included in the calculation and is therefore seen as non-aromatic, even though it fulfils all 4 criteria of aromaticity: cyclicity, continuously conjugated double bonds, planarity and 4n+2 π e^- (Hückel's rule)[91]. This aspect might lead to some deviations from the real results, and it would be recommended to be corrected in future research. As these amino acids might alter the reactivity of neighbouring functional groups and as they are sterically bulky (because of the presence of the aromatic ring), their frequency might give information if there is any influence on the formation/ability for detection of a crosslink.

On the other hand, it is possible that they could show a stabilizing effect through pi stacking if both peptide sequences involved in the crosslink are presenting an aromatic amino acid. Not only that, but aromatic residues can undergo non-covalent pi-pi contact interactions with other residue types such as non-aromatic side-chains or peptide backbone.[92] This multifactorial problem of steric hindrance, stabilizing pi-pi interactions, torsion angles, etc. and the role of the conformation of the crosslinks in the context of how it affects the reaction yield or the fragmentation is to the best of the author's knowledge not tackled yet. One can assume that the peptides adopt a linear structure because of their short length, but crosslinks could adopt more complex ones.

For exemple, an energy minimization job with MM2[93](with a minimum RMS gradient of 0.01) was performed on an imaginary crosslink containing more aromatic residues. It does provide a low accuracy for molecules with many polar functional groups[94] but it can give a strong hint about the more complex 3-dimensional structure a crosslink can adopt. In the case presented below, the crosslinker tends to bend in order for the non-covalent interactions to take place and an energetically more favourable conformation is formed.

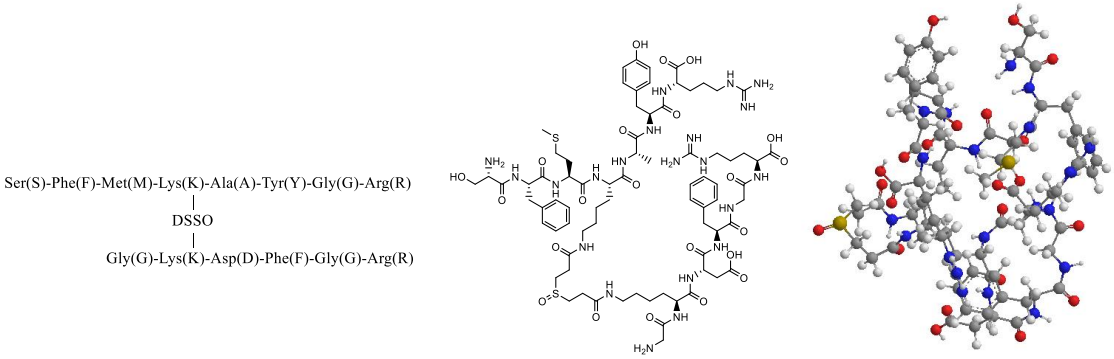


Figure 22: Crosslink in three different visualisation formats. In the first representation, three letter codes (respectively one letter codes) are used; in the second one, the structural formula; in the third formula, the three-dimensional ball-and-stick model is showing the molecule after running an MM2 energy minimization job (total energy of -99.8459 kcal/mol)

Calculation is explained based on the following example:

AAYMTKEMPR Peptide A (length = 10, number of aromatic amino acids = 1)

|

STWPQNKPQFR Peptide B (length = 11, number of aromatic amino acids = 2)

$$Frequency\ of\ aromaticity\ (Crosslink) = \frac{[Frequency\ of\ aromaticity(Peptide\ A) * Length(Peptide\ A) + Frequency\ of\ aromaticity(Peptide\ B) * Length(Peptide\ B)]}{Length(Peptide\ A) + Length(Peptide\ B)}$$

$$Frequency\ of\ aromaticity\ (Crosslink) = \frac{\left[\left(\frac{1}{10}\right)*10+\left(\frac{2}{11}\right)*11\right]}{10+11}$$

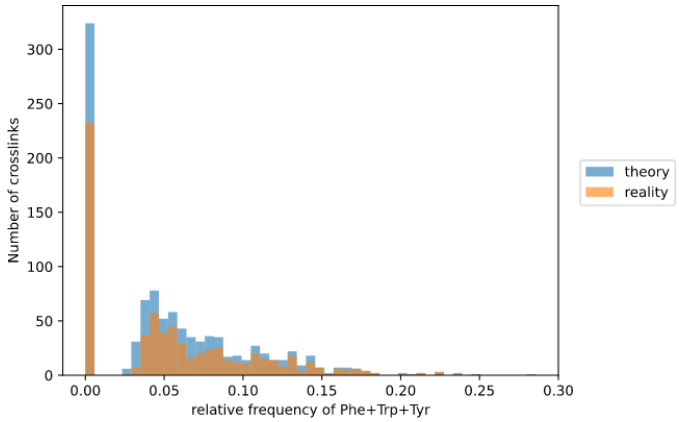


Figure 23: Aromaticity frequency distribution on the crosslink level where areas are not normalised.

Molecular weight

The following approach presents the distribution of the crosslink population based on their molecular weight.

For the molecular weight, the mass of the connected peptides and the crosslinker are summed up. Importantly, the program ignores all other chemical modifications, which might alter the calculated mass vs. real mass.

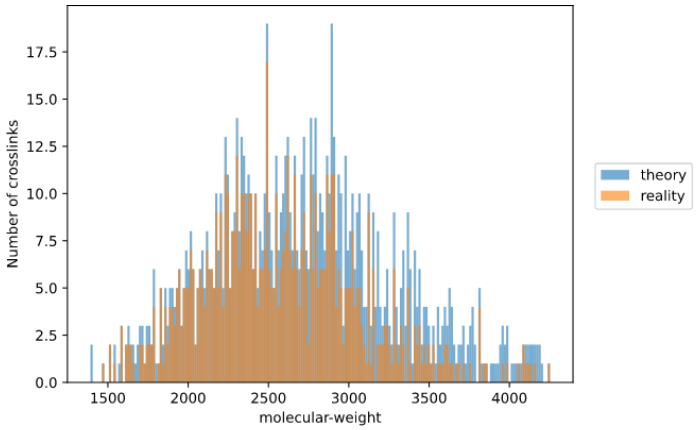


Figure 24: Molecular weight distribution on the crosslink level where areas are not normalised.

Neighbours of the crosslinkers binding amino acid

The purpose of the next procedure is to calculate the frequency matrix (with regard to the amino acid residue types) of the three closest neighbours from both sides of the linker's binding sites across the whole correct XL data set.

It is known that homobifunctional N-hydroxy succinimide (NHS) esters react primarily with lysine residues, but side reactions with serine, threonine and tyrosine might occur. In a study concentrated on enhancing the reactivity and deciphering the reaction mechanism of these side reactions, the

authors synthetized some peptide models where they showed how changes in the pH value and amino acid residues impact the crosslinking yield.[95] The presence of histidine in the close proximity of serine or tyrosine, more specifically in the structure "-HXT-" or "-HXS-", enhances the yield of the reaction.[95] With this new gained knowledge, it is valuable to analyse which residues in the close proximity of the crosslinker reactive residue can positively or negatively influence the reaction. The function counts the frequencies of the amino acids where their positions are accounted for as well.

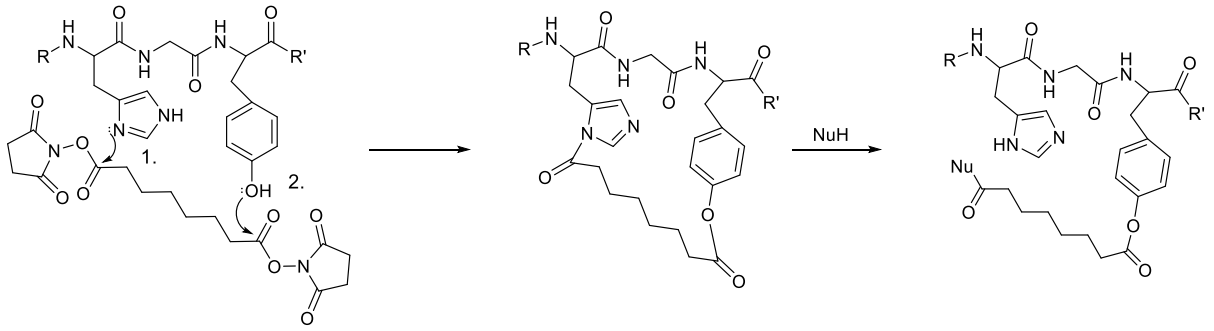


Figure 25: Proposed mechanism[95] for the high prevalence of Type 1 crosslinks with tyrosine. In the initial article, the nucleophile was water, which results in a Type 1 crosslinks. In the mechanism above, the notation "NuH" was used to include the possibility of another nucleophile, for example lysine, to react with the intermediate state and facilitate the reaction. This represents a suitable example of how neighbouring residues (depending on their proximity) can affect the linkage reaction.

To address if the direct chemical environment of crosslinker reactive amino acids influences its reactivity, three amino acids to the left and to the right of the crosslinker bound amino acid are reported. Below, the position frequency matrix structure from the analysed structures is shown, where each row represents the following characteristics:

- first row - third neighbour from the left (farthest left)
- second row - second left neighbour (middle left located)
- third row - first left neighbour (closest left)
- fourth row - first neighbour on the right (closest right)
- fifth row - second neighbour from the right (middle located right)
- sixth row - third neighbour on the right (farthest right)

```
The most frequent amino acid residues neighboring the binding amino acid are:

GGA ~binding amino acid~ A--

Position frequency matrix of the closest neighbours to the binding amino acid from left (N-Terminus-direction) to right (C-Terminus-direction)
Each row of the matrix represents a distinct position reported to the binding amino acid.

  A       C       D       E       F       G       H       I       K       L       M       N       P       Q       R       S       T       V       W       Y       X   *   -
0.125000 0.013057 0.016053 0.046233 0.012414 0.251284 0.012200 0.029966 0.000000 0.040026 0.046447 0.046447 0.003853 0.030822 0.128425 0.071490 0.032962 0.081122 0.00000 0.012200 0.0 0.0 0.000000
0.093108 0.000000 0.046447 0.027825 0.040454 0.183861 0.012628 0.045591 0.000000 0.058219 0.033818 0.035103 0.033176 0.054580 0.072132 0.072988 0.084760 0.069349 0.00899 0.026969 0.0 0.0 0.000000
0.148330 0.044092 0.082192 0.108091 0.008990 0.064212 0.017123 0.085616 0.000000 0.048587 0.061858 0.030394 0.031036 0.036387 0.112158 0.034461 0.032320 0.054152 0.00000 0.000000 0.0 0.0 0.000000
0.205051 0.010702 0.028896 0.065068 0.031892 0.070848 0.037671 0.061430 0.036601 0.097817 0.016053 0.000000 0.041738 0.088399 0.067423 0.031464 0.042594 0.029324 0.00000 0.037029 0.0 0.0 0.000000
0.101884 0.000000 0.067637 0.014983 0.011772 0.069563 0.019264 0.074272 0.041310 0.067851 0.025471 0.029324 0.011772 0.063570 0.063142 0.048159 0.093964 0.075985 0.00000 0.016053 0.0 0.0 0.104024
0.063142 0.000000 0.031678 0.054152 0.037243 0.032106 0.012842 0.032320 0.047517 0.073630 0.014341 0.082620 0.029966 0.015197 0.134204 0.052654 0.012842 0.049229 0.00000 0.019692 0.0 0.0 0.204623
```

Figure 26: Neighbours of the crosslinker's binding position weighted on the number of CSMs for each crosslink.

It is reported in the one-letter amino acid code and the special characters "*", "X" and "-".

X	Any amino acid
*	Translation stop
-	Empty space (beyond the end of the peptide sequence)

Retention time vs physicochemical properties

In the figure below there is a display of the retention time versus the above-described physiochemical properties of each correctly annotated CSM.

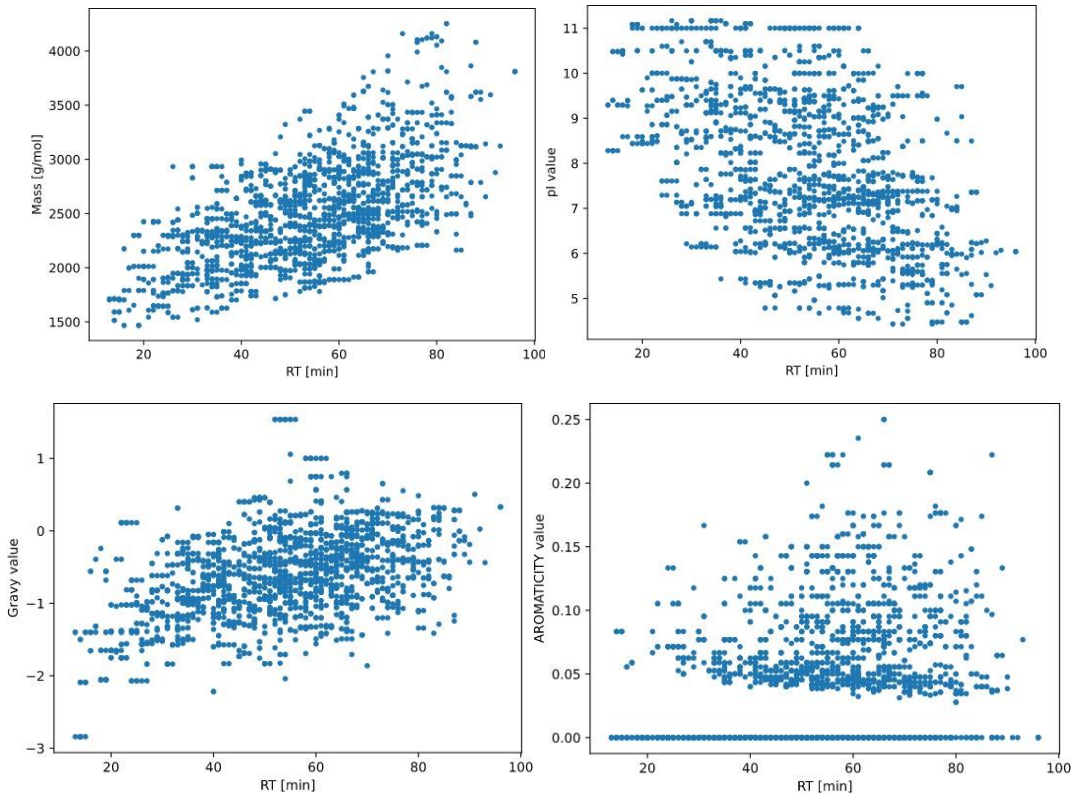


Figure 27: Physicochemical properties (molecular weight-top left; isoelectric point-top right; GRAVY value-bottom left; aromaticity fraction-bottom right) as a function of retention time (RT)

Moderate dependencies between the physicochemical properties and retention times can be observed. Higher molecular masses of the crosslinks lead to higher retention times. The calculated pI values decrease with the increasing elution times. In the case of GRAVY values, it can be stated that, in our data sets, higher hydrophobicity slightly increases the elution time. An interesting phenomenon can be observed in the case of the aromaticity: a pronounced unconditional heteroscedasticity.[96,97] The variance of the data tends to increase over time and it is unconditional because the pattern repeats itself on more data sets. It remains unclear if this is due to the high mass of aromatic residues or because most of the analysed sample is eluting at lower retention times.

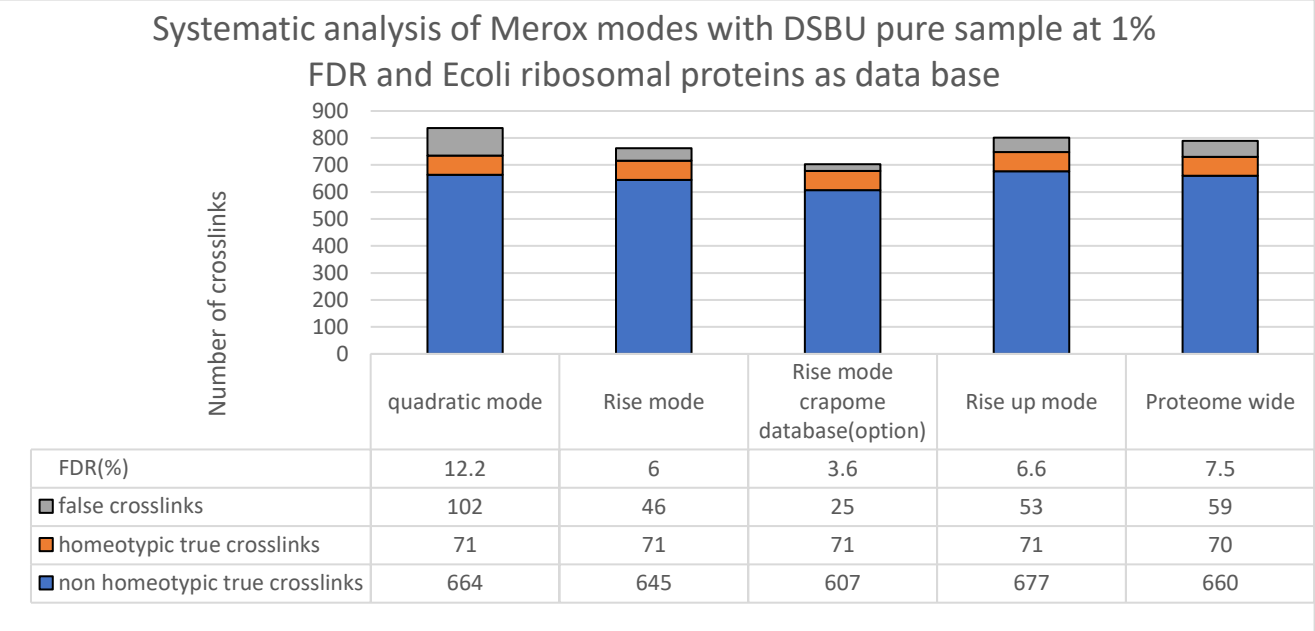
A more concise version of the software technicalities and the correct way to use it can be found on Github[98].

3. Results

3.1. Harmonization Studies

A harmonization study regarding the software settings of MeroX, Annika, XlinkX2 and pLink2 was done in order to achieve optimal settings for each software in the peptide library system. The chosen crosslinker was DSBU. Due to its longest spacer arm from the evaluated linker, it produces the highest

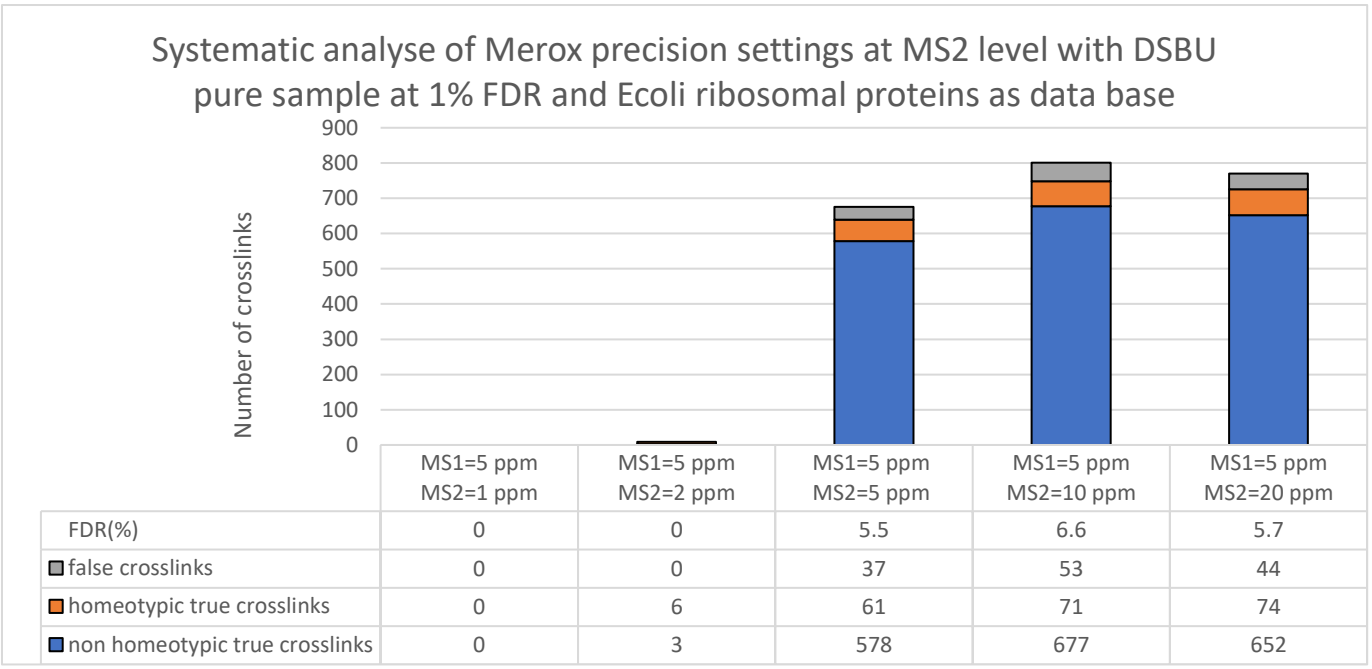
number of crosslinks and is a viable choice for this type of analysis. Each crosslinking search engine has its own particularities and different parameters can be defined by the user. The scope of this study was to define common settings, but also to find the specific parameters where each software performs at its best. This assures that every search engine is not negatively discriminated in any way when confronted with the model system.



	quadratic mode	Rise mode	Rise mode crapome database(option)	Rise up mode	Proteome wide
FDR(%)	12.2	6	3.6	6.6	7.5
false crosslinks	102	46	25	53	59
homeotypic true crosslinks	71	71	71	71	70
non homeotypic true crosslinks	664	645	607	677	660

Figure 28: Systematic analysis of Merox settings with DSBU pure sample at 1% FDR and E. coli ribosomal proteins as library support. Different algorithmic modii

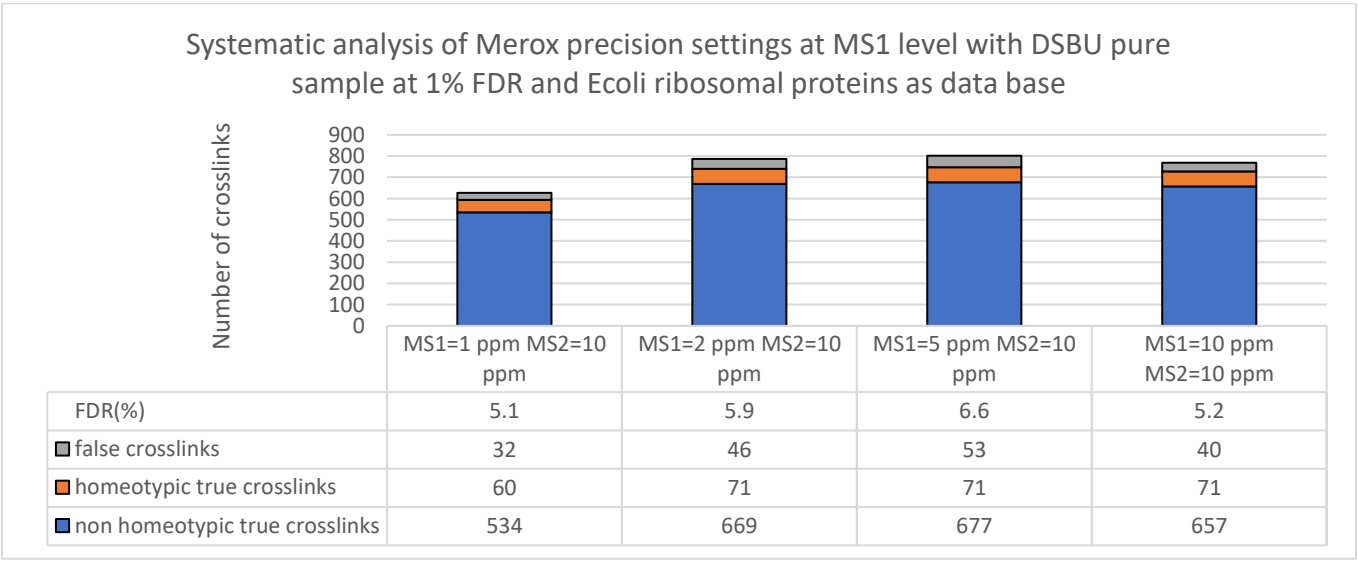
In the figure above, it is clear that the Rise up mode delivers the best quantitative results yielding 677 non-homeotypic true crosslinks and 71 homeotypic true crosslinks. The time-consuming quadratic mode seems to deliver the highest number of crosslinks, but with an FDR of 12.2%. As mentioned before in the algorithmic details, the "quadratic mode" does not rely on the presence of a reporter ion to confirm the existence of a crosslinker. One would normally not choose this modus for cleavable crosslinker such as DSBU, but for uncleavable ones where no reporter ions can exist anyway. Both the number of unique crosslinkers and the higher FDR are explained through the absence of this extra "layer of security". The synthetic library highlights the algorithmic limitations of the quadratic mode and proves the unreliability of the results in a more complex system, as underlined by authors of the software. An ”additional filter” provided by the Crapome data base can prove very useful in “conservative” crosslinking experiments, as it presents a real FDR of 3.6%, which is very close to the theoretically calculated value of 1%, but with worse quantitative results.



	MS1=5 ppm MS2=1 ppm	MS1=5 ppm MS2=2 ppm	MS1=5 ppm MS2=5 ppm	MS1=5 ppm MS2=10 ppm	MS1=5 ppm MS2=20 ppm
FDR(%)	0	0	5.5	6.6	5.7
false crosslinks	0	0	37	53	44
homeotypic true crosslinks	0	6	61	71	74
non homeotypic true crosslinks	0	3	578	677	652

Figure 29: Systematic analysis of Merox precision settings with DSBU pure sample at 1% FDR and E.coli ribosomal proteins as library support: The precursor precision (at MS1 level) remains constant at 5 ppm, while the fragment ion precision is varied from 1ppm to 20ppm

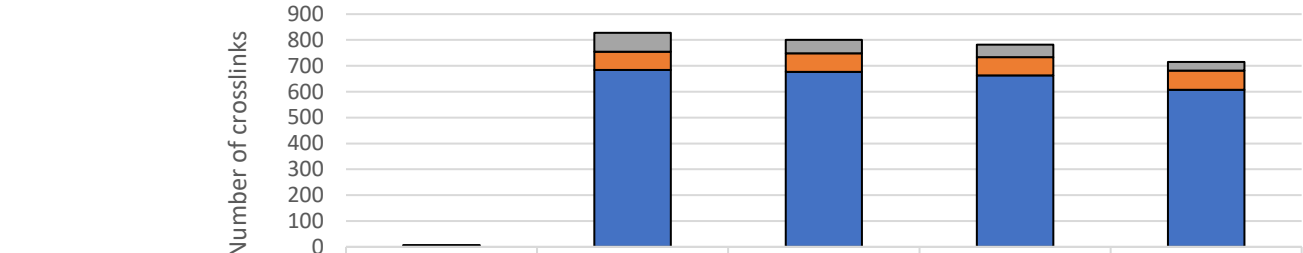
The precursor precision and fragment ion precision values allow a certain, user-defined deviation of the real detected masses from the theoretically calculated masses of the precursor ion at MS1 level and the resulting fragment ions at MS2 level from the theoretically calculated values. Having a too stringent system (1 ppm or 2 ppm) is not beneficial because no crosslinks or just a few crosslinkers can be identified. The setting MS1 = 5 ppm, MS2 = 5 ppm shows the lowest FDR value (5.5%), but with a high cost of losing 109 true crosslinks in comparison to the recommended values of 5 ppm precision for MS1 and 10 ppm precision at MS2 proposed in the standard settings. Interestingly, allowing a fragment ion precision of 20 ppm leads to a slight decrease in the found crosslinks and a better FDR value.



	MS1=1 ppm MS2=10 ppm	MS1=2 ppm MS2=10 ppm	MS1=5 ppm MS2=10 ppm	MS1=10 ppm MS2=10 ppm
FDR(%)	5.1	5.9	6.6	5.2
false crosslinks	32	46	53	40
homeotypic true crosslinks	60	71	71	71
non homeotypic true crosslinks	534	669	677	657

Figure 30:Systematic analysis of Merox precision settings with DSBU pure sample at 1% FDR and E.coli ribosomal proteins as data base: The fragment ion precision (at MS2 level) remains constant at 10 ppm, while the precursor precision is varied from 1 ppm to 10 ppm

The precision of 1 ppm at the MS1 offers the lowest FDR and does not drastically decrease the amount of found crosslinks as in the case of the same precision stringency at MS2 level. This underlines the vital importance of a precise detection of the precursor mass ions. The proposed standard setting (MS1= 5 ppm MS2= 10 ppm) once again outperforms the other options. Reducing the stringency level shows again a counterintuitive behaviour by obtaining a lower number of crosslinks with a lower FDR.

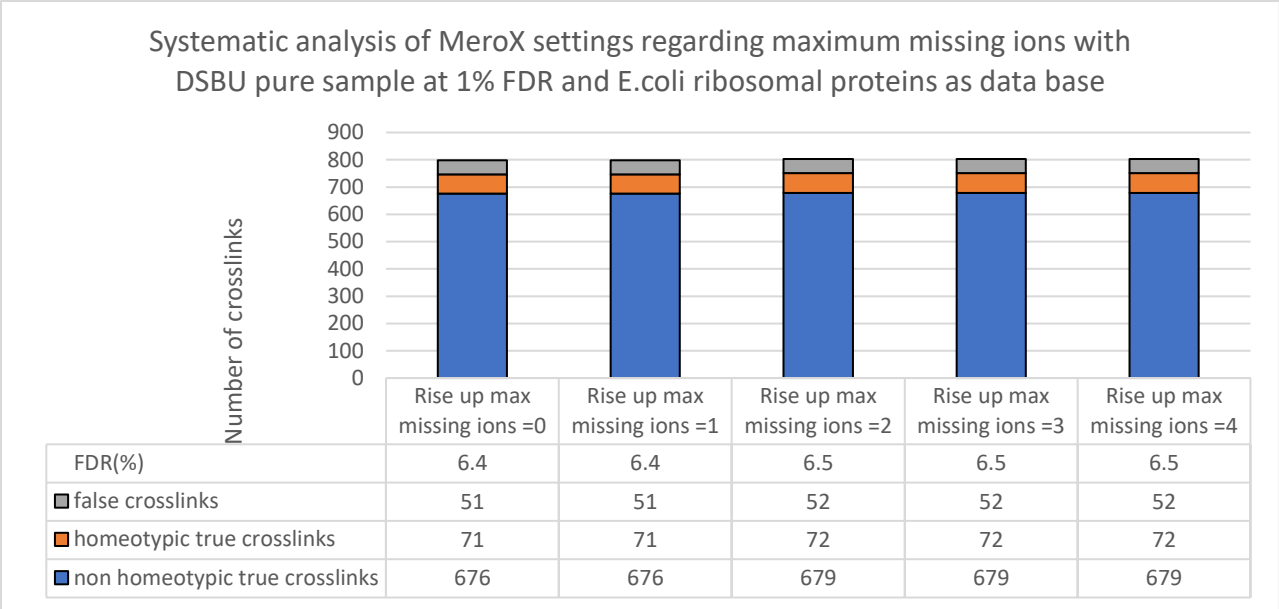


	MS1=1 ppm MS2=2 ppm	MS1=4 ppm MS2=8 ppm	MS1=5 ppm MS2=10 ppm	MS1= 6 ppm MS2=12 ppm	MS1= 10 ppm MS2=20 ppm
FDR(%)	0	8.8	6.6	6.1	4.6
false crosslinks	0	73	53	48	33
homeotypic true crosslinks	5	71	71	71	74
non homeotypic true crosslinks	2	684	677	663	608

Figure 31:Systematic analysis of Merox precision settings with DSBU pure sample at 1% FDR and E.coli ribosomal proteins as data base; The precursor precision (at MS1 level) and fragment ion precision (at MS2 level) are varied while maintaining a constant ratio of 1:2 between them.

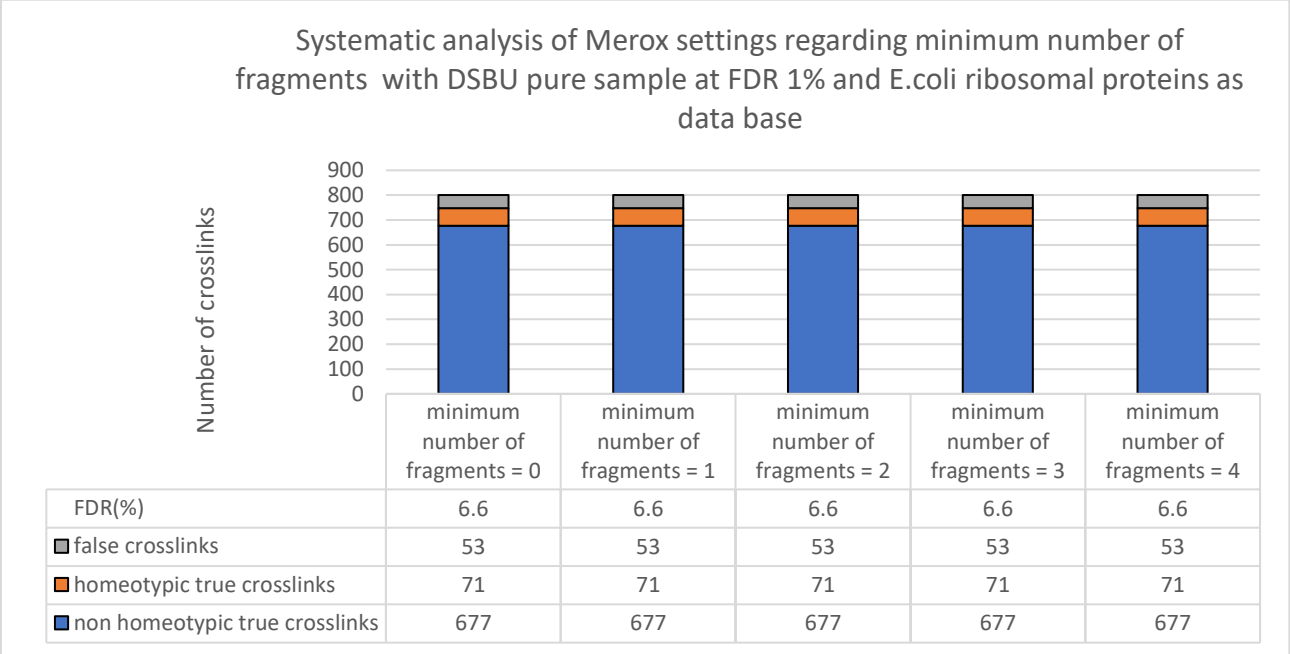
In the scenario where both precisions levels are changed but restricted to the same ratio of 1:2 due to the higher precision reached by the mass spectrometer at MS1 level, the best performance is still achieved by the proposed standard settings. Even though the pair 4 ppm-8 ppm delivers 7 more crosslinks, this comes with the expense of 20 new false crosslinks.

The number of maximum missing ions (from 0 to 4) and the minimum number of present fragments (see Figure 7 and Figure 8) do not influence the results. This shows the capacity of the search engine to compensate the loose settings of some parameters and still maintain the same search-quality.



	Rise up max missing ions =0	Rise up max missing ions =1	Rise up max missing ions =2	Rise up max missing ions =3	Rise up max missing ions =4
FDR(%)	6.4	6.4	6.5	6.5	6.5
false crosslinks	51	51	52	52	52
homeotypic true crosslinks	71	71	72	72	72
non homeotypic true crosslinks	676	676	679	679	679

Figure 32: Systematic analysis of MeroX settings regarding maximum missing ions with DSBU pure sample at FDR 1% and E. coli ribosomal proteins as data base



	minimum number of fragments = 0	minimum number of fragments = 1	minimum number of fragments = 2	minimum number of fragments = 3	minimum number of fragments = 4
FDR(%)	6.6	6.6	6.6	6.6	6.6
false crosslinks	53	53	53	53	53
homeotypic true crosslinks	71	71	71	71	71
non homeotypic true crosslinks	677	677	677	677	677

Figure 33: Systematic analysis of MeroX settings regarding minimum number of fragments with DSBU pure sample at FDR 1% and E.coli ribosomal proteins as data base

In the workframe of Proteome Discoverer, it is easier to automatise more runs, so more data points were collected when the MS1 and MS2 settings in Annika were benchmarked. When MS1 is kept constant at 5 ppm and MS2 is varied the number of XL IDs drops at 420 correct links at 1 ppm, which is not a dramatic decrease. There are also three local maxima (in the function of correct links depending on the ppm evolution): at MS2 = 3 ppm with 571 correct XLs, at MS2 = 7 ppm with 565 correct XLs and at MS2 = 10 ppm with 575 correct links. This indicates self-compensatory mechanisms built in the algorithm, which can be beneficial.

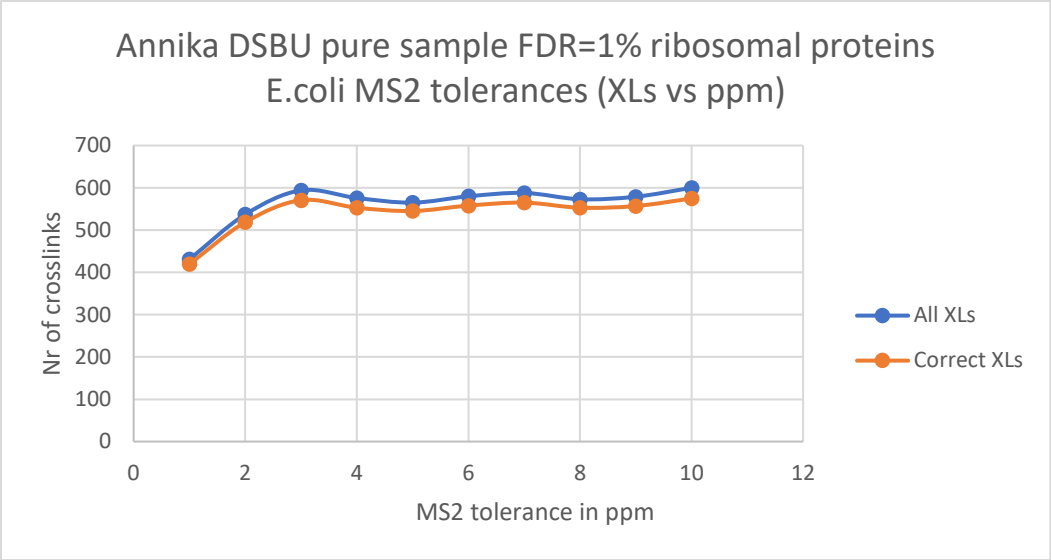


Figure 34: Annika crosslink search engine DSBU pure sample, FDR 1% ribosomal proteins E coli as data base, MS2 tolerances are varied, while MS1 tolerance is kept constant at 5 ppm. The number of found crosslinks and the true crosslinks are represented here as a function dependent of different tolerances (from 1 ppm to 10 ppm). The function shows an ascendent trend from 1 ppm to 3 ppm and reaches then a plateau with minimal variations.

For MS1 tolerances, the performance of Annika decreases dramatically when a MS1 tolerance of 1 ppm is chosen. Only 214 correct links are identified which is approximately 37% of the links found at the best setting chosen (10 ppm with 583 correct XLs).

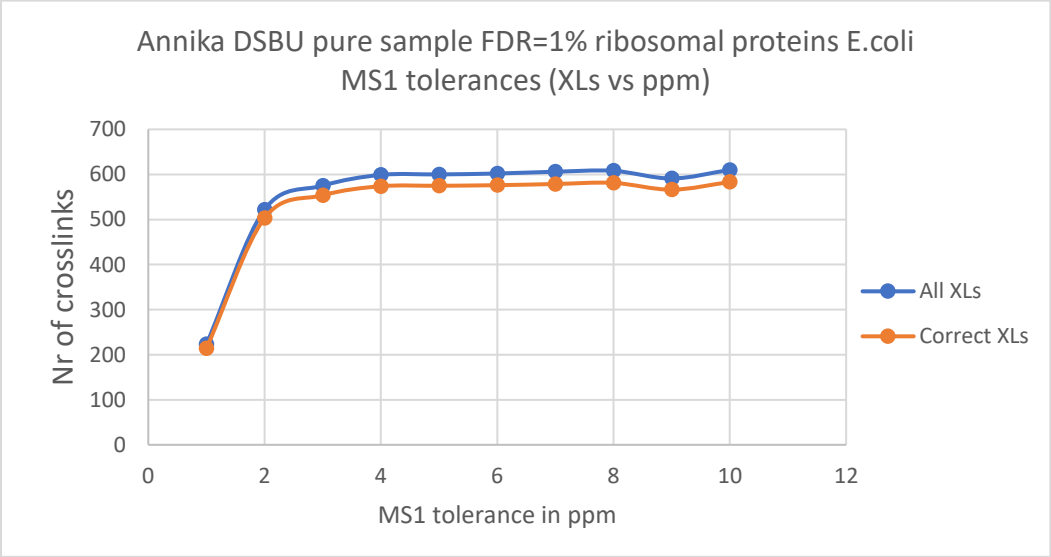


Figure 35: Annika crosslink search engine DSBU pure sample, FDR 1% E coli ribosomal proteins as data base, MS1 tolerances are varied while MS2 tolerance is kept constant at 10 ppm. All found crosslinks and the true crosslinks are represented as functions dependent on different MS1 tolerances in ppm (from 1 ppm to 10 ppm).

Regarding the different modii, it is shown in the figure below that the combined mode (of shotgun and evidence mode) produces the highest number of crosslinks (600 XL IDs) and is therefore the recommended option.

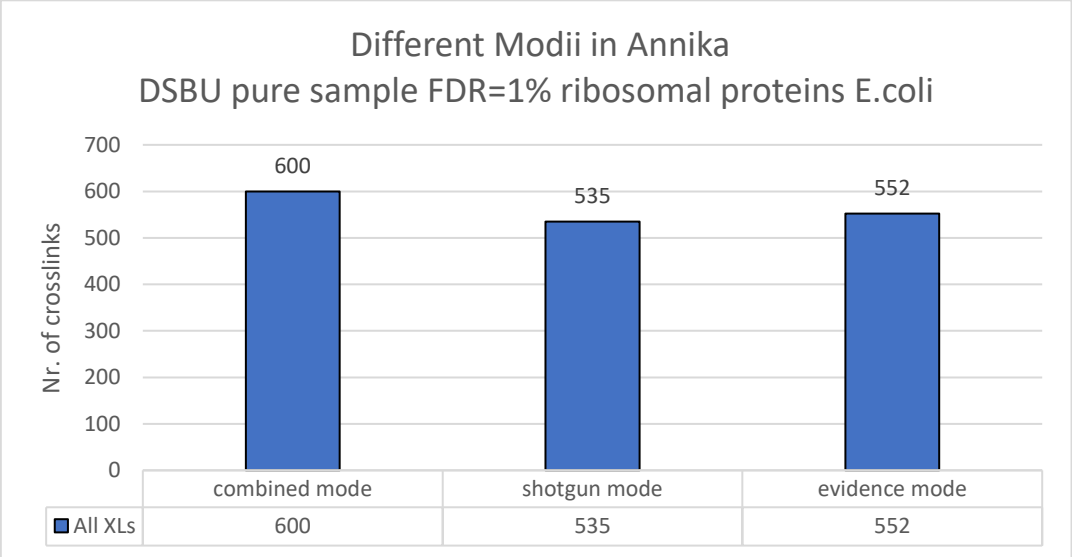


Figure 36: Annika search engine, different modii DSBU pure sample FDR 1% ribosomal proteins E.coli as data base

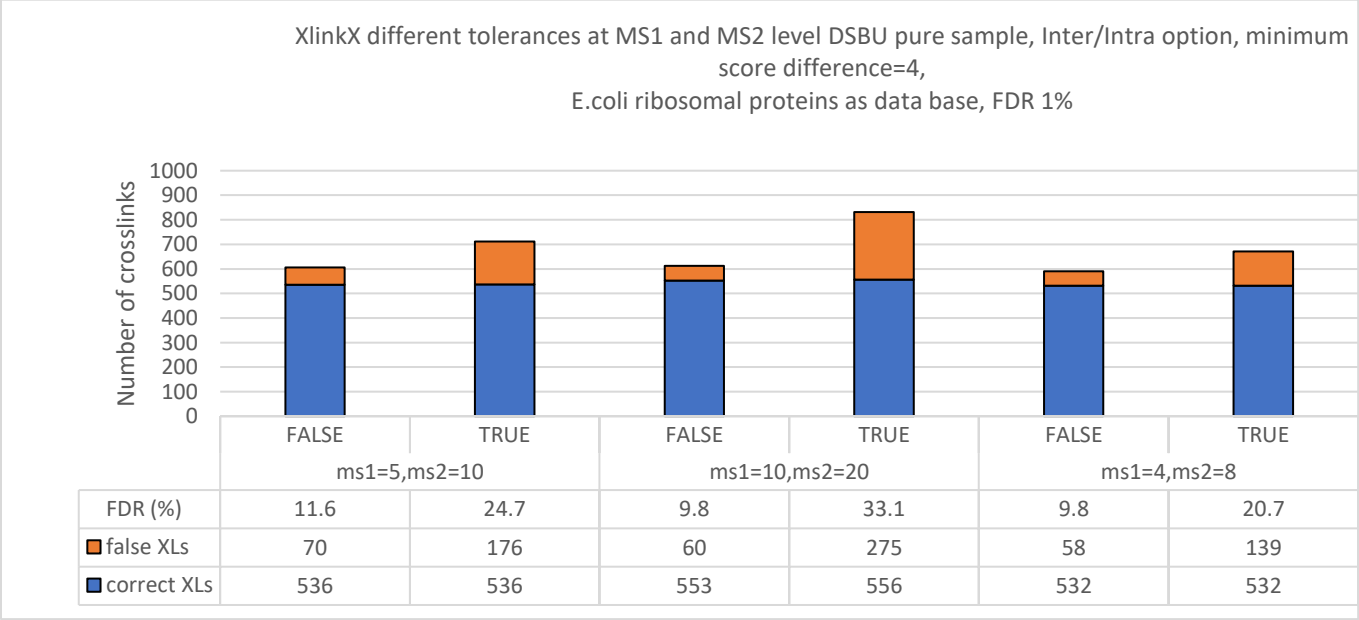


Figure 37: XlinkX different tolerances at MS1 and MS2 level DSBU pure sample, Inter/Intra option, minimum score difference=4, E.coli ribosomal proteins as data base, FDR 1%

A very interesting phenomenon was observed when analysing the different behavioural patterns in the performance of XlinkX. Varying the tolerances at MS1 and MS2 level does not produce any drastic changes in the results. The software maintains its capacity to find roughly the same number of correct crosslinked peptides at an almost constant FDR. In the last version of the XL search engine, a new feature was introduced: the option to postscore interlinks and intralinks separately. The necessity of this new extra option can be explained by the intrinsically higher scores that Intralinks have. With a peptide library formed of peptides originating from different proteins, it can be shown that most of

the additional crosslinks found are in fact false. The aspect could have not been made visible with the synthetic peptide library developed by Beveridge et al.[78] since all peptides used originate from one single protein and therefore all found links would be categorized as intralinks by every XL search algorithm. If this option is activated in the standard settings (MS1 5ppm and MS2 10 ppm), it leads to FDR values of over 20% and affects the reliability of the results.

3.2. Benchmarking of XL-search engines

Before reporting the results and the findings of the study, the workflow must be made clear and the differences to the Beveridge's peptide library[78] must be highlighted. Already in the design of the peptides changes have been made. All the sequences from Beveridge's synthetic library originate from Cas9 protein from the organism *Streptococcus pyogenes*. Although the relevance of the protein is unquestionable, as it is commonly used as a tool for genome engineering (and new insights regarding the crosslinking behaviour of these peptides are valuable and attract interest), it does present a flaw: only "quasi-intralinks" will be discovered. Because of its high mass (158 kDa) and a high number of residues (1368), it was possible to develop a whole library out of a single protein sequence.[99–101] Another small, but significant change in the design is the sequence WGGGGR instead of the sequence YGGGGR. Both sequences mimick the digestion of a real protein in the *in vivo* crosslinking, but tryptophan presents a considerably higher molar extinction coefficient, allowing for a better quantification and aliquotation. Finally, the biotin moiety was intentionally eliminated (and was the corresponding streptavidin treatment). This resulted in a significant background signal from the highly concentrated Ac-WGGGGR sequence in the final mix after digestion, which has the potential to interfere with crosslinking identification and reduce the dynamic range. However, fortunately, this does not seem to be the case as the exhibited signal has distinct characteristics at a specific retention time in the Total Ion Chromatogram (TIC). Instead of one single library with 426 possible links and 95 peptides included, three chemically different libraries were constructed (with 100, 64 and 43 peptides respectively) and the assignment of each peptide in 2 groups allowed a higher barrier in maximum numbers of links that can be reached (1018, 512 and 280 possible links). Chemically, the workflow was mostly maintained constant (concentration and reaction times proved to be perfected already), but other crosslinker reagents were tried. The benchmarking of the software with a focus on FDR, but also the comparison of results and the analysis of different properties was fully automatized.

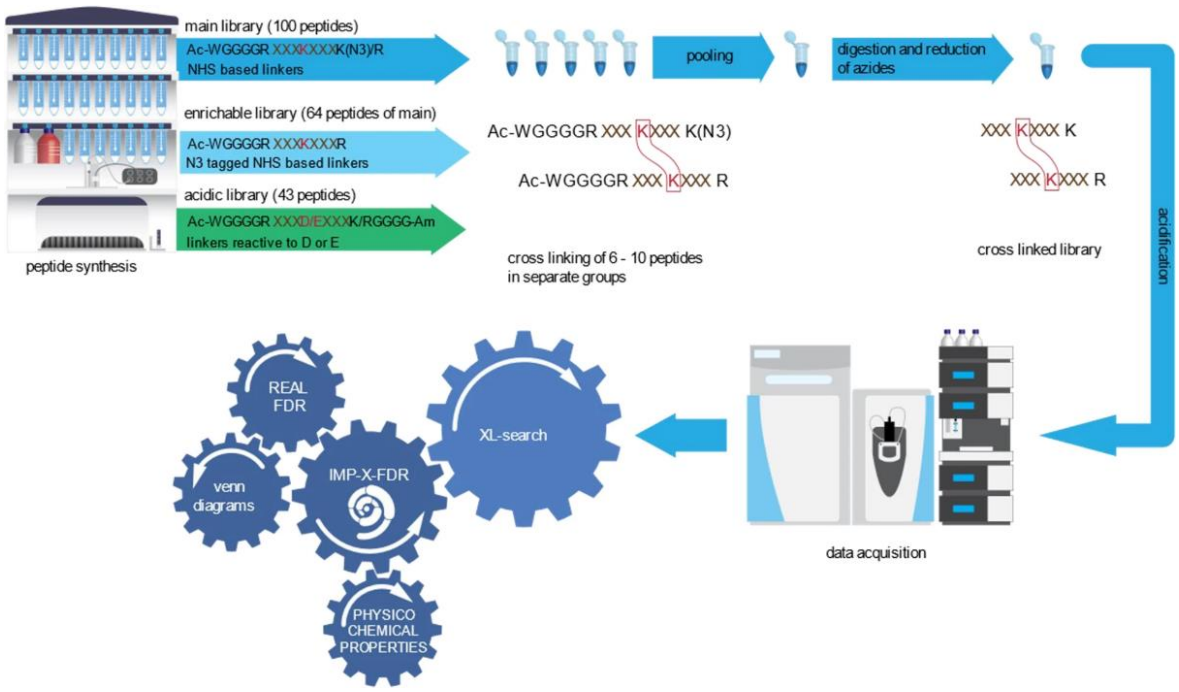


Figure 38: workflow represents schematically and simplified each step of the process. It starts with designing, synthetising, purifying and aliquoting the protected peptides of the libraries. It is followed by library-specific crosslinking experiments, where each group stays isolated from each other. The single groups are then mixed, the solution acidified and data is obtained through a HPLC-MSMS (stepped HCD) run. The acquired data runs first through a XL-search algorithm than through IMP-X-FDR for automatic benchmarking of XL-results. Figure adapted from Matzinger, M. Vasiu, A. et al [79]

In the case of DSSO applied to the main library, MeroX and MS Annika perform the best, obtaining on average 620 and 614 true crosslinks, respectively, out of 1018 chemically possible. Annika's algorithm does present a lower FDR value (2.7% compared to 5.7%). This could be a consequence of the fact that the algorithm was trained and perfected on DSSO, while MeroX was based on DSBU training data. XlinkX, with an average of 503 crosslinks, and pLink2, with 438 crosslinks, have a worse outcome. In the case of pLink, the results are somehow surprisingly good considering that the algorithm is specialised neither on HCD data nor on cleavable linkers. The postprocessing procedure to bring the FDRs to an experimentally validated value of 1% does not drastically affect the quantity of results in the case of MeroX, MS Annika and XlinkX (moderate increase of the score of lowest accepted crosslink).

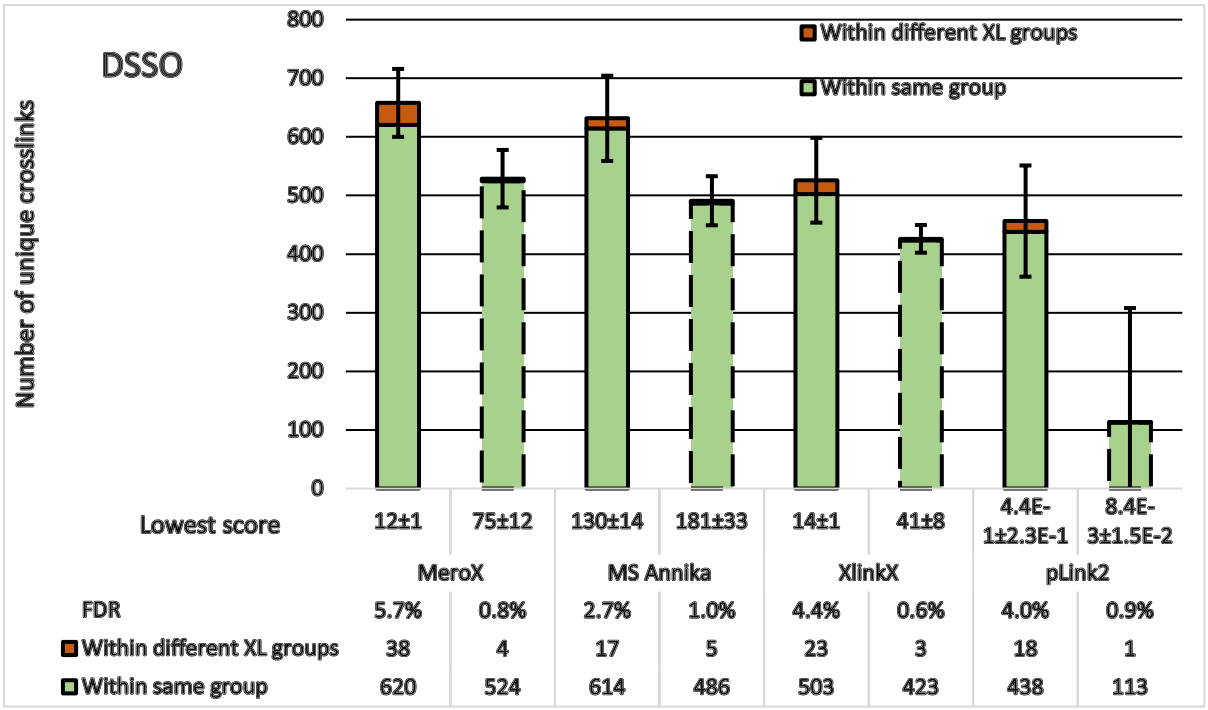


Figure 39: The mean unique crosslink number reported from each algorithm (MeroX, MS Annika, XlinkX and pLink2) benchmarked (sample size n= 3) in the case of DSSO-main peptide library. The solid lines refer to the results obtained with a set FDR value of 1%. The long-dashed lines represent the average results after the post-processing in order to bring them to an experimentally validated FDR value of 1%. The error bars show the standard deviation from each sample with regard to the total number of XL independent of the true/false criterium. The lowest score row shows (with the SDs attached) the score of the worst graded accepted link in the data set.

There is a very interesting three-step decrease in the FDR level when the replicates from DSSO are overlayed. Three distinctive data sets must be considered: links that are unique to one replicate (with the highest real FDR values up to 34.8%), links common to 2 of the replicates (where the FDR decreases dramatically and is near the expected values), and links common to all three replicates. The last groups present the highest result confidence, with an FDR under the expected value of 1%. Although costly and hardly feasible for laboratories which do not own their MS machine, it can be stated that more replicates assure a higher quality of the results and could be considered good practice.

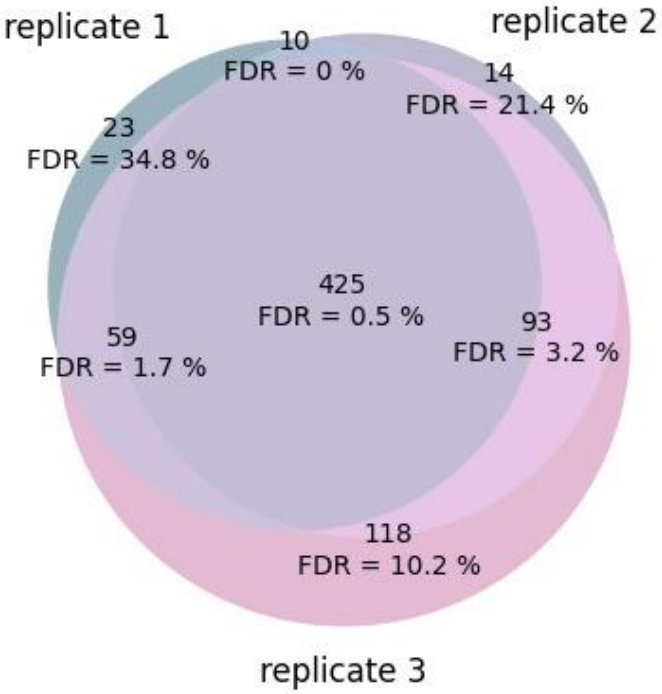


Figure 40: Venn Diagram of the three DSSO replicates (with the main library) analysed by MS Annika (with a set FDR of 1%). The FDR displayed show the actual false discovery rates of the different data sets.

A cost-friendlier way to reduce the FDR of the sample without producing multiple replicates, but still with a moderate impact on the quantity of the findings, would be to analyse the same sample with multiple XL-search algorithms. In the figure bellow it is illustrated how a combination of two software (Annika vs MeroX, XlinkX and pLink2) brings the FDR score to 1.3-1.4%. The loss in results can reach up to 202 XLs, but the sets present in only one software reach numbers up to an FDR=31.1%.

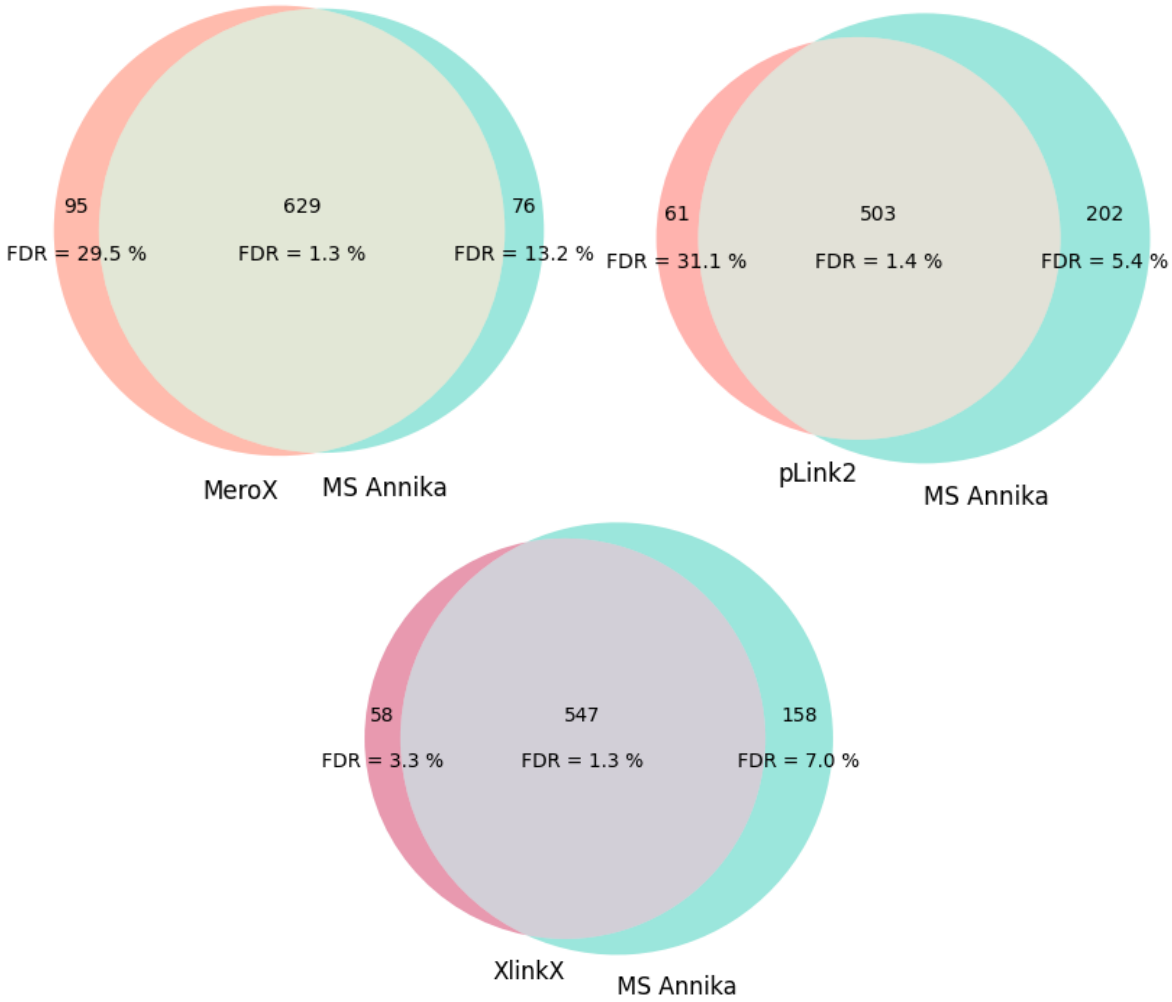


Figure 41: Venn diagram of the replicate number 3 of DSSO-main library experiment with regard to the unique number of links and the specific FDR value to each particular data set.

DHSO linked with the acidic library yields relatively low score independent of the software utilised. Most probably, this is because of the low reactivity of hydrazines rather than a cross-software recognition incapacity (fragmentation chemistry remains unchanged and therefor enough signals are produced). MeroX finds on average again more unique links ($\bar{x}_{\text{correct crosslinks}}$= 76) with an insignificantly higher FDR of 5% (compared to 4.9% from MS Annika). As a whole, all algorithms perform the same with a range between 71 and 80 found links, but the false discovery rates are higher in the case of XlinkX with 15.5% and pLink2 with 13.7%. Interestingly, all post-processed results show an FDR score of 0.0%, but come with a quantitative loss of 70% and 72% in the case of XlinkX and pLink2, respectively.

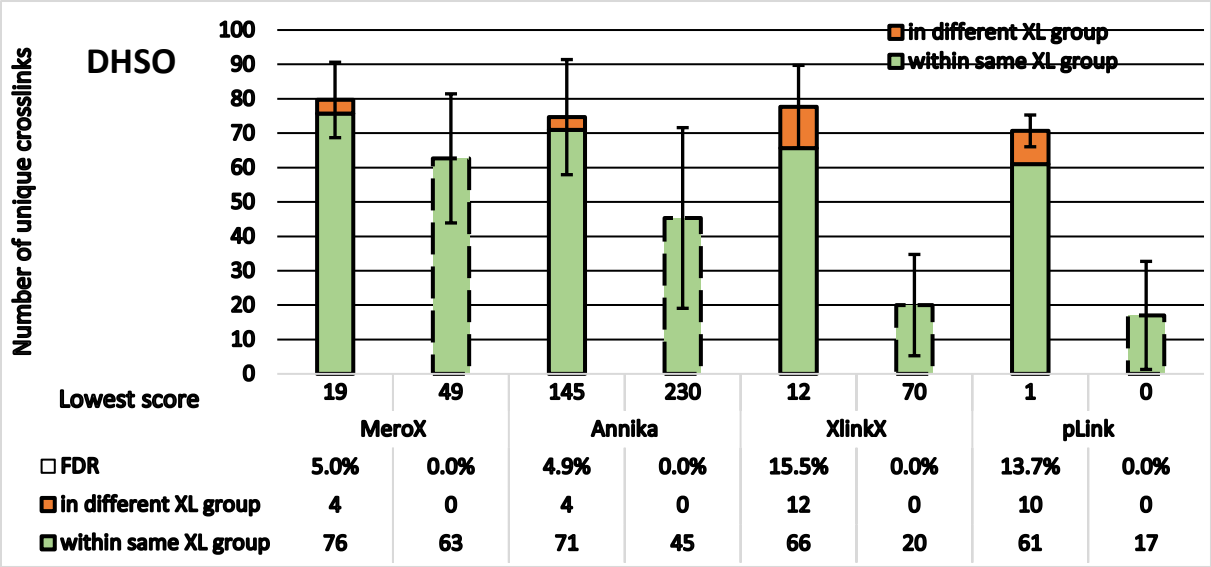
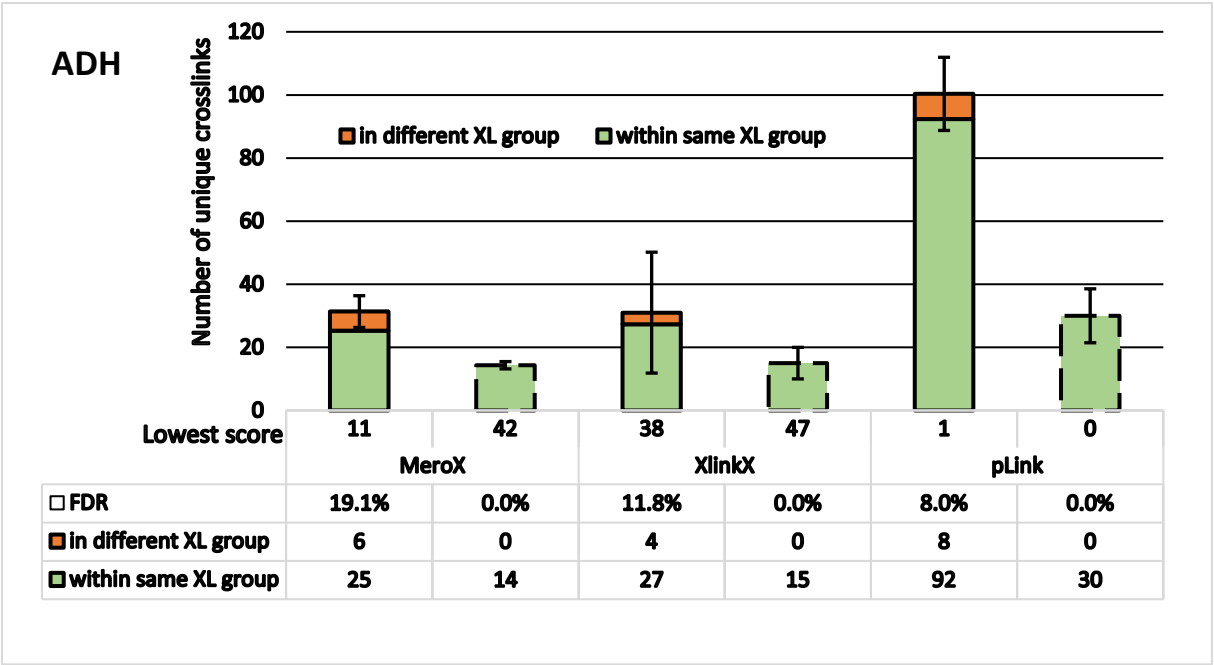


Figure 42: The mean unique crosslink number reported from each algorithm (MeroX, MS Annika, XlinkX and pLink2) benchmarked (sample size n= 3) in the case of DHSO-acidic peptide library. The solid lines refer to the results obtained with a set FDR value of 1%. The long-dashed lines represent the average results after the post-processing to bring them to an experimentally validated FDR value of 1%. The error bars show the standard deviation from each sample with regard only to the total number of links. The lowest score row shows (with the SD attached) the score of the worst graded accepted link in the data set.

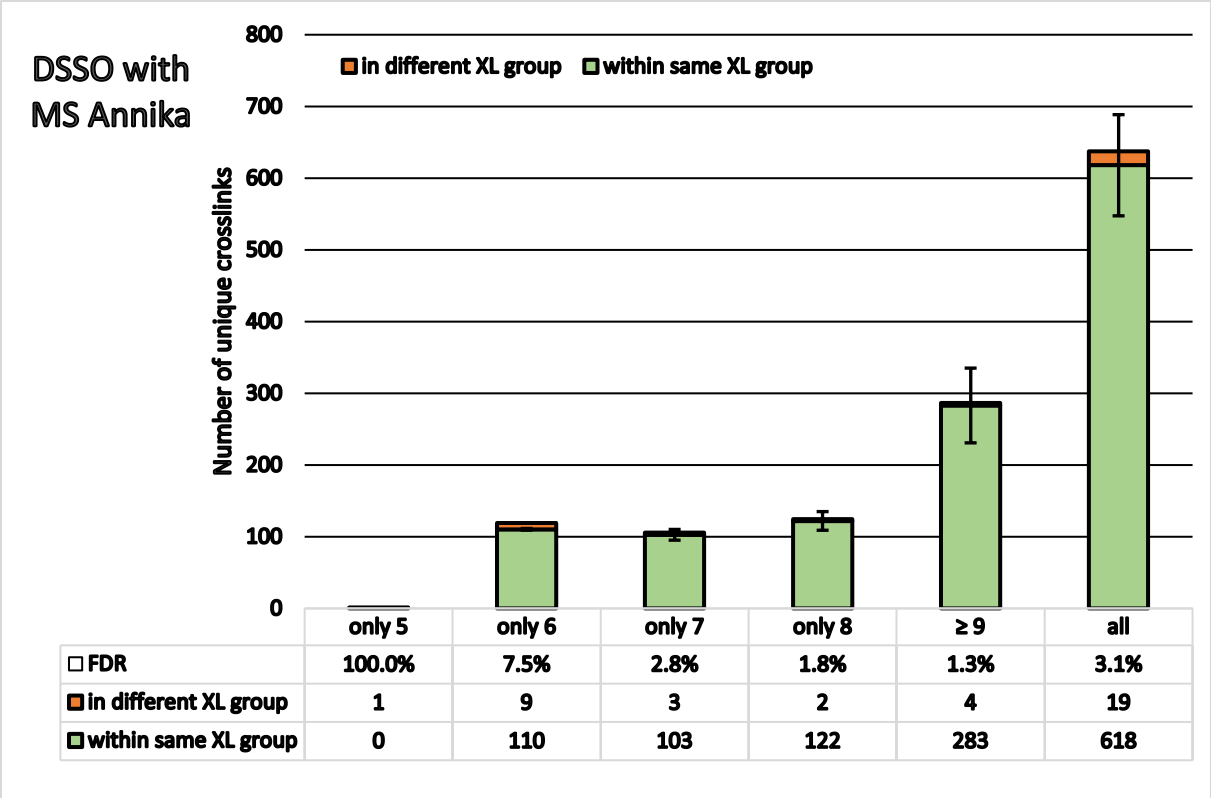
ADH was the most challenging linker reagent to benchmark with the chosen software because of its uncleavable nature. MeroX and XlinkX produce very poor results. In the case of XlinkX, the average number of links has ~59% drop in correct crosslinks compared to its performance with DHSO and yields only 9.6% (correct XLs) of the practically allowed links in the acidic library. On the other hand, pLink2 shows an average of 100 XLs, but still with a relatively high FDR of 8%. pLink2 outperforming all other engines is expected when referring to its fine-tuning on uncleavable reagents. While it is true that the adipic acid dihydrazide has a shorter spacer arm (11.1 Å)[17] than DHSO (12.4 Å)[102], this significant drop in crosslinks is a multifactorial variable mostly driven by the uncleavable nature of the linker. Similarly, as in DHSO, the post processed ADH results have all a FDR of 0.0%.



Lowest score	11	42	38	47	1	0
	MeroX		XlinkX		pLink	
FDR	19.1%	0.0%	11.8%	0.0%	8.0%	0.0%
in different XL group	6	0	4	0	8	0
within same XL group	25	14	27	15	92	30

Figure 43: The mean unique crosslink number found with XL-search engines (MeroX, XlinkX and pLink2) tested (sample size n= 3) in the case of uncleavable ADH-acidic peptide library. The solid lines refer to the results obtained with a set FDR value of 1%. The long-dashed lines represent the average results after the post-processing to bring them to an experimentally validated FDR value of 1%. The error bars show the standard deviation from each sample with regard only to the total number of links. The lowest score row shows (with the SD attached) the score of the worst graded accepted link in the data set. MS Annika is not present here because of its capacities are specialised only on cleavable linkers.

The influence of the accepted peptide length on the correctness of results was tested as well. As expected, with a shorter peptide sequence, the false discovery rate increases. The wrongly annotated spectra in the case of shorter peptides are justified through fewer generated fragments. The number of redundant sequences becomes higher as the database increases.[79,103] This redundancy problem occurs more often in shorter sequences. Even though this aspect cannot be followed in the synthetized peptide library, because an ambiguous crosslink (a link annotated to two or more proteins) is still judged as correct or false only with regard to the belonging or not to the same group. The reason why the FDR is 100% for the peptide length = 5 is because no such peptide is present in the designed library. A minimal acceptable peptide length would be 6 or even 7. The FDR in the data set of XLs containing 6 residues long peptides is comparatively high (7.5%), but its exclusion would lead to ~18% loss of data.



	only 5	only 6	only 7	only 8	≥ 9	all
FDR	100.0%	7.5%	2.8%	1.8%	1.3%	3.1%
in different XL group	1	9	3	2	4	19
within same XL group	0	110	103	122	283	618

Figure 44: Mean number of XLs obtained from DSSO with a certain restrain from the results obtained in MS Annika integrated with Proteome Discoverer in the case of DSSO linked with the main library it was filtered depending on the peptide length. The peptide length refers to the shorter peptide sequence in the XL-pair. The chosen FDR was 1% and the search was done against a database containing ribosomal proteins from E. coli organism identified with a shotgun analysis.

There was an interest to check if the FDR differs between inter-crosslinks and intra-crosslinks. As expected, lower FDR scores are obtained in the case of intralinks. Some XL-search software does possess the option to choose a separate FDR calculation in the case of intra/inter-links in order for them to be treated with different stringency levels. In all cases, the FDR was either not at all or minimally impacted by the two settings of this option, except for XlinkX. Better results are delivered if the separate FDR calculation filter is turned off. An approximate 18.5% is reached for inter-XLs (filter turned on) compared a FDR of ~4.5% with a minimal loss of only 7 correct links (filter turned off). This aspect was completely unobservable in Beveridge's Library as all its sequences originate from one single protein.

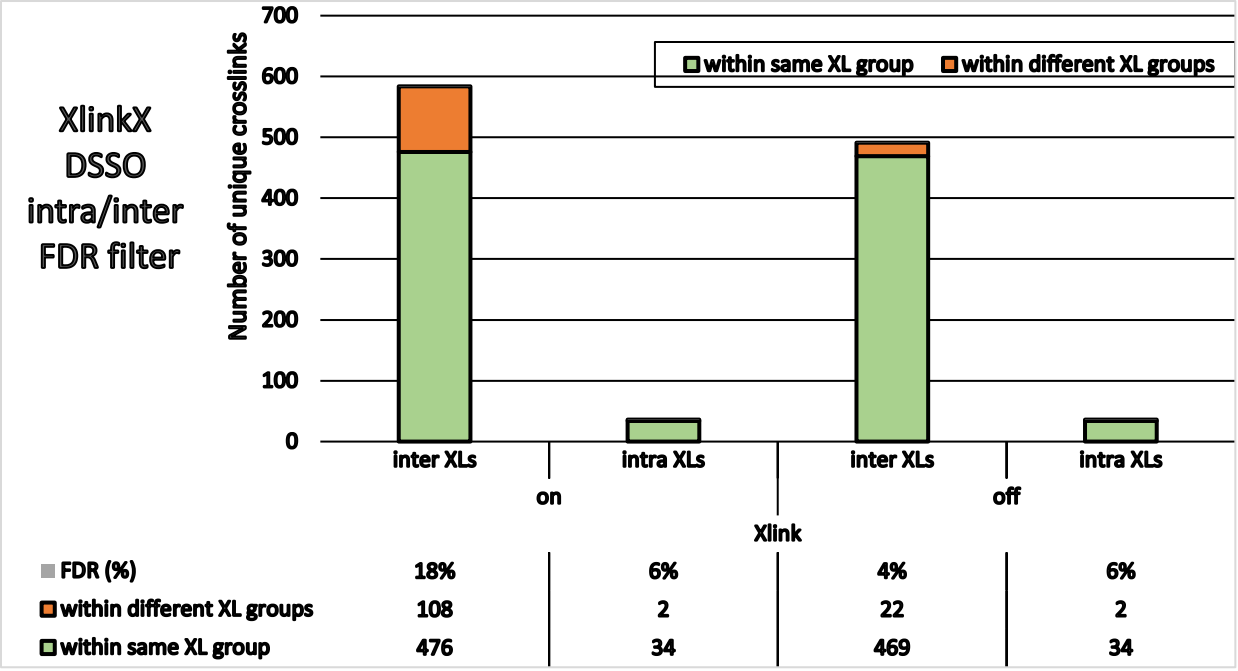


Figure 45:Average number of unique links detected in the DSSO-main library experiment with a HCD MS2 acquisition strategy. Interlinks and Intralinks are displayed separately. The analysis is shown just in the case of XlinkX because no discrepancies were observed in the case of other software. The sample size is n=3 with a set FDR =1% and against ribosomal peptides from E. coli as database.

3.3 Study case – unexpected development of the FDR processes

In some of the analyses, an anomalous behaviour was observed, which highlights one of the sensible points of FDR calculation. It is exemplified beautifully in the analysis below of the main library-DSBU experiment with an FDR setting of 1% and E. coli ribosomal proteins as database at disposal. There is a smooth decrease in the number of correct (heterotypic), correct homeotypic and false crosslinks with a gradually increasing score. Both homeotypic and false links slowly vanish as the score reaches very high values (above 300, which would be considered a very high score for Annika). However, in the post-score cut-off diagram where the number of links is reduced from a real FDR of 4.2% to a desired value of 1%, the number of correct links massively drops. This shows that a significantly high cost of quantity is paid over the sought-after quality. In the FDR development function dependent on the score, a so-called ”saw teeth” in the range of higher scores can be observed. Prolonged positive evolutions of the FDR value followed by a sudden drop (still higher than the value where the positive evolution started) are very cost-inefficient, because many correct high-score links are sacrificed in order to eliminate only one false link.

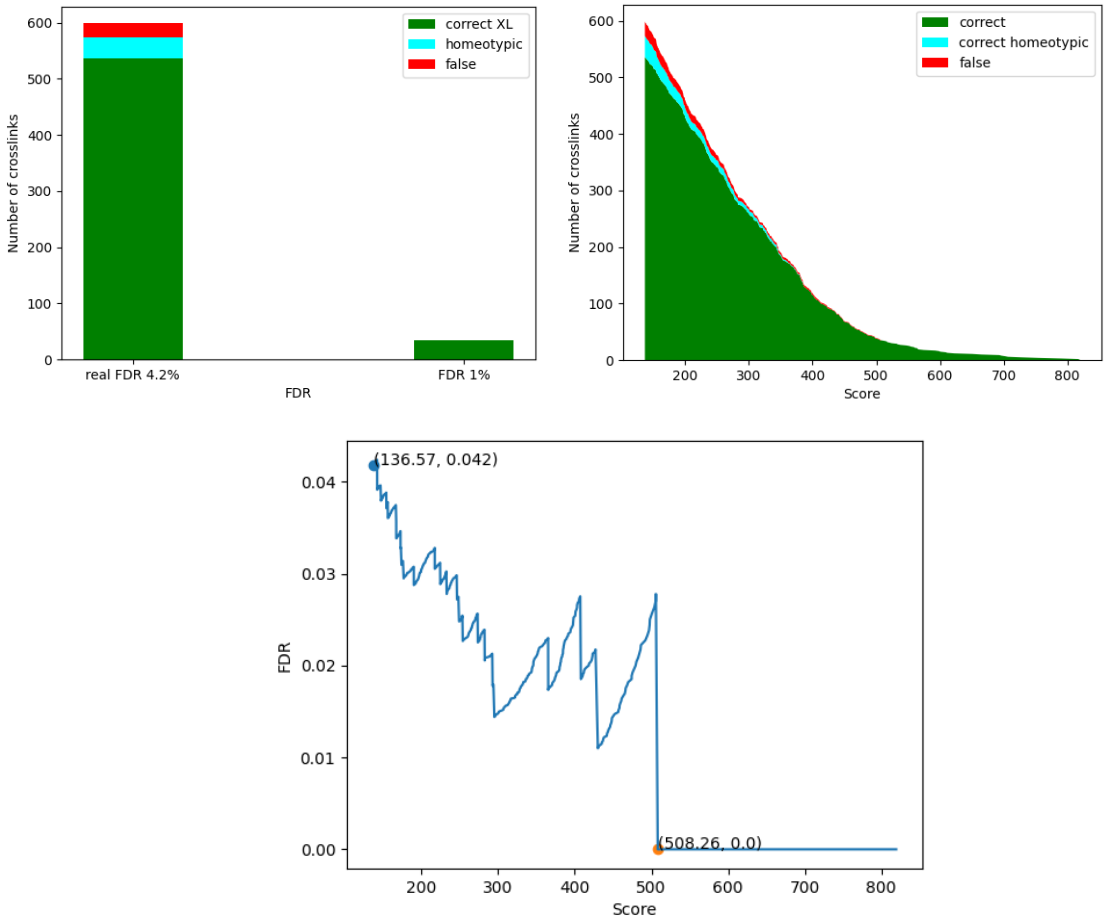


Figure 46: The diagrams above refer to replicate 1 from the main peptide library linked with disuccinimidyl urea and analysed with Annika as an integrative workflow in Proteome Discoverer. The set FDR was 1% and ribosomal proteins found in a MS-shotgun analysis were used as data base; (upper-left) crosslinks vs FDR with and without post-score cut-off; (upper-right) number of remaining crosslinks depending on the score evolution; (central-bottom) real FDR depending on the evolution of the score.

With the purpose of ensuring that these false crosslinks are rather a result of algorithmic misinterpretation of the signals done by the crosslink software, and not caused by a human mistake, a more detailed look was taken at the spectrum with a score of 508.26 (Annika noting scale). Peptide α does exist (ID number 3, present in group 1 and 13), but peptide β is not present in the peptide library. Furthermore, every single peptide from each utilized group in every library was rechecked via MALDI analysis in order to exclude possible contaminations (of one peptide in other peptide solution) caused by humans. As shown in the fasta-format entry of the origin-protein below, the "TK" would be the extension of the sequence present in the main library. This is most likely a misinterpretation by Annika as (KAGNVAADGVIK-KAGNVAADGVIKTK), even though the real crosslink would fit to the homeotypic pair (KAGNVAADGVIK- KAGNVAADGVIK). For a manual validation, one must look if most peaks present in the spectrum are annotated, and if there are peaks with a high relative intensity which have not been annotated (possible indicator of a miss-annotation).

>sp|P0A6P1|EFTS_ECOLI Elongation factor Ts OS=Escherichia coli (strain K12) OX=83333 GN=tsf PE=1 SV=2
MAEITASLVKELRERTGAGMMDCKKALTEANGDIELAIENMRKSGAIKAAK**KAGNVAADGVIK**TKIDGNYGIILEV
NCQTDFVAKDAGFQAFADKVLDAAVAGKITDVEVLKAQFEEERVALVAKIGENINIRRVAALEGDVLGSYQHGARI

GVLVAAKGADEELVKHIAMHVAASKPEFIKPEDVSAEVVEKEYQVQLDIAMQSGKPKEIAEKMVEGRMKKFTGEVS
LTGQPFVMEPSKTVGQLLKEHNAEVTGFIRFEVGEGIEKVETDFAAEVAAMSKQS

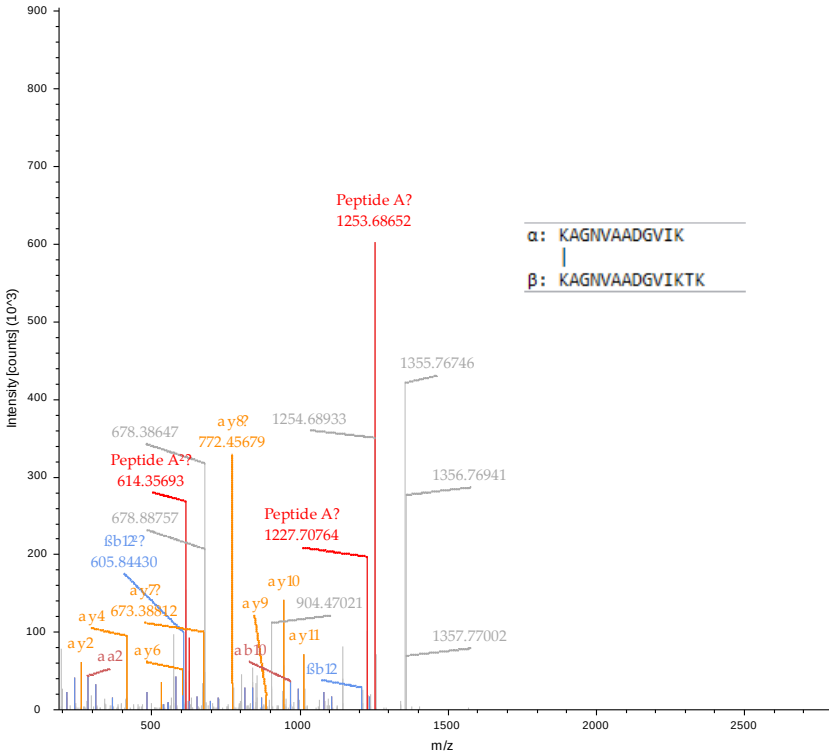


Figure 47: Spectrum annotated to the best scored false crosslink from the first replicate of the main library-DSBU experiment. It is represented as intensity counts vs m/z. Spectrum is a capture from the Proteome Discoverer's graphical user interface and the unannotated signals are shown in grey colour.

On the other hand, this seems to be restricted to one software. MeroX does not identify this crosslink, nor annotates the spectrum.

Another prominent example is the homeotypic crosslink MAKLTK-MAKLTK found in all three replicates of the main library-DSSO experiment analysed by MS Annika. The peptide does not exist in any of the synthetic libraries, but there is a peptide with the same molecular weight in the main library with the sequence MAKTIK (ID number 16, present in group 2 and 16). In this case, because one of the peptides is so small, the chance was high to find a similar sequence with an identical molecular mass. The fragmentation patterns produced by shorter sequences cannot be very complex. However, the software should have been able to correctly differentiate between those two peptides if a fragmentation would have occurred between threonine and isoleucine.

3.4. Physicochemical analysis of different crosslink data sets

We analysed then the sample obtained from crosslinking the main synthetic peptide library with DSSO following the usual crosslinking-protocol in different amounts in order to gain new insights about the necessary amount of sample to detect as many links as possible with the least amount of sample. The different amounts were prepared via a dilution series. Not only that, but the

physicochemical properties were investigated as well to establish which properties make a crosslink more likely to be detected in reduced amounts. This type of analysis, where the absolute number of crosslinks/CSMs are not counted, but rather the properties of the CSMs in a population unveiled some meaningful trends. Because of this reason, the same analysis was also applied to the data where different linkers were benchmarked, and the sample amount was kept constant.

It is worth mentioning that the measurements of the same sample in different quantities were performed with an Exploris machine (FAIMS turned on), while all the linker-benchmarking experiments were measured with QExactive HF-X. The data shown below represents the properties only from the true crosslinks (unless specified otherwise). For a better comparison, the properties of all theoretically allowed links were included as well and noted with “theory” or “theory DSSO”.

The calculated GRAVY values of the links in the gradually increasing sample amount experiment show a similar spread of data for all amounts and theoretical results. While the data from 25 ng to 800 ng shows a slight shift towards lower GRAVY values (and implicitly favours more hydrophobic XLs), the results from 10 ng are strongly shifted upward with the median having a GRAVY index around 0. This is rather unexpected because hydrophobic peptides tend to stick to the walls of the Eppendorf Tubes (even to the low binding Eppendorf Tubes that are now the standard in almost each laboratory). When comparing the data across different crosslinkers, no changes in the spread, median of shifting stand out.

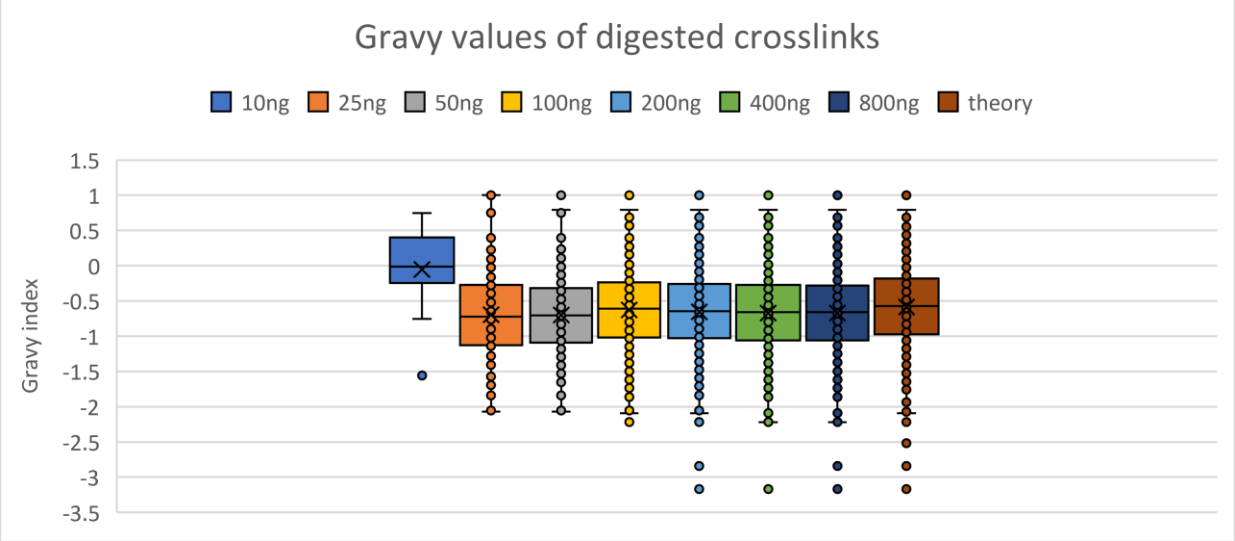


Figure 48: Boxplot representations of the GRAVY values calculated by IMP-X-FDR physicochemical cross-link properties of the main peptide library crosslinked with DSSO and analysed on Exploris MS in different amounts; “theory” represents the boxplot representation of the values for all theoretically possible crosslinks; the analysis was done using the digested form of the synthetic peptides.

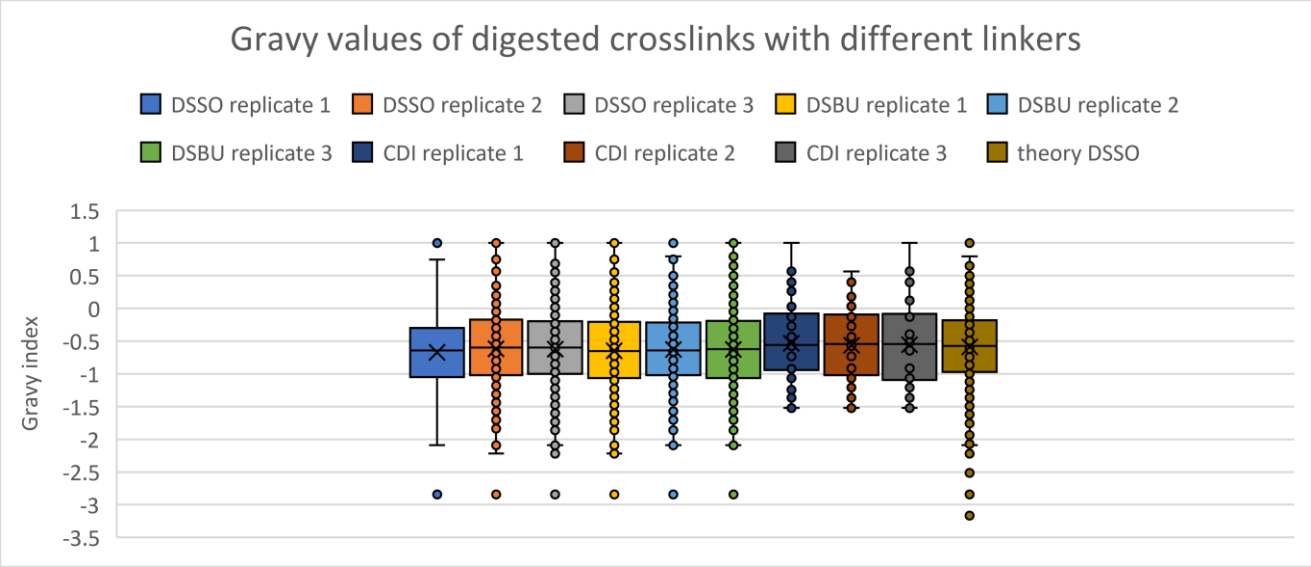


Figure 2: Boxplot representations of the GRAVY values calculated by IMP-X-FDR physicochemical cross-link properties of the main peptide library (ca 1000 ng) crosslinked with different linkers and analysed on QExactive HF-X; "theory DSSO" represents the boxplot representation of the values for all theoretically possible crosslinks; the analysis was done using the digested form of the synthetic peptides.

The importance of the right amount of analysed sample is underlined as well when analysing the frequency of tryptophan, phenylalanine, and tyrosine. The samples where 10 ng, 25 ng and 50 ng of crosslink mix was injected show different population parameters compared to the rest, having on average a lower frequency of these aromatic residues. Throughout different utilized linkers, the differences seem to vanish with the exception of 0 length crosslinker (CDI). For 1,1'-carbonyldiimidazole, median is higher in all replicates and the data is skewed and has a shorter upper quantile denoting a conglomeration of links with higher percentages of aromatic residues. It must not be forgotten that the absolute results in the case of CDI differ strongly from the others (ca. 50 crosslinks identified in comparison to >500 by using DSSO or DSBU).

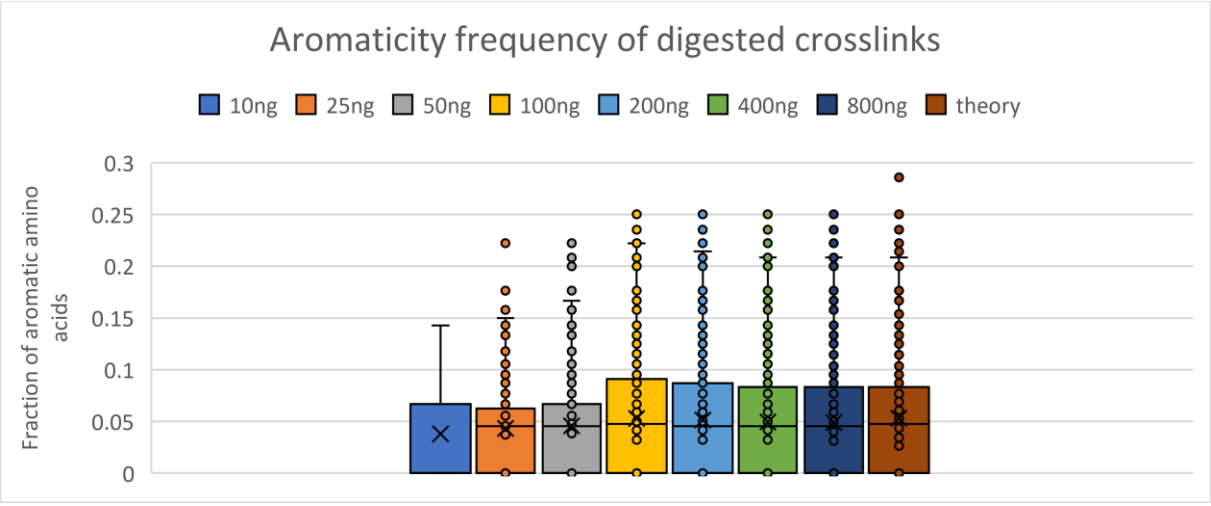


Figure 3: Boxplot representations of the fraction of aromatic amino acids calculated by IMP-X-FDR physicochemical cross-link properties of the main peptide library crosslinked with DSSO and analysed on Exploris MS in different amounts; "theory" represents the boxplot representation of the values for all theoretically possible crosslinks; the analysis was done using the digested form of the synthetic peptides.

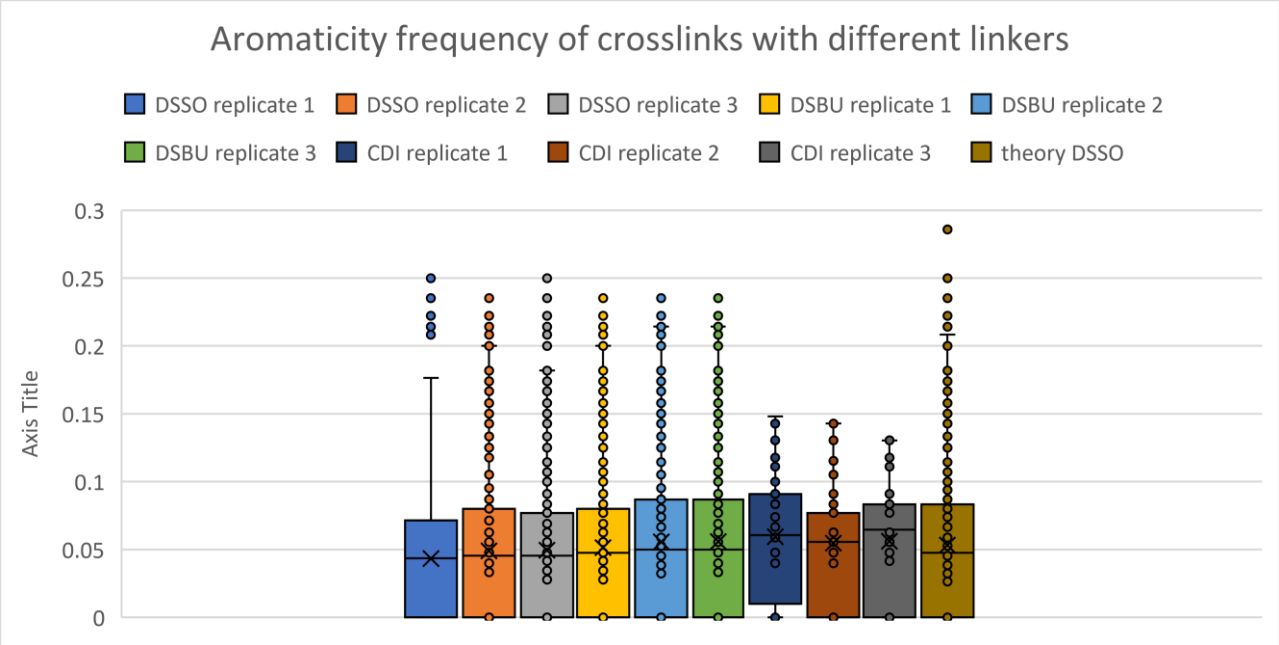


Figure 4: Boxplot representations of the fraction of aromatic amino acids calculated by IMP-X-FDR physicochemical cross-link properties of the main peptide library (ca. 1000 ng) crosslinked with different linkers and analysed on QExactive HF-X; "theory DSSO" represents the boxplot representation of the values for all theoretically possible crosslinks; the analysis was done using the digested form of the synthetic peptides.

Since all peptides in the data sets presented (except for the acidic library) contain one lysine and one arginine by default, it is expected to see both median and average with a calculated isoelectric point above 7. From 10 ng to 100 ng there is a descending pI value in the median and average that stabilizes starting from 200 ng. The theoretical results have a lower median. This suggests that the presence of acidic residues has an inhibitory effect on the crosslink detection (possibly because of weaker ionization) especially when lower sample amounts are being used. The use of various linkers does not impact which peptides are being crosslinked based on their respective isoelectric points.

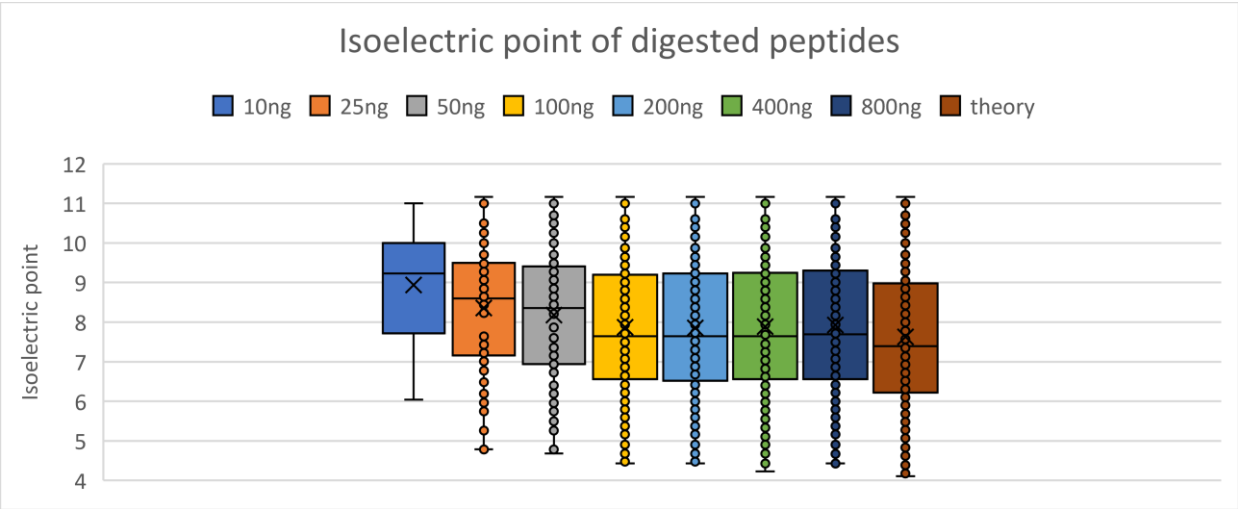


Figure 5: Boxplot representations of the isoelectric points calculated by IMP-X-FDR physicochemical cross-link properties of the main peptide library crosslinked with DSSO and analysed on Exploris MS in different amounts; "theory" represents the boxplot representation of the values for all theoretically possible crosslinks; the analysis was done using the digested form of the synthetic peptides.

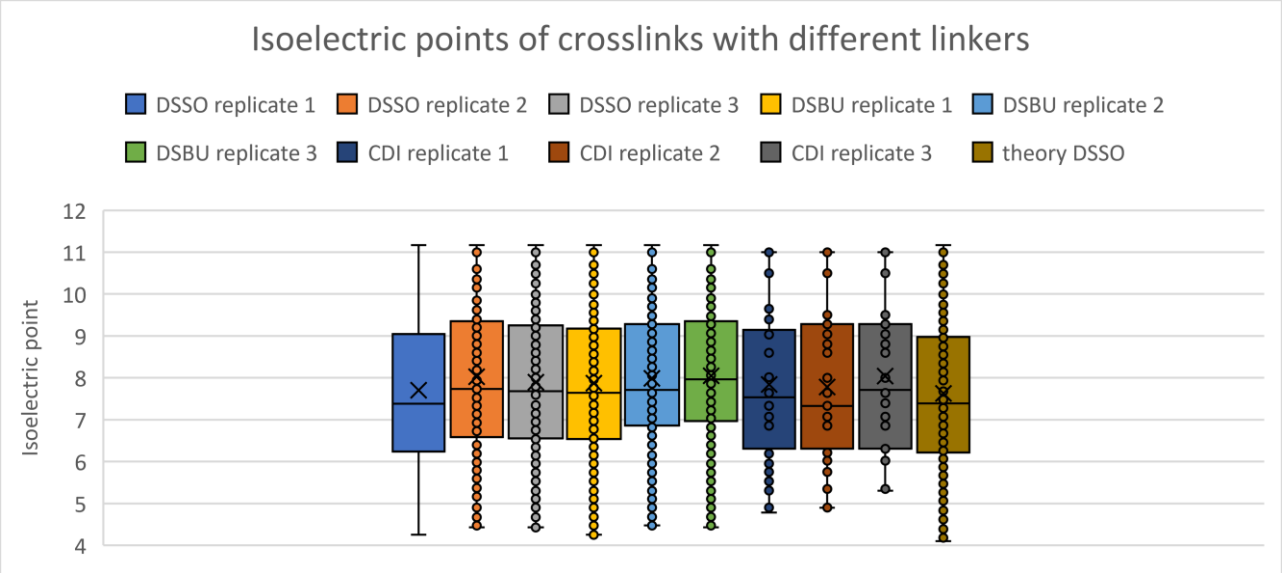


Figure 6: Boxplot representations of the isoelectric points calculated by IMP-X-FDR physicochemical cross-link properties of the main peptide library (ca. 1000 ng) crosslinked with different linkers and analysed on QExactive HF-X; "theory DSSO" represents the boxplot representation of the values for all theoretically possible crosslinks; the analysis was done using the digested form of the synthetic peptides.

The molecular weight is the property that most influences the possibility of a crosslink to be successful ionized and hence detected by a crosslink search engine. In the experiment, where different amounts of sample are used for measurements, the median and average mass of the detected XLs increase gradually. From 10 ng to 50 ng, only the low weight crosslinks are being identified. The trend adjusts starting from 100 ng, but still differs strongly from the theoretical results. One must wisely decide which amount should be measured taking into account two aspects: it must not overload the machine and it should be high enough to get a detailed image of the non-covalent chemical interactions.

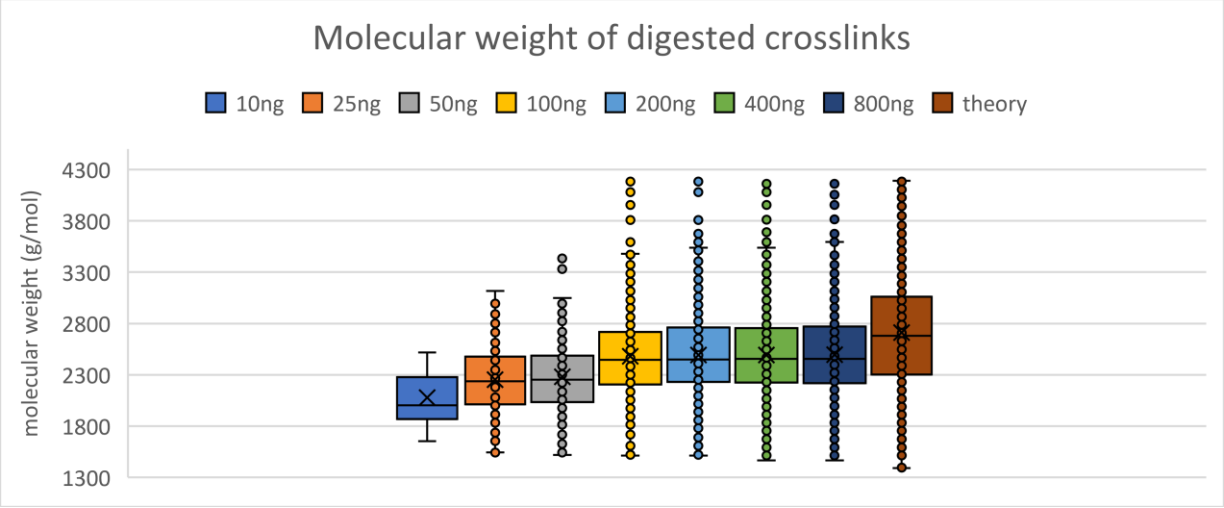


Figure 7: Boxplot representations of the molecular weights calculated by IMP-X-FDR physicochemical cross-link properties of the main peptide library crosslinked with DSSO and analysed on Exploris MS in different amounts; "theory" represents the boxplot representation of the values for all theoretically possible crosslinks; the analysis was done using the digested form of the synthetic peptides.

When looking at the molecular weight of links when different linkers are engaged, it is visible that the variability of the data between different replicates of the same sample (where a specific linker was used) is low demonstrating robustness. Between different linkers, changes in the populations are

observable. The interquartile range of the DSBU populations are wider in comparison to the ones from DSSO, and the IQRs from the CDI samples are the largest. The medians and averages are also slightly larger in the data obtained by using DBSU instead of DSSO. The medians of the samples where CDI is used are lower than the rest indicating that practically no high-mass link was found. None of these linkers reassemble perfectly the box-plot of the theoretical results (with DSSO as crosslinker) which has a shift towards higher masses. The false crosslinks from the first replicate of the experiment with DSSO as linker were isolated and measured separately. The median and average have an abrupt shift toward very low molecular weights. This is explicable through the fact that shorter sequences create fewer and shorter fragments. The chances are higher that a random mass from another sequence would fit under such circumstances.

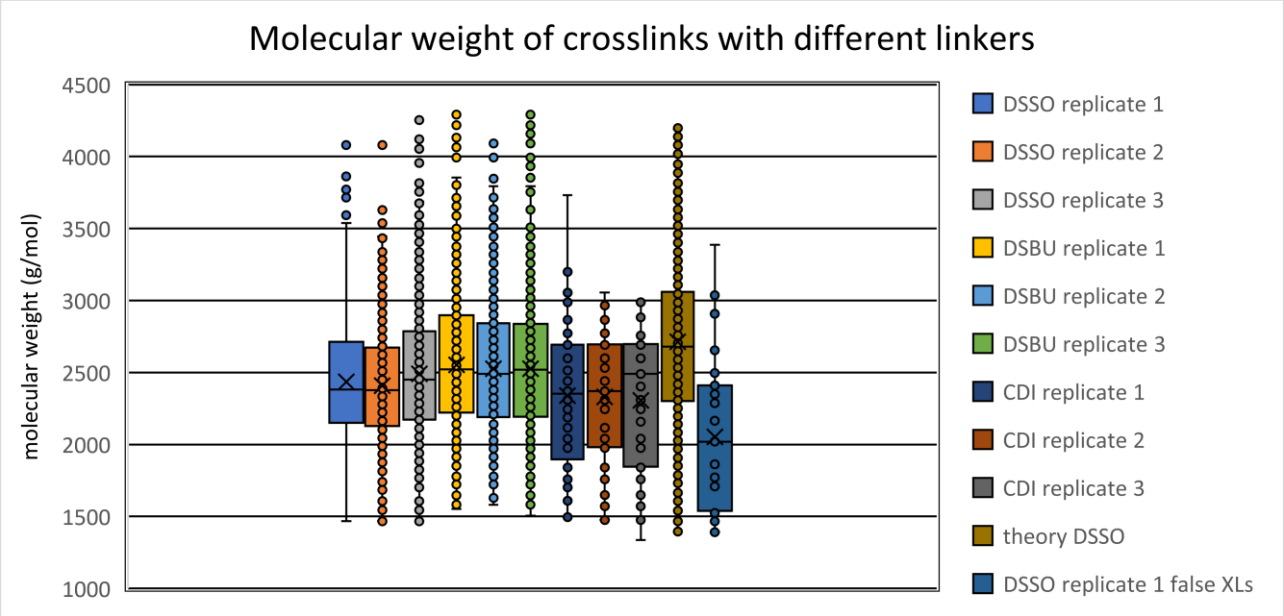


Figure 8: Boxplot representations of the isoelectric points calculated by IMP-X-FDR physicochemical cross-link properties of the main peptide library (ca 1000 ng) crosslinked with different linkers and analysed on QExactive HF-X; "theory DSSO" represents the boxplot representation of the values for all theoretically possible crosslinks; "DSSO replicate 1 false XLs" sows only the false crosslinks founds by Annika in the first DSSO replicate in contrast to all the other data where just correctly identified links are shown; the analysis was done using the digested form of the synthetic peptides.

3.5. Main motifs

In the case of benchmarking the linkers, the main motif theoretically calculated seem to be mostly reflected in the experimentally validated results. This is a very important aspect, because it shows that DSSO, DSBU, CDI yields with lysine residues are not influenced by the order or types of neighbouring residues. Strong deviations would have meant a sequence-dependent preferential rection kinetics, which is undesirable. Some deviations do occur when different sample amounts are measured.

Sample origin	Main motif
DSSO replicate 1	GGAKA--
DSSO replicate 2	GGAKAA-
DSSO replicate 3	GGAKA--
DSBU replicate 1	GGEKAA-
DSBU replicate 2	GGAKAA-
DSBU replicate 3	GGAKAA-
CDI replicate 1	GGIKAD-
CDI replicate 2	GGIKAA-
CDI replicate 3	GGIKADR
Theory	GGAKAA-

Sample origin	Main Motif
DSSO 10 ng	AVIKAVR
DSSO 25 ng	GGRKAR-
DSSO 50 ng	GGEKAA-
DSSO 100 ng	GGAKAT-
DSSO 200 ng	GGAKAT-
DSSO 400 ng	GGAKAA-
DSSO 800 ng	GGAKAT-
Theory	GGAKAA-

Due to the fact that hydrazine-based linkers are rather new, and their application area is narrowed down by the higher efficiency of the NHS-esters, which target lysines primarily, their reaction chemistry was not intensively researched. By looking at the most abundant neighbouring residues described in the table below, it was also intended to establish if some specific patterns confer a higher yielding of some specific crosslinks. DHSO experiment results for the acidic library are consistent with the ones obtained where the acidic peptide library was pooled (a bigger data sets for increased statistical relevance) demonstrating robustness of the method. Not only that, but the motifs are also mostly in agreement with the theoretically calculated ones. The same statements are true when using ADH as reagent (same chemistry, but different in fragmentation behaviour). There does not seem to be a specific residue which positively influences crosslink formation.

Sample origin	Main motif
DHSO acidic library replicate 1	GATD/EAVG
DHSO acidic library replicate 2	RATD/EAVG
DHSO acidic library replicate 3	RATD/EAVG
DHSO acidic library Theory	GATD/EAVG
DHSO pooled acidic library replicate 1	RATD/EAVG
DHSO pooled acidic library replicate 2	AATD/EAVG
DHSO pooled acidic library replicate 3	RAAD/EAVG
DHSO pooled acidic library Theory	GAAD/EAGG

The amino acid residues coloured in red were not found in any crosslink, even though their respective uncrosslinked peptides were able to be identified in the mix (when searching for peptides).

Histidine, isoleucine, phenylalanine, tryptophan, tyrosine and glutamine are reproducibly hampering the formation of crosslinks when they are located in the proximity of the binding amino acid (D/E). This behaviour is strengthened by its presence in both DHSO experiment with the acidic library as well as in the pooled peptide library. Out of these 6 residue types, 4 of them present electron-rich aromatic

cycles which can act as nucleophiles. Further research should be conducted in this case to determine if the above-mentioned residues participate in a reaction that has been neglected until now. Single synthetic peptides (for a low complexity) and close monitoring of the fragmentation products while maintaining otherwise the same reaction conditions could provide more insights into the chemistry of this linker with amino acid residues or pairs of amino acids in proximity.

DHSO with the acidic peptide library							
Favouring XL formation ≥60%				D/E	C	C	N
Hindering XL Formation ≥ -60%		F/H/I	F/R/W		N/Q/S/Y	Q/Y	T/W/Y

Figure 49: Effect of the different amino acid residues in the proximity of the crosslinker-binding amino acid (maximum length =3) of the crosslink formation in the acidic library. The percentages correspond to the frequencies of the specific residue type at a given position relative to the crosslinking-binding site normalized by frequencies obtained from all (in silico generated) possible crosslinks from the acidic library system. Amino acids marked in red were not found in any crosslink identified even though the theoretical results prove it as possible, and the linear peptides are identifiable in the non-crosslinked linear peptide mix. The sample size is equal to 3 and the measurements were conducted on separate days.

DHSO on pooled acidic peptide library							
Favouring XL formation ≥60%		Q	G	D/E	C/T		
Hindering XL Formation ≥ -60%		F/H/I/N	F/W		N/Q/S/Y	H/P/Q/Y	T/W/Y

Figure 50: Effect of the different amino acid residues in the proximity of the crosslinker-binding amino acid (maximum length =3) of the crosslink formation in the pooled acidic library (only one group). The percentages correspond to the frequencies of the specific residue type at a given position relative to the crosslinking-binding site normalized by frequencies obtained from all (in silico generated) possible crosslinks from the pooled acidic library system. Amino acids marked in red were not found in any crosslink identified even though the theoretical results prove it as possible, and the linear peptides are identifiable in the non-crosslinked linear peptide mix. The sample size is equal to 3 and the measurements were conducted on separate days.

Theoretical exemple: the amino acid serine is found on the second position in the direction of the N-terminus (-XSXD/EXXX-) in 15% of all peptide sequences of the crosslinks (weighted on CSM level), but the expected frequency obtained from in-silico generated crosslinks on this specific position of the serine is 10%. This would mean that serine on the second position is favouring the formation of the crosslink by 50%.

It must be proved that the peptides with specific neighbours close to the binding site are indeed identifiable, but as linear peptides. In other words, that these peptides are not in some other ways negatively discriminated (depending on charge or bad fragmentation). For this the same strategy was

employed as with crosslinks. This means that only ions with a charge equal or higher than 3 were selected for fragmentation. The results show that more IDs of histidine-containing peptides than theoretically expected could be found. For the other amino acids (except for isoleucine) that could negatively impact XL-formation, the obtained average frequency lies between the expected values. The percentages reinforce the possible negative impact on interpeptidal crosslinking, especially histidine which might cross-react with the active form of carboxylic group and form an intrapeptidal crosslink.

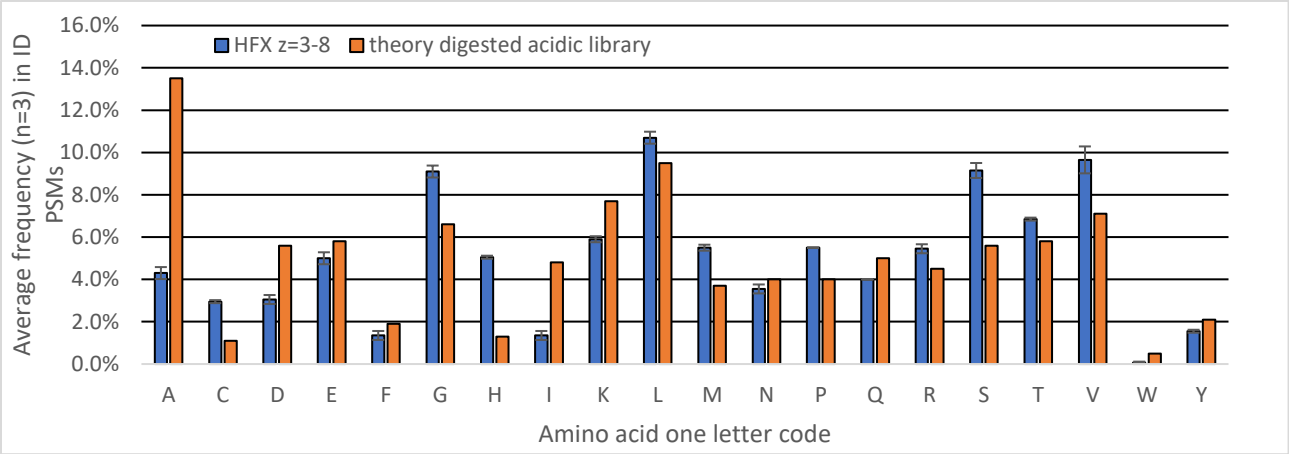


Figure 51: The frequency of amino acids in peptides which were not crosslinked compared to the theoretical expected frequency of the amino acids when equimolarity is assumed. Error bars correspond to the SD form the sample (n=3)

4. Conclusions and Outlook

As an integrative technique in the proteomics research and structural biology, XL-MS has reached a high prominance and popularity due its always-improving nature, but it still plagued by a lack of standardised protocols. There is a large diversity of utilised workflows, linkers, software, and different machines (e.g. Orbitrap Fusion Lumos, Orbitrap Fusion Tribid, Q Exactive Plus or Orbitrap Elite). Not only that, but various resolving power settings, as well as precursor ion tolerance and product ion tolerance must be chosen. Some research groups do specialize in in-solution digestion and others in in-gel digestion. All these parameters may strongly impact the final results. This thesis contributes to the harmonization and benchmarking of different XL-MS protocols.[104]

MeroX shows the best results on DSBU, but demonstrates a commendable level of proficiency across all evaluated linkers. It also has a lot of versatility by offering a large palette of parameters, but it can be somehow tedious to program multiple analysis-jobs. MS Annika outperforms all other engines in the case of DSSO when the FDR is considered as quality criterium. However, MS Annika is not capable of analysing any uncleavable linkers, which XlinkX can with very high time costs. When considering the ID numbers, Merox delivers the best results with a mean of 620 unique links. pLink2 can deliver very fast and more reliable results in case uncleavable linkers are the reagent of choice. Overall, it seems that all XL-search algorithms tested are fine-tuned on one specific linker and tend to perform better with crosslinkers that present the same chemistry.

All algorithms do underestimate their false discovery rate, but in this master thesis we proposed a useful ways of reducing the FDRs. One method would be the generation of more replicates. The more

replicates, the smaller the FDR will become in the overlap region. A bioinformatic workaround is to simply analyse the sample with multiple programs and to utilize the common XLs.

Bioinformatic workarounds had to be found in some cases, as for CDI when using the Proteome Discoverer 2.5 platform. One cannot define a fragment with the mass of 0 Da, even though it is needed and has to choose the lowest mass possible: 1E-5 Da. That could hinder XlinkX and MS Annika to make more accurate findings.

IMP-X-FDR put forward a concept, namely that the identification of a link is not solely a matter of algorithm performance or the chemical reactivity of different reagents with their residue targets, but rather a multifactorial problem. The correct assignment of a link could depend on its mass, GRAVY index, and order of amino acids in the sequence. It is already known that the identification of a peptide/link depends on its ionizability, and the presence of such a tool only accelerates the understanding of what make a XL visible.

Through this resulted data, a ground truth basis for FDR calculation is produced. Future XL-search engines and new software updates of already existing ones can be benchmarked and finetuned based on the data generated in this study. The raw data was published as well to serve future developers and can be found in the PRIDE repository[105] with the ID PXD029252. Finally, the analyses of the obtained data and findings were published[79].

5. List of abbreviations

HEK – human embryonic kidney 293
BSA - bovine serum albumin
DSSO- Disuccinimidyl Sulfoxide
DSBU - Disuccinimidyl Dibutyric Urea
DSBSO - Azide-tagged acid-cleavable disuccinimidyl bissulfoxide
CDI - N,N'-Carbonyldiimidazole
ADH – Adipic Acid Dihydrazide
DHSO – 3,3’-Sulfinyldi(propanehydrazide)
DMTMM- 4-(4,6-dimethoxy-1,3,5-triazin-2-yl)-4-methyl-morpholinium chloride

6. References

1. Cammarata, M. B., Maclas, L. A., Rosenberg, J., Bolufer, A. & Brodbelt, J. S. Expanding the Scope of Crosslink Identifications by Incorporating Collisional Activated Dissociation and Ultraviolet Photodissociation Methods. *Anal. Chem.* **90**, 6385 (2018).

2. Hage, C., Iacobucci, C., Rehkamp, A., Arlt, C. & Sinz, A. The First Zero-Length Mass Spectrometry-Cleavable Cross-Linker for Protein Structure Analysis. *Angew. Chem. Int. Ed. Engl.* **56**, 14551–14555 (2017).

3. Merkley, E. D. *et al.* Distance restraints from crosslinking mass spectrometry: Mining a molecular dynamics simulation database to evaluate lysine–lysine distances. *Protein Sci.* **23**, 747–759 (2014).

4. Kahraman, A. *et al.* Cross-Link Guided Molecular Modeling with ROSETTA. *PLoS One* **8**, e73411 (2013).

5. Tang, X., Munske, G. R., Siems, W. F. & Bruce, J. E. Mass Spectrometry Identifiable Cross-Linking Strategy for Studying Protein−Protein Interactions. *Anal. Chem.* **77**, 311–318 (2004).

6. Chavez, J. D. *et al.* Chemical Crosslinking Mass Spectrometry Analysis of Protein Conformations and Supercomplexes in Heart Tissue. *Cell Syst.* **6**, 136-141.e5 (2018).

7. Chavez, J. D. *et al.* Systems structural biology measurements by in vivo cross-linking with mass spectrometry. *Nat. Protoc. 2019 148* **14**, 2318–2343 (2019).

8. 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).

9. Merkley, E. D. *et al.* Distance restraints from crosslinking mass spectrometry: Mining a molecular dynamics simulation database to evaluate lysine–lysine distances. *Protein Sci.* **23**, 747–759 (2014).

10. Heitz, J. R., Anderson, C. D. & Anderson, B. M. Inactivation of yeast alcohol dehydrogenase by N-alkylmaleimides. *Arch. Biochem. Biophys.* **127**, 627–636 (1968).

11. Iacobucci, C., Piotrowski, C., Rehkamp, A., Ihling, C. H. & Sinz, A. The First MS-Cleavable, Photo-Thiol-Reactive Cross-Linker for Protein Structural Studies. *J. Am. Soc. Mass Spectrom.* **30**, 139–148 (2019).

12. 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. 2018 1312* **13**, 2864–2889 (2018).

13. Dormán, G. & Prestwich, G. D. Benzophenone Photophores in Biochemistry. *Biochemistry* **33**, 5661–5673 (1994).

14. Belsom, A., Mudd, G., Giese, S., Auer, M. & Rappsilber, J. Complementary Benzophenone Cross-Linking/Mass Spectrometry Photochemistry. *Anal. Chem.* **89**, 5319–5324 (2017).

15. Ziemianowicz, D. S., Bomgarden, R., Etienne, C. & Schriemer, D. C. Amino Acid Insertion Frequencies Arising from Photoproducts Generated Using Aliphatic Diazirines. *J. Am. Soc. Mass Spectrom.* **28**, 2011–2021 (2017).

16. West, A. V. *et al.* Labeling Preferences of Diazirines with Protein Biomolecules. *J. Am. Chem. Soc.* **143**, 6691–6700 (2021).

17. Leitner, A. *et al.* Chemical cross-linking/mass spectrometry targeting acidic residues in proteins and protein complexes. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 9455–9460 (2014).

18. Mohammadi, A., Tschanz, A. & Leitner, A. Expanding the Cross-Link Coverage of a Carboxyl-Group Specific Chemical Cross-Linking Strategy for Structural Proteomics Applications. *Anal. Chem.* **93**, 1944–1950 (2021).

19. Sutherland, B. W., Toews, J. & Kast, J. Utility of formaldehyde cross-linking and mass spectrometry in the study of protein–protein interactions. *J. Mass Spectrom.* **43**, 699–715 (2008).

20. 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).

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

22. Piersimoni, L. & Sinz, A. Cross-linking/mass spectrometry at the crossroads. *Anal. Bioanal. Chem.* **412**, 5981–5987 (2020).

23. Kendrew, J. C. *et al.* A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181**, 662–666 (1958).

24. Kendrew, J. C. *et al.* Structure of myoglobin: A three-dimensional fourier synthesis at 2 . resolution. *Nature* **185**, 422–427 (1960).

25. Perutz, M. F. *et al.* Structure of Hæmoglobin: A three-dimensional fourier synthesis at 5.5-. resolution, obtained by X-ray analysis. *Nature* **185**, 416–422 (1960).

26. Pietzsch, J. The Nobel Prize in Chemistry 1962 - Perspectives: Cracking the Phase Problem. https://www.nobelprize.org/prizes/chemistry/1962/perspectives/.

27. 1962 - John Kendrew & Max Perutz - MRC Laboratory of Molecular Biology. http://www2.mrc-lmb.cam.ac.uk/achievements/lmb-nobel-prizes/1962-john-kendrew-max-perutz/.

28. Wüthrich, K. NMR with Proteins and Nucleic Acids. *Europhys. News* **17**, 11–13 (1986).

29. Wüthrich, K. NMR studies of structure and function of biological macromolecules (Nobel Lecture). *Journal of Biomolecular NMR* vol. 27 13–39 https://www.nobelprize.org/prizes/chemistry/2002/wuthrich/lecture/ (2003).

30. Callaway, E. Molecular-imaging pioneers scoop Nobel. *Nature* **550**, 167 (2017).

31. Protein DataBank. PDB Statistics: Number of Released PDB Structures per Year. https://www.rcsb.org/stats/all-released-structures (2022).

32. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583 (2021).

33. Kryshtafovych, A. *et al.* Evaluation of the template-based modeling in CASP12. *Proteins Struct. Funct. Bioinforma.* **86**, 321–334 (2018).

34. Croll, T. I., Sammito, M. D., Kryshtafovych, A. & Read, R. J. Evaluation of template-based modeling in CASP13. *Proteins Struct. Funct. Bioinforma.* **87**, 1113–1127 (2019).

35. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins Struct. Funct. Bioinforma.* **87**, 1011–1020 (2019).

36. DeepMind. AlphaFold : a solution to a 50-year-old grand challenge in biology. *Https://Deepmind.Com/Blog* https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology%0Ahttps://deepmind.com/blog/article/alphafold-a-

solution-to-a-50-year-old-grand-challenge-in-biology (2020).

37. Gowers, R. *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *Proc. 15th Python Sci. Conf.* 98–105 (2016) doi:10.25080/majora-629e541a-00e.

38. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **32**, 2319–2327 (2011).

39. Theobald, D. L. Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. *Acta Crystallogr. Sect. A Found. Crystallogr.* **61**, 478–480 (2005).

40. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins Struct. Funct. Genet.* **57**, 702–710 (2004).

41. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–895 (2010).

42. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).

43. Kelley, L. A. & Sternberg, M. J. E. Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.* **4**, 363–373 (2009).

44. Piersimoni, L., Kastritis, P. L., Arlt, C. & Sinz, A. Cross-Linking Mass Spectrometry for Investigating Protein Conformations and Protein-Protein Interactions-A Method for All Seasons. *Chem. Rev.* **122**, 7500–7531 (2022).

45. Yu, C. & Huang, L. Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology. *Anal. Chem.* **90**, 144–165 (2018).

46. Liu, F., Rijkers, D. T. S., Post, H. & Heck, A. J. R. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nat. Methods* **12**, 1179–1184 (2015).

47. Liebeskind, B. J., Young, R. L., Halling, D. B., Aldrich, R. W. & Marcotte, E. M. Mapping Functional Protein Neighborhoods in the Mouse Brain. *bioRxiv* 2020.01.26.920447 (2020) doi:10.1101/2020.01.26.920447.

48. Chavez, J. D. *et al.* Chemical Crosslinking Mass Spectrometry Analysis of Protein Conformations and Supercomplexes in Heart Tissue. *Cell Syst.* **6**, 136-141.e5 (2018).

49. Schweppe, D. K. *et al.* Mitochondrial protein interactome elucidated by chemical cross-linking mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 1732–1737 (2017).

50. 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. *Anal. Chem.* **93**, 11442–11450 (2021).

51. Wang, Y., Hu, Y., Höti, N., Huang, L. & Zhang, H. Characterization of In Vivo Protein Complexes via Chemical Cross-Linking and Mass Spectrometry. *Anal. Chem.* **94**, 1537 (2022).

52. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331 (1999).

53. Joerger, A. C. & Fersht, A. R. The p53 Pathway: Origins, Inactivation in Cancer, and Emerging Therapeutic Approaches. *Annu. Rev. Biochem.* **85**, 375–404 (2016).

54. Arlt, C. *et al.* An Integrated Mass Spectrometry Based Approach to Probe the Structure of the Full-Length Wild-Type Tetrameric p53 Tumor Suppressor. *Angew. Chemie* **129**, 281–285 (2017).

55. Sarpe, V. *et al.* High sensitivity crosslink detection coupled with integrative structure modeling in the mass spec studio. *Mol. Cell. Proteomics* **15**, 3071–3080 (2016).

56. Hoopmann, M. R. *et al.* Kojak: Efficient analysis of chemically cross-linked protein complexes. *J. Proteome Res.* **14**, 2190–2198 (2015).

57. Yllmaz, Ş., Busch, F., Nagaraj, N. & Cox, J. Accurate and Automated High-Coverage Identification of Chemically Cross-Linked Peptides with MaxLynx. *Anal. Chem.* **94**, 1608–1617 (2022).

58. Fischer, L., Chen, Z. A. & Rappsilber, J. Quantitative cross-linking/mass spectrometry using isotope-labelled cross-linkers. *J. Proteomics* **88**, 120–128 (2013).

59. Mendes, M. L. *et al.* An integrated workflow for crosslinking mass spectrometry. *Mol. Syst. Biol.* **15**, e8994 (2019).

60. Tayri-Wilk, T. *et al.* Mass spectrometry reveals the chemistry of formaldehyde cross-linking in structured proteins. *Nat. Commun. 2020 111* **11**, 1–9 (2020).

61. Rasmussen, M. I., Refsgaard, J. C., Peng, L., Houen, G. & Højrup, P. CrossWork: Software-assisted identification of cross-linked peptides. *J. Proteomics* **74**, 1871–1883 (2011).

62. Lima, D. B. *et al.* Characterization of homodimer interfaces with cross-linking mass spectrometry and isotopically labeled proteins. *Nat. Protoc. 2018 133* **13**, 431–458 (2018).

63. Pirklbauer, G. J. *et al.* MS Annika: A New Cross-Linking Search Engine. *J. Proteome Res.* **20**, 2560–2569 (2021).

64. Giese, S. H., Fischer, L. & Rappsilber, J. A Study into the Collision-induced Dissociation (CID) Behavior of Cross-Linked Peptides. *Mol. Cell. Proteomics* **15**, 1094–1104 (2016).

65. Trnka, M. J., Baker, P. R., Robinson, P. J. J., Burlingame, A. L. & Chalkley, R. J. Matching cross-linked peptide spectra: only as good as the worse identification. *Mol. Cell. Proteomics* **13**, 420–434 (2014).

66. 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**, (2019).

67. Roepstorff, P. & Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.* **11**, 601–601 (1984).

68. Meng, F. *et al.* Informatics and multiplexing of intact protein identification in bacteria and the archaea. *Nat. Biotechnol.* **19**, 952–957 (2001).

69. Liu, F., Lössl, P., Scheltema, R., Viner, R. & Heck, A. J. R. Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. *Nat. Commun. 2017 81* **8**, 1–8 (2017).

70. Intro. to Signal Processing:Deconvolution. https://terpconnect.umd.edu/~toh/spectrum/Deconvolution.html.

71. Turner, J. N., Lasek, S. & Szarowski, D. H. Confocal Optical Microscopy. *Encycl. Mater. Sci. Technol.* 1504–1509 (2001) doi:10.1016/b0-08-043152-6/00271-0.

72. 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**, 2864–2889 (2018).

73. Graham, M., Combe, C., Kolbowski, L. & Rappsilber, J. xiView: A common platform for the downstream analysis of Crosslinking Mass Spectrometry data. *bioRxiv* 561829 (2019)

doi:10.1101/561829.

74. Riffle, M., Jaschob, D., Zelter, A. & Davis, T. N. ProXL (protein cross-linking database): A platform for analysis, visualization, and sharing of protein cross-linking mass spectrometry data. *J. Proteome Res.* **15**, 2863–2870 (2016).

75. Bullock, J. M. A., Thalassinos, K. & Topf, M. Jwalk and MNXL web server: model validation using restraints from crosslinking mass spectrometry. *Bioinformatics* **34**, 3584–3585 (2018).

76. Ferrari, A. J. R. *et al.* TopoLink: evaluation of structural models using chemical crosslinking distance constraints. *Bioinformatics* **35**, 3169–3170 (2019).

77. Moore, R. E., Young, M. K. & Lee, T. D. Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* **13**, 378–386 (2002).

78. 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**, (2020).

79. Matzinger, M. *et al.* Mimicked synthetic ribosomal protein complex for benchmarking crosslinking mass spectrometry workflows. *Nat. Commun.* **13**, (2022).

80. Gasteiger, E. *et al.* Protein Identification and Analysis Tools on the ExPASy Server. *Proteomics Protoc. Handb.* 571–607 (2005) doi:10.1385/1-59259-890-0:571.

81. Wilkins, M. R. *et al.* Detailed peptide characterization using PEPTIDEMASS--a World-Wide-Web-accessible tool. *Electrophoresis* **18**, 403–408 (1997).

82. Petritis, K. *et al.* Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. *Anal. Chem.* **78**, 5026–5039 (2006).

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

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

85. matplotlib.pyplot.hist — Matplotlib 3.7.0 documentation. https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.hist.html.

86. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).

87. Protein GRAVY. https://www.bioinformatics.org/sms2/protein_gravy.html.

88. Expasy - ProtParam tool. https://web.expasy.org/protparam/.

89. Hopp, T. P. & Woods, K. R. A computer program for predicting protein antigenic determinants. *Mol. Immunol.* **20**, 483–489 (1983).

90. Cornette, J. L. *et al.* Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* **195**, 659–685 (1987).

91. Erich Hiickel in Stuttgart, V. 204 Quantentheoretische Beitrgge zum Benzolproblem. I. Die Elekfronenkonfigurafion des Benzols und verwandfer Verbindungen~). (1931).

92. Vernon, R. M. C. *et al.* Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *Elife* **7**, (2018).

93. Allinger, N. L. Conformational Analysis. 130. MM2. A Hydrocarbon Force Field Utilizing V1and

V2Torsional Terms1,2. *J. Am. Chem. Soc.* **99**, 8127–8134 (1977).

94. MM2* Force Field. http://gohom.win/ManualHom/Schrodinger/Schrodinger_2015-2_docs/maestro/help_Maestro/macromodel/energy_mm2.html.

95. Mädler, S., Bich, C., Touboul, D. & Zenobi, R. Chemical cross-linking with NHS esters: A systematic study on amino acid reactivities. *J. Mass Spectrom.* **44**, 694–706 (2009).

96. Breusch, T. S. & Pagan, A. R. A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica* **47**, 1287 (1979).

97. Miscellanea: On Heteros*edasticity on JSTOR. https://www.jstor.org/stable/1911250.

98. GitHub - fstanek/imp-x-fdr: https://doi.org/10.1101/2021.10.21.465295. https://github.com/fstanek/imp-x-fdr.

99. Le Rhun, A., Escalera-Maurer, A., Bratovič, M. & Charpentier, E. CRISPR-Cas in Streptococcus pyogenes. *RNA Biol.* **16**, 380 (2019).

100. Ferretti, J. J. *et al.* Complete genome sequence of an M1 strain of streptococcus pyogenes. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 4658–4663 (2001).

101. UniProt. cas9 - CRISPR-associated endonuclease Cas9/Csn1 - Streptococcus pyogenes serotype M1 - cas9 gene & protein. *Cas9_Strp1* http://www.uniprot.org/uniprot/Q99ZW2 (2015).

102. Gutierrez, C. B. *et al.* Developing an acidic residue reactive and sulfoxide-containing MS-cleavable homobifunctional cross-linker for probing protein-protein interactions. *Anal. Chem.* **88**, 8315–8322 (2016).

103. 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).

104. Iacobucci, C. *et al.* First Community-Wide, Comparative Cross-Linking Mass Spectrometry Study. *Anal. Chem.* **91**, 6953–6961 (2019).

105. 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).