universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Imitated synthetic E. Coli ribosomal complex and XL-MS optimisation"

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 066 862
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Chemie
Betreut von / Supervisor:	Univ.-Prof. Dr. Christopher Gerner
Mitbetreut von / Co-Supervisor:	

Acknowledgements

First, I would like to thank Dr. Manuel Matzinger, who was always there along the way and always supported me. We were able to form a very successful team. I am very grateful for having you as a mentor. I am forever grateful that I had the chance to work with you and your work ethic, vigilance, attention to detail and professionalism are truly admirable. You are simply making science better.

Additionally, I am very grateful to Karl Mechtler, who allowed me to be part of his research group and gave me the freedom and support for the research projects.

I thank Univ.-Prof. Dr. Christopher Gerner for supervising my thesis and supporting me along the way.

Furthermore, I would like to thank Dr. Elisabeth Roitinger, Ines Steinmacher, Susanne Opravil, Dr. Karel Stejskal, Michael Schutzbier and Dr. Claudia Ctorteka. You are the backbone of this group and not only true professionals, but also very warm-hearted people.

Finally, I would like to express my gratitude to my parents, Andreas, Teodora and Daniel. You are giving meaning, joy and hope to my world every day.
I wish everyone happiness, love and health.

Abstract

Crosslinking mass spectrometry is an integrative part of proteomics research and is capable of providing insights into both protein networks and contact surfaces between proteins, as well as the three-dimensional structure of a protein complex and its dynamics.

This master's thesis begins with a detailed introduction to mass spectrometry in the field of proteomics, coupled with chemical crosslinking. The development history of mass spectrometric devices, their types and instrumentation is also explained. In addition, various crosslinker reagents are presented, their chemistry, and the technical aspects to consider when performing a crosslinking mass spectrometry workflow.

The state-of-the-art method of Beveridge et al. of assessing and optimizing crosslink workflows through a large and versatile synthetic peptide library was adopted and further developed. Not only were improvements made in the chemical design of the peptides, but the peptide library portfolio was also expanded. Three new peptide libraries were created, which were used to test the linkers DSSO, DSBU, CDI, DSBSO, ADH, and DHSO, and to refine the reaction conditions to maximize the number of cross-links between the covalently bound synthetic peptides. Probably the most important point, however, was a recalculation of the false discovery rate values and their minimization.

Various enrichment strategies were also evaluated with promising results. Even when the original sample was spiked with HEK peptides at a ratio of 1:100, more links could be found when the correct enrichment method was selected, compared to the purified peptide mix without any background proteome added.

Finally, a new highly promising XL-MS protocol based on adding fatty acids to the sample prior to the linking reaction was also developed to reduce the number of non-specific links and to further increase the total number of links obtained. The new protocol has already been tested on BSA and GroEL proteins.

Zusammenfassung

Crosslinking-Massenspektrometrie ist ein integrativer Teil der Proteomik-Forschung und ist in der Lage, Einblicke sowohl in die Protein-Netzwerke als auch in die Kontaktoberflächen zwischen Proteinen sowie in die dreidimensionale Struktur eines Proteinkomplexes und dessen Dynamik zu geben.

Diese Masterarbeit beginnt mit einer ausführlichen Einführung in die Massenspektrometrie im Bereich der Proteomik, gekoppelt mit chemischem Crosslinking. Es wird die Entwicklungsgeschichte der massenspektrometrischen Geräte, deren Arten und Instrumentalisierung erklärt. Darüber hinaus werden verschiedene Crosslinker-Reagenzien präsentiert, deren Chemie und die technischen Aspekte, die man betrachten muss, wenn ein Crosslinking-Massenspektrometrie-Workflow durchgeführt wird.

Die state-of-the-art Methode von Beveridge et al. Wurde übernommen und weiterentwickelt. Dabei wurden nicht nur Verbesserungen im chemischen Design der Peptide erzielt, sondern auch das Peptid-Bibliotheks-Portfolio erweitert. Es wurden drei neue Peptid-Bibliotheken erstellt, mit deren Hilfe die Linker DSSO, DSBU, CDI, DSBSO, ADH und DHSO getestet wurden und die Reaktionsbedingungen verfeinert wurden, um die Anzahl der Vernetzungen zwischen den kovalent gebundenen synthetischen Peptiden zu maximieren. Der wahrscheinlich wichtigste Punkt war jedoch eine Neuberechnung der False-Discovery-Rate-Werte und deren Minimierung.

Verschiedene Anreicherungsstrategien wurden ebenfalls evaluiert, die zu vielversprechenden Ergebnissen führten. Selbst bei einer Impurifizierung der ursprünglichen Probe mit HEK-Peptiden im Verhältnis 1:100 konnten mehrere Links gefunden werden, wenn die richtige Anreicherungsmethode ausgewählt wurde, als in der reinen, nicht impurifizierten Probe.

Schließlich wurde auch ein neues, vielversprechendes XL-MS-Verfahren entwickelt, das auf dem Hinzufügen von Fettsäuren zur Probe vor dem Linking-Reagenz basiert, um die Anzahl der unspezifischen Links zu reduzieren und die Anzahl der erreichten Links weiter zu steigern. Das neue Protokoll wurde bereits an BSA- und GroEL-Proteinen getestet.

Table of Contents

1.Theory

1.1. Mass Spectrometry

1.1.1. History and milestones

Like most of the instruments used nowadays by chemists, mass spectrometry's birthplace originates in the field of physics. And as with most discoveries that leave a lasting mark in science and reshape it, the "inventor" of the first mass spectrometer was in fact looking for something else. The 28-year-old J.J. Thomson, newly appointed with a Cavendish Professorship at the Cambridge University, had to choose a new program of experimental research. A hot topic at that time was "the transmission of electricity through gases" and Thomson joined the current. Back then, one of the research questions was to answer the nature of the cathode rays and, for those who believed that it was made of particles, there was a steep race to measure the mass of those particles. In 1897, Thomson, together with his assistant Everett, succeeded to design and construct an apparatus that was measuring e/m (different than m/e, which is being measured today) of those particles. Two years later, they managed to concomitantly determine e/m and e, thus unveiling the mass of an electron. His work was awarded with a Nobel Prize in Physics in 1906. Eventually, he (together with Aston) also constructed the first mass spectrometer in history, which was able to measure the mass of charged atoms.[1]

Fast-forwarding the history to the 1980s and, by doing so, skipping a lot of ground-breaking work and inventions, the researchers were able to routinely analyse small organic molecules. The field started to be embraced more and more by chemists. The new obstacle at that point was to measure large molecules, like proteins or nucleic acids. The challenge was that the ionization procedure was based back then on gas-phase collision between the analyte and the charged particles. It was extremely difficult to have large molecules in the gas phase without heavily fragment or decompose them. Some techniques, like fast atom bombardment or plasma desorption, managed to produce some results, but they would require high amounts of sample and still fail to measure larger proteins. MALDI and ESI made then their appearance in 1988 and solved the problem. These techniques were developed almost simultaneously, and they prevail the field to this day. Electron Spray Ionization (ESI) was developed by the chemist John B. Fenn, while the existence of MALDI is largely due to Hillenkamp's and Karas' contributions. Now, let us teleport to the Nobel Prize Gala for chemistry in 2002, where the prize was shared between Kurt Wüthrich for "his development of nuclear magnetic resonance spectroscopy for determining the three-dimensional structure of biological macromolecules in solution"; and John B. Fenn and Koichi Tanaka for "the development of methods for identification and structure analysis of biological macromolecules".[2] It must be mentioned that Tanaka's technique was different than MALDI in some regards. Even though it also used a laser to reach ionization of the analyte, a suspension of nanoparticles made from metal in glycerol is employed to absorb the energy from the laser and carry a fraction of it to the analyte. Also, the sample to analyse is not embedded into the suspension as it is the case for MALDI but layered on the surface. The MS community embraced the MALDI method thanks to its better sensitivity. As a final note in this very brief history of mass spectrometry, the community should recognize and appreciate the development of MALDI, even though it did not receive the high recognition that it deserved.[1]

1.1.2. Principles and instrumentation

The analytes must be in the gas phase and ionized during the mass spectrometric analysis. Conceptually, a mass spectrometer is composed of an ion source, a mass analyser which separates ions based on charge to mass ratios (m/z), and a detector that counts the number of ions at every m/z value.

For the analysis of peptides and proteins, the most suitable techniques are Electrospray ionization (ESI) and Matrix Assisted Laser Desorption/Ionization (MALDI).[3]

The process of an analyte being transferred from solution into the gas phase through ESI involves three main steps. The first step is the formation of charged droplets from a high voltage capillary tip. When the analyte solution is carried through the emitter, a redox reaction of the solvent takes place which produces an electron flow to or from the metal capillary (based on positive or negative polarity). Water, ethanol, or acetonitrile are very suitable polar solvents because they can simply go through electrochemical reactions as they pass through the spraying nozzle. The created charges are repelled by the capillary of the same polarity and drawn to the liquid surface at the capillary exit. Because the charges at the surface are destabilized, the meniscus turns into a cone (Taylor cone) under the influence of the high electric field. The Taylor cone expels a jet of liquid in the direction of the other electrode and the charged jet decomposes into small droplets. The second step is Coulomb fission and disintegration of the charged droplets. The equidistant positioning of the charges at the surface of a droplet is due to the minimisation of the potential energy. Inside of a charge droplet, there are two forces which act in opposite directions: the surface tension responsible for the spherical form and Coulomb repulsion between the like charges at the surface. The evaporation of the solvent takes place on the way from the spraying nozzle to the “heated” capillary. Therefore, the size of the droplet reduces until it reaches the Rayleigh limit, and the Coulomb fission occurs. The third and last step is the formation of gas-phase analyte ion from the very small highly charged droplet. In order to explain this process, two main mechanisms were proposed: the residue model and the ion evaporation model, but up until now none of them was unambiguously proven right.[4]

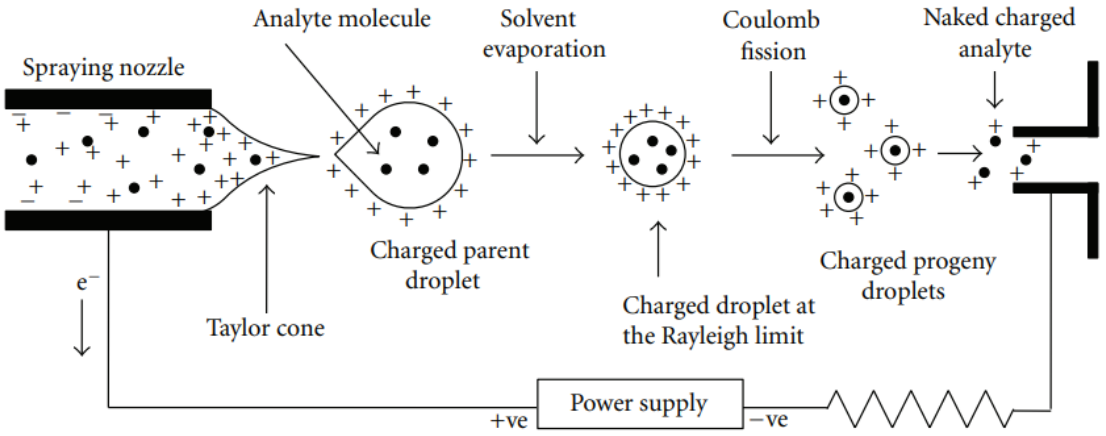


Figure 1: Schematic characterization of electrospray ionization technique. Adapted from Shibdas Banerjee and Shyamalava Mazumdar[5]

1.2. MALDI

Matrix-assisted laser desorption/ionization (MALDI) is the second soft ionization technique able to ionize non-volatile, high-molecular-weight analytes prior to mass spectrometry analysis. The method is straightforward in operation, has rapid analysis times and shows not only a high mass accuracy, but also a high mass resolution. The method is routinely employed in fast determination of proteins and peptides, but has various applications in measuring synthetic polymers, glycans and other macromolecules as well. During MALDI, a laser beam hits a spot on a specially designed stainless-steel plate. On the plate, there are usually uniformly distributed, circular spots where the sample mix is applied. The sample mix is composed of the analyte to be studied and a normally large excess of matrix material.[6] There are two postulated analyte protonation pathways in MALDI: the Lucky Survival model and the gas phase protonation model, which are extensively discussed in other studies.[7] Nowadays MALDI is usually coupled with Time-of-Flight (TOF) analysers for analysis of the generated ions, where the time needed from the place where ions were generated to the place where the detector is situated is used to calculate the mass-to-charge ratio. Nitrogen lasers at 337 nm and Nd:YAG lasers at 355 nm are the most common lasers typically employed. The special resolution of the mass spectrometer is dictated by the diameter of the laser beam on the sample surface, and it measures normally 20 μm^3 on the commercial machines but can be adapted to reach down to 5 μm^3. When talking about matrices, it all started when tryptophan was used as a matrix for the non-absorbing alanine with a laser beam of 266 nm. This way, alanine could be ionised because tryptophan absorbs at that wavelength (although the absorption maximum is not at 266 nm). Nowadays the most common matrices used for MALDI are y, 2,5-dihydroxybenzoic acid (DHB), α-cyano-4-hydroxycinnamic acid (CHCA), and sinapic acid (SA). It must be understood that the choice of the matrix is dependent on the analysed sample. If proteins are studied, the recommended matrix is either SA or DHB. For peptides, one should use CHCA or DHB, while for carbohydrates only DHB. If the composition of the sample is unknown, then DHB is the best option because it fits a broad chemical space. There are also two major disadvantages when using the common matrices. The first is the weak reproducibility due to inhomogeneities in the sample and can sometimes be avoided by using a combination of more matrix-materials. The second inconvenience is

that the matrix usually produces noise.[6] For example, DHB presents peaks at m/z 137.0 [DHB - H2O + H]$^+$, 154.0 [DHB·]$^{+\bullet}$, 155.0 [DHB + H]$^+$, 177.0 [DHB + Na]$^+$, 199.0 [DHB - H + 2Na]$^+$, and 273.1 [2DHB - H2O - OH]$^+$.[8] These interferences are not necessarily a problem for peptide and protein analysis and the m/z ranges do not overlap. A general representation of MALDI is shown below.

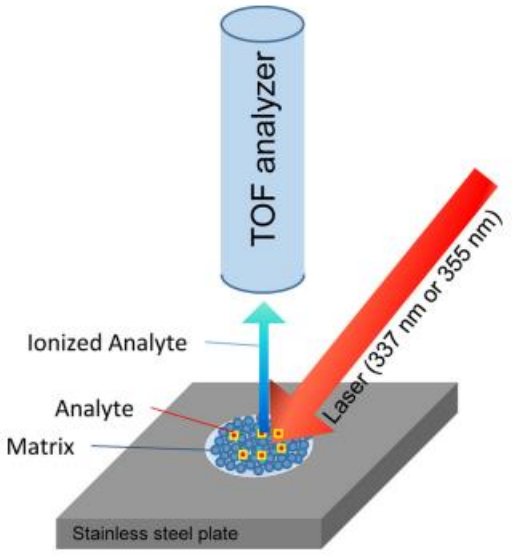


Figure 2: Schematic representation of the MALDI technique. Adapted from Kim, Jeongkwon. (2015). Sample Preparation for Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry. Mass Spectrometry Letters. 6. 27-30. 10.5478/MSL.2015.6.2.27.[6]

1.3.Q-Exactive Orbitrap

With the massive improvement brought by Alexander Makarov in 2000 by developing a new type of mass analyser (Orbitrap) which utilises trapping in an electrostatic field, the proteomics field evolved enormously. This allowed the construction and commercialization of the Q Exactive mass spectrometers by Thermo Fisher Scientific. Because these machines are so widely spread and to this date almost indispensable, they deserve a detailed description.

The so-called Q Exactive is a hybrid instrument equipped with a HESI ionization source, a series of S lenses, injection flatapole, actively guiding bent flatapole, quadrupole, HCD collision cell, C trap and an Orbitrap in the back as principal components.[9]

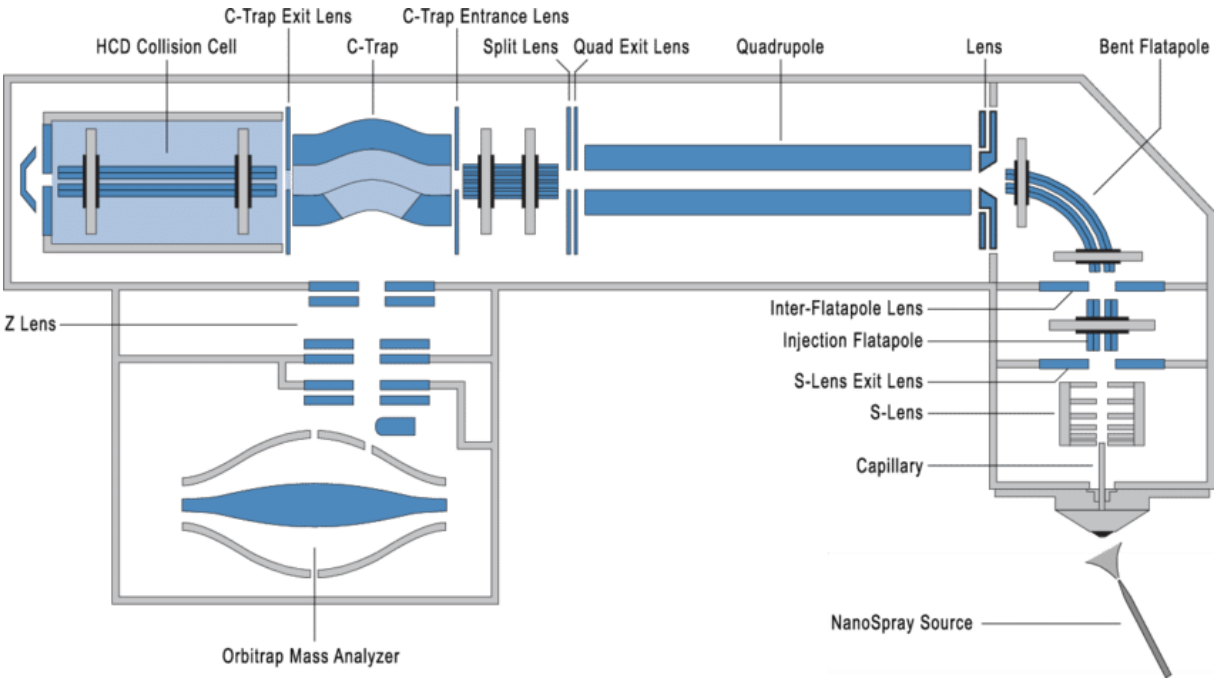


Figure 3: Construction details of the Q Exactive. This instrument is based on the Exactive platform but incorporates an S-lens, a mass selective quadrupole, and an HCD collision cell directly interfaced to the C-trap. Note that the drawing is not to scale. Adapted from "Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer" [10]

The Orbitrap is a new type of mass analyser that utilises trapping in an electrostatic field. Ion stability is reached in this case only because ions are orbiting around the axially placed electrode. While orbiting, the ions present a harmonic oscillation behaviour, and their frequency is proportional to the $\sqrt{m/z}$. Using fast Fourier Transformation, the oscillations are converted into mass spectra.[9]

During full scan mode of Q Exactive, ions are being generated at the ESI source, captured and focused through the S lenses in order to achieve a high ion transmission. The uncharged neutral molecules are colliding in the bent flatapole, and therefore filtered out. After the ions have been cleaned up, they traverse the quadrupole (without further filtering in the full-scan mode), gathered into packs in the C-trap, stabilized and directly sent to the orbitrap analyser to be detected. As mentioned before, the orbital motion of the ions has a frequency of rotation, which is dependent on their m/z. Therefore, ions with different m/z ratio will present distinct frequencies of oscillation. In the case of bottom-up proteomics, including crosslinking experiments, the preferred mode is MS/MS (also called tandem mass spectrometry). Here, ions are once again produced at the ESI source, detained and focused through the lenses. After filtering out the uncharged neutral molecules in the bent flatapole, only ions of interest with desired molecular masses are passed through the quadrupole. The rest is being filtered out. The ions of interest are then entering the collision cell. In the HCD cell, collision induced fragmentation occurs and the fragments are formed. They are then accumulated into packets and stabilized in the C-trap and finally sent to the orbitrap for detection.[10,11,12]

1.4. Crosslinking mass spectrometry as separate study field

Crosslinking mass spectrometry has emerged as an extremely potent way to study protein-protein-interactions (PPI), the tertiary structure of a protein, protein complex topologies and protein dynamics.[13,14]

Due to its high applicability in many key domains of the proteomics study, the researcher must firstly define the research question and determine the size of the system. In a low-complexity system with highly purified proteins, the topologies of protein complexes can be determined with a low resolution. Insights to the tertiary structure or the multitude of conformational states a protein can adopt are also achievable. In a highly complex sample, for example organelles, living cells or tissue, the complex interaction networks of proteins and, up to a certain level, the interaction sites of the partners can be elucidated.[13–16]

The system wide studies have become increasingly accessible as cleavable crosslinks emerged. As such, more highly sensitive mass spectrometers make their way to the market and many variants of crosslink search engines appeared.

The principle is simple: two amino acid residues in close spatial proximity are bound together covalently by a linker and therefore freezing this spatial propinquity. Most of the linkers target preponderantly lysine and show some limited activity towards serine, threonine and tyrosine. Others target amino acids with a carboxyl group within their side chain (aspartic acid and glutamic acid), but show lower reaction yields. This aspect already can constitute a problem as lysine is not homogenously distributed through the protein sequence and two interacting protein sites do not necessary contain lysines in their sequence. As a solution to this reduced amount of data, non-specific crosslinkers were created.

Another problem arising from using lysine as crosslinking site is a high miscleavage rate of the proteins in the digestion step. If the lysine residue is attached to the linker, trypsin is not able to recognise it anymore. The obstacle can be overcome by using a combination of more proteases cleaving at different sites as described in the enzyme subchapter. An incomplete digestion and the formation of the crosslinker itself (two peptides linked together with or without a spacer) lead to bulky molecules. They are generally harder to ionize and to fragment. In the sense of comparability with other workflows and because trypsin performs generally well, many researchers still prefer the trypsin-digestion protocol.

The proportion between crosslinked peptides and linear peptides (not crosslinked peptides) is very low. Especially in the case of in-vivo studies, linkers tend to hydrolyse before they passed the cell membrane and/or organelle walls and find two reactive residues. Many linkers are designed with 2 N-hydroxysuccinimide (NHS) esters and are unstable in an aqueous medium.[14]

One of the biggest challenges with regard to the XL-MS breakthroughs from the last years is its heuristic approach. Due to this, the random hits that appear in a very large search-space of the database may lead to questionable results.[14]

The employment of cleavable crosslinkers aims to further maximise the confidence and minimize the running time of the search algorithms from $O(n^2)$ to $O(2n)$. Of note O() is commonly used notation in computer science and refers to the time complexity of an algorithm[17]. In such a linker, there is a functional group that is being cleaved by either collision induced dissociation (CID), higher-energy collisional dissociation (HCD) or electron-transfer dissociation (ETD) method. The dissociation of the labile functionality takes place at lower or approximately equal potentials to the ones necessary for fragmenting the peptide bonds. Sulfoxides[18] with a lower dissociation potential and urea groups[19] with comparable potentials have shown very good results.[14]

Also worthy of mentioning, an interlink between two peptide sequences originating from the same protein can be perceived from the crosslink-search engines as an intralink and vice versa. The software has virtually no means of differentiating between these two cases. This aspect was also tackled by creating a special crosslinking protocol, but the results were rather surprising.

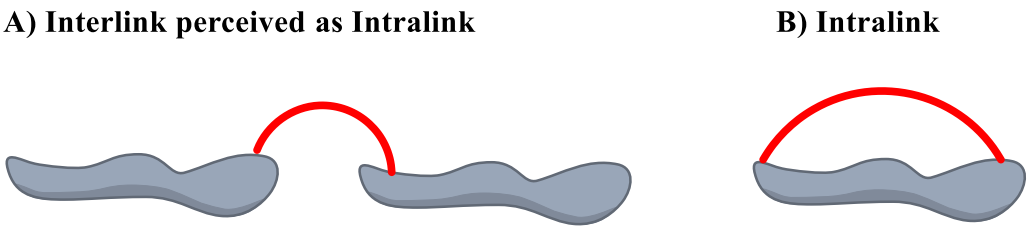


Figure 4: Interpretation of interlinks originating from the same protein as an intralink

1.5. Classification of Crosslinkers

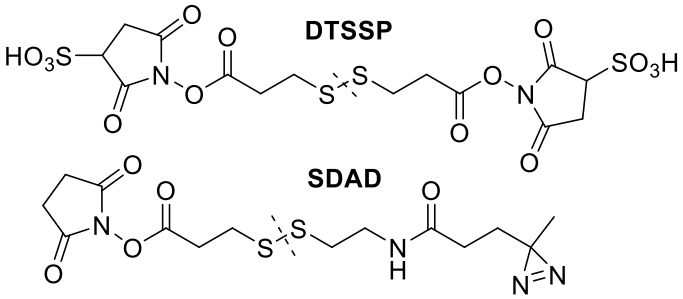


Figure 5: 2 examples of thiol-cleavable crosslinkers. DTSSP is homobifunctional and amine-reactive, while SDAD is a heterobifunctional linker and is at one end amine reactive and photo-reactive at the other end (eliminating the necessity of another amino-group in the proximity of the one-sided reacted linker).

Crosslinkers cleavable by reduction have been used to map protein interfaces. They have a disulfide bridge incorporated in their spacer. When this type of crosslinker is utilized, the linked pair is covalently bound, separated from the rest of the sample by a non-reducing SDS-PAGE. Additionally, the linked partners are enzymatically digested. The sample is then analysed in a mass spectrometer in non-reduced and reduced form. The resulted obtained through peptide mapping analysis before and after the reduction are then compared. The signal only present in the non-reduced form represents the alleged linked peptide pair. In the reduced form one or both peptides with their chain containing a thiol group.[20]

This workflow is simplistic but has the advantage that it can be used also with low-resolution mass spectrometers. It proved to be successful in 2000, where the researchers used 3,3'-dithio-bis (sulfosuccinimidyl propionate) to reveal a "head to tail" conformation of the homodimeric DNA binding protein ParR and an interaction between CD28-IgG and CD80-Fab.[12]

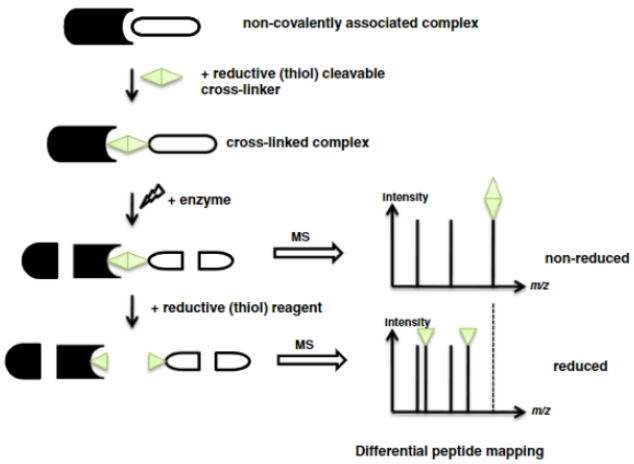


Figure 6: Schematic representation of the crosslinking workflow using thiol-cleavable linkers. Adapted from [12]

SDAD (Succinimidyl 2-([4,4'-azipentanamido]ethyl)-1,3'-dithiopropionate) contains a traditional NHS ester moiety, a diazirine moiety and a disulfide group and is a heterobifunctional cleavable crosslinker. In this particular linker case, a bait protein can be activated by reaction with the NHS ester before other potential interacting proteins are added to the sample. UV-irradiation favours the formation of a highly reactive carbene on the diazirine site, which immediately reacts with any interaction partner in spatial proximity. Additionally, the sample is digested and reduced for the MS-peptide mapping as in the workflow explained above. This idea circumvents the problem of interactome coverage and the necessity of having two lysines at both interacting sites. It paved the way for the first study of interactions between Fdh-N and Fdh-O (two homologous respiratory formate dehydrogenases) in the periplasmic space of anaerobically grown Escherichia Coli. A schematic representation of the workflow (excluding the digestion step before reduction for simplicity) is shown below.[21]

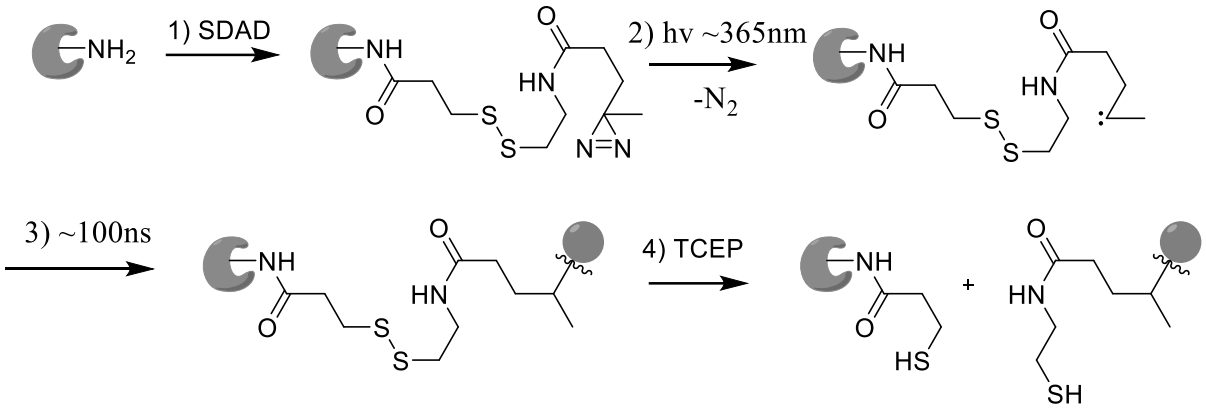


Figure 7: Schematic representation of heterobifunctional amine-/photoreactive linker SDAD[21,22]*. Drawn in Chemdraw*

A very promising crosslinker class is represented by the MS-cleavable linkers that create specific fragment ion signatures. In comparison to the linkers cleavable by reduction, these have the potential of fully automated workflows due to the cleaving step occurring during the MS/MS experiments. Even though there are some examples of crosslinks cleavable under infrared multiphoton dissociation (IRMPD)[23] or electron transfer dissociation (ETD) conditions[24], the vast majority of linkers can be cleaved under CID conditions and represent the accepted norm out of practical reasons.

The Edman linker was created by exploiting the knowledge of easily cleavable peptide bond between asparagine and proline, and the already successful crosslinkers with an Asp-Pro incorporated in the spacer. It has thiourea group and Gly-Pro peptide bond. The strongly nucleophilic sulfur atom can initiate a nucleophilic attack on the neighbouring amide and produce the cleavage. The Edman linker

cleaves with high yields at both low energy conditions (5-100 eV) and high ones (thousands eV) used in CID methods. It possesses one labile bond but produces the characteristics doublets due to its asymmetry.[25,26]

Disuccinimidyl dibutyric urea (BuUrBu or more commonly called DSBU) is the result of intense research effort and it is a viable crosslinker for performing proteome wide experiments. DSBU is homobifunctional and symmetric with two cleavable bonds determined by the urea group at the centre of the spacer. As a result of its design, it exhibits specific fragment ion patterns created by mass increments of 85 Da (Bu fragment) and 111 Da (BuUr fragment) at the linked peptides during the CID-MS2 experiments. The cleavage takes place at one of the two -NH-C=O bonds and two doublets with a difference of 26 Da each will be seen in the spectrum (ideally). Additionally, one can discriminate between looplinks, regular crosslinks and dead-end links. A visual representation of the fragmentation process is shown below. It is visible that none of the four signals have the same intensity. This is due to the fact that the two -NH-C=O do not necessarily fragment in the same percentage and the ionization efficiency can also differ. The linker was also used with success to study the tetrameric p53, the so-called "Guardian of the genome".[27,28] Its spacer arm has a length of 12.5Å and it leads to a higher result density with a lower resolution compared to DSSO or CDI.

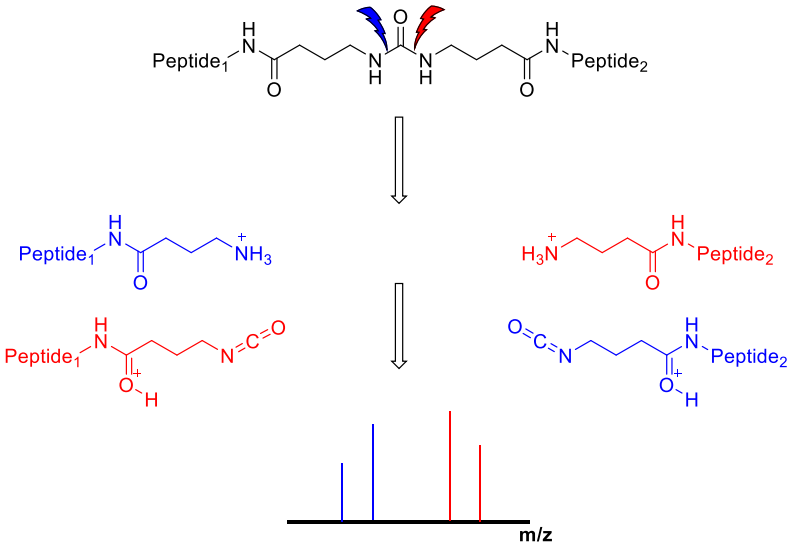


Figure 8: Representation of fragmentation pathway of DSBU-crosslinked peptides. Drawn in ChemDraw. Adapted from [20]

DSSO, developed in 2010 and at this point commercially available, has two symmetric CID-cleavable sites, which make a highly efficient identification of the crosslinked peptides based on their characteristic fragmentation patterns possible. Just as with DSBU, disuccinimidyl sulfoxide permits a discrimination between crosslink-type 0, 1 and 2 based on their signals.[29]

DSSO can be obtained through a two-step synthesis. Firstly, 3,3'-thiodipropionic acid is mixed with N-hydroxysuccinimide (1:2 molar equivalents) in a dioxan solution and stirred under Ar atmosphere. Then 2 molar equivalents of N,N'-dicyclohexylcarbodiimide in dioxan are added dropwise and stirred for another 12h. The obtained sulfide precursor is then purified, dissolved in chloroform and oxidized with a molar equivalent of m-chloropebenzoic acid. The synthesis procedure has a yield of 64%. For the detailed synthesis procedure, consult the paper.[29]

In the Figure 8.B, the proposed fragmentation scheme of a regular type 2 interlink is shown. One of the two C-S bonds adjacent to the sulfoxide group divides the linked peptide pair into a pair of peptide fragments. In the case α_A/β_S, the α-peptide has an alkene moiety (+54 Da) and the β-peptide presents

a sulfenic acid (+104 Da) group. In the most common case, where peptide α and peptide β are different, both α_A/β_S and α_S/β_A are expected to be observed and therefore four individual peaks appear in the MS/MS spectrum. If peptide α and peptide β have an identical amino acid sequence, only one pair of peaks will be seen in the analysed spectrum.[20,29]

As shown in Figure 8.C, during the analysis of a type 0 link, where one end of DSSO is hydrolysed and one attached to a peptide, only two fragment ions (α_A and α_S) are identifiable. Regarding the precursor ion, it has an increment of 176 Da to the peptide mass. The intralink (type 1 crosslink) bears a defined mass modification of 158 Da due to DSSO attaching to two different amino acid residues in the same peptide string (shown in Figure 8.D). During CID fragmentation, only one peak will be produced with the identical mass as the precursor. Figure 8.E illustrates the sulfenic acid-modified fragment undergoing a second fragmentation step through water loss. This produces an unsaturated thiol moiety (+86 Da). Because this is often the case, it must be individually considered in the crosslink search engines, as it produces another mass difference between the peaks. The mathematical relationships between different fragment ions are stated in Figure 8.F.[20,29]

Structurally distinct than previously discussed linkers, CDI, marketed as the first zero-length mass-spectrometry cleavable crosslinkers, is providing geometrical constraints in order to reach higher resolution computational modelling of the analysed protein systems. This is due to its very short spacer length of 2.6 Å. Every crosslinker has a defined length and can be seen as molecular ruler. 1,1'-carbonyldiimidazole imposes very strict distance constraints between the covalently linked residues. In their search for reactive, very-short linkers, the authors screened 4 different candidates: disuccinimidyl carbonate (DSC),1,1'-carbonyldiimidazole (CDI),1,1'-carbonyl-di-1,2,4-triazole(CD-1,2,4-T), and 1,1'-carbonyldipyrazole(CDP). Their crosslinking-fitness was analysed with three proteins and CDI proved to be the most effective reagent. The other candidates either show a lower reactivity towards the targeted residues or a faster hydrolysis kinetics rate. Chemically speaking, it is an electrophilic reactive amide. The nitrogen atom involved in the amid bond is also part of an aromatic five-membered heterocycle. For the Huckel's rule to be fulfilled ($2n+e^-$ per cyclic system), the nitrogen atom contributes with its lone pair to the aromatic system. This results in a weakened partial double-bond effect on the amid bond and thus a lowered electronic density on the nitrogen atom.[30–33]

A unique feature that CDI has is its capacity of 'recycling' partially hydrolysed crosslinks. Thus, no dead-end links are produced, because the carbamic acid decarboxylates restoring the initial structure of the targeted residue. This can be used as a remarkable advantage, because the inexistence of the type 0 links also eliminates one of the most common causes of misassignments. Other than that, CDI is also commercially available at a very low cost, making it accessible to more laboratories.[33,34]

DHSO (together with DMTMM) appears as a complementary tool in the class of the MS/MS cleavable crosslinkers and helps the expansion of the covered protein interaction-regions in the field. Having a sulfoxide moiety in the centre, dihydrazide sulfoxid expresses the same specific fragmentation patterns as DSSO. DHSO exclusively targets carboxyl-containing side chains (aspartic and glutamic acid residues).[20,35]

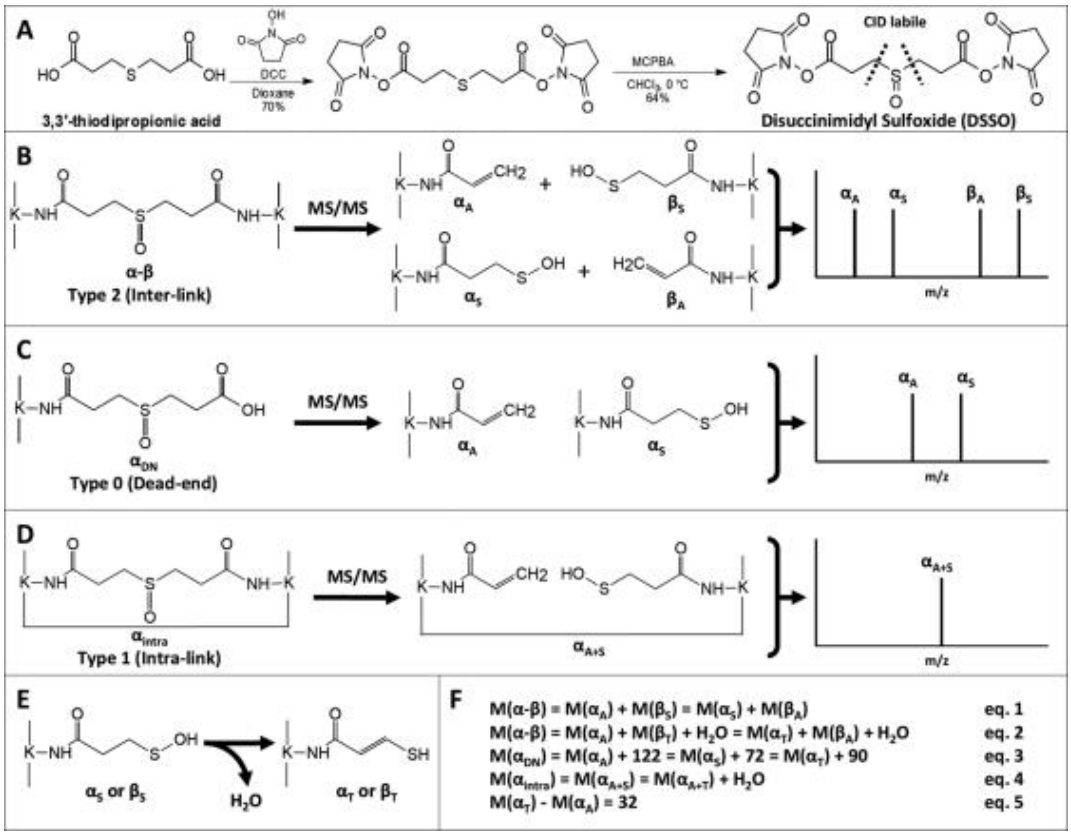


Figure 9: Schematic Representation of DSSO-crosslinks fragmentation pathways. A) Synthesis of the crosslinker. B) Signal assignment of type 2 crosslink fragments. C) Signal assignment of type 0 crosslink fragments. D) Signal assignment of type 1 crosslink fragments. E) Conversion from a sulfenic acid modified fragment to an unsaturated thiol-modified fragment through water loss.[20] F) Equations for the masses of the produced fragments. Reprint from [29]

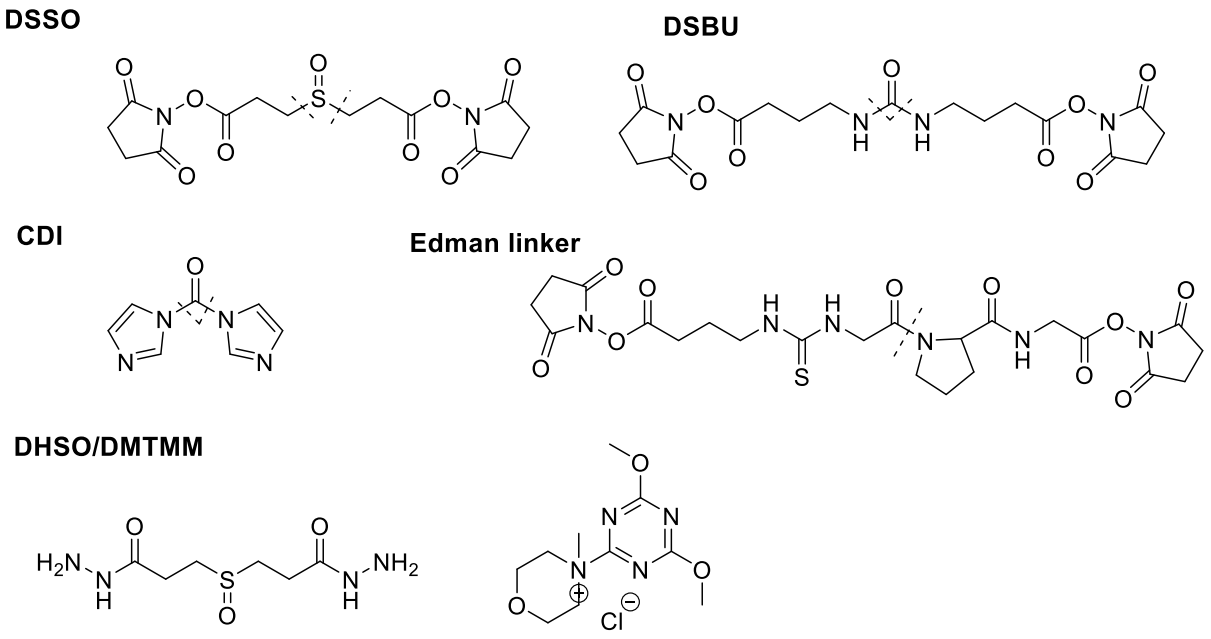


Figure 10: Examples of CID-cleavable (homobifunctional and heterobifunctional) crosslinkers. The cleavage sites are noted with a dashed line. DSSO, DSBU, CDI, Edman linker are amine-reactive NHS esters and DHSO (in combination with DMTMM) is an acid-reactive linker.

An interesting branch in crosslinks design is represented by the trifunctional linkers. Two of the functionalities are binding to the residues covalently while the third one is used for affinity enrichment of linked products. Since their development in 2005 with the apparition of PIR (Protein Interaction Reporter), a lot of bioinformatic strategies and in vivo applications were made possible. Even though the spacer arm of PIR is 43 Å, it was observed that residues in much closer proximity can be bound due to the flexibility of the linker. PIR bears a biotin, which can be enriched with the regular biotin-streptavidin strategy, but it is relatively bulky.[36–39] To avoid the steric problems that can arise from the bulkiness of biotin, one can use DSBSO. In DSBSO, an azide group is incorporated in the linker and can be selectively enriched by applying a copper free click chemistry reaction.[40]

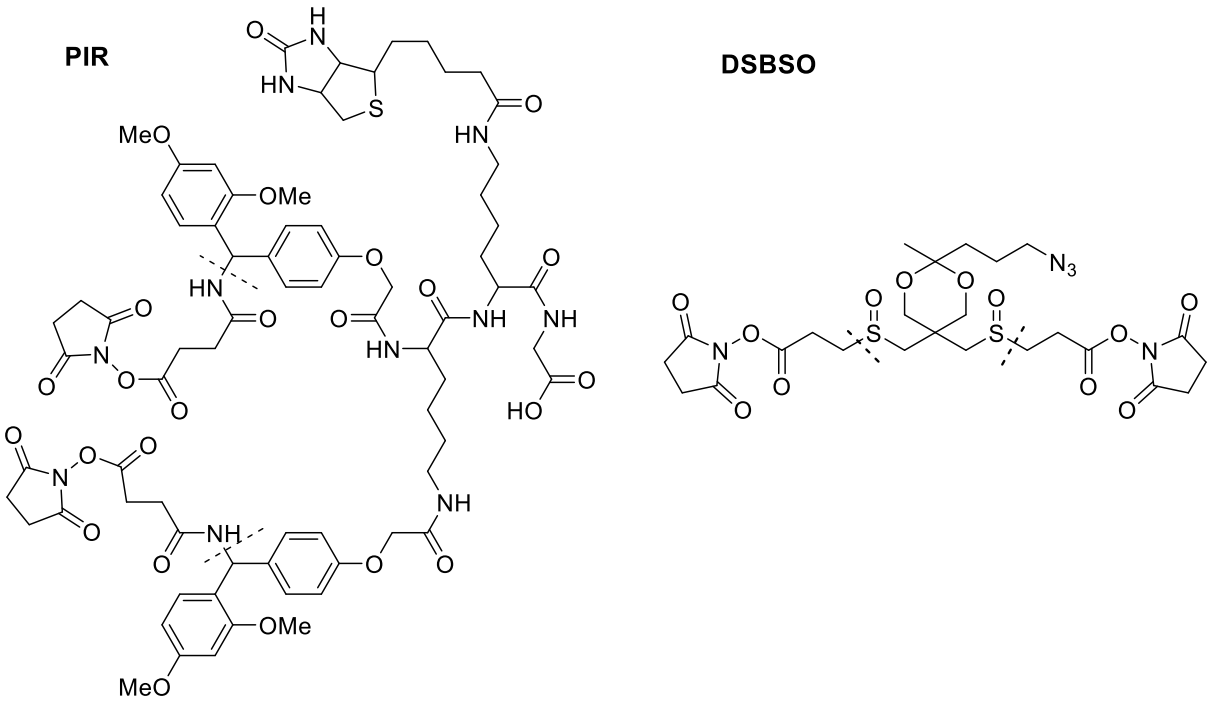


Figure 11: Examples of cleavable heterotrifunctional crosslinkers. The cleavable sites are noted with a dashed line.

Before describing the non-cleavable crosslinker class and their particularities, it must be stated that they still are predominantly applied in XL-MS studies, detrimental to the cleavable linkers. In a study published in 2020, it was shown that the use of uncleavable reagents exceeds the use of cleavable crosslinking reagents by far, with ~77% of the crosslinking reagents used being not cleavable.

In the diagram below one can observe the evolution of the number of research articles published through the years, where XL-MS methods were applied, and which type of linker was employed for the study. This fact can have several possible explanations: non-cleavable linkers appeared first and are better known to the researchers; most studies conducted concentrate on examining interactions in low or middle complex samples; or continuous improvement and adaptation to the newest mass spectrometers of the search engines specialized in non-cleavable reagents like plink/plink2. On the other hand, cleavable reagents are by far the preferred option when performing studies on highly complex samples.

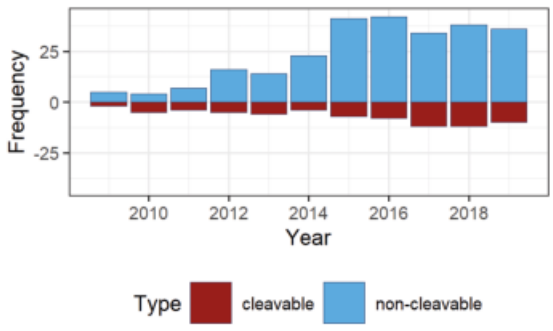


Figure 12: Tendency of employment of the cleavable and noncleanable crosslinkers in the scientific papers over the years. Reprint from [41]

DSS and BS3 were concomitantly developed in 1982 and are being intensively used in the proteomics field[42]. The only difference between the two linker is that DSS presents the usual, NHS ester moiety

and BS3 an sulfo-NHS ester moiety in order to further enhance the solubility in aqueous solutions. Their spacer arm has a length of 11.4 Å. These reagents combined were utilized in approximately 59% of all XL-MS studies published in the last decade (either alone or together with other linkers) making them widely accepted in structural biology studies.[41]

The first efficient crosslinker developed targeting the acidic amino acids was in 2013 with the published work ADH/DMTMM and PDH/DMTMM. Previously, there have been attempts to crosslink acidic amino acids with dihydrazides in combination with EDC (1-ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride), but this requires a mildly acidic medium (pH 5.5) which can have a denaturating effect on the protein sample. The adipic acid dihydrazide can be applied to the protein sample at a neutral pH (7-7.5), has a spacer length of 11.1 Å and four positions in its arm which can be exchanged with deuterium atoms for the heavy form.[43]

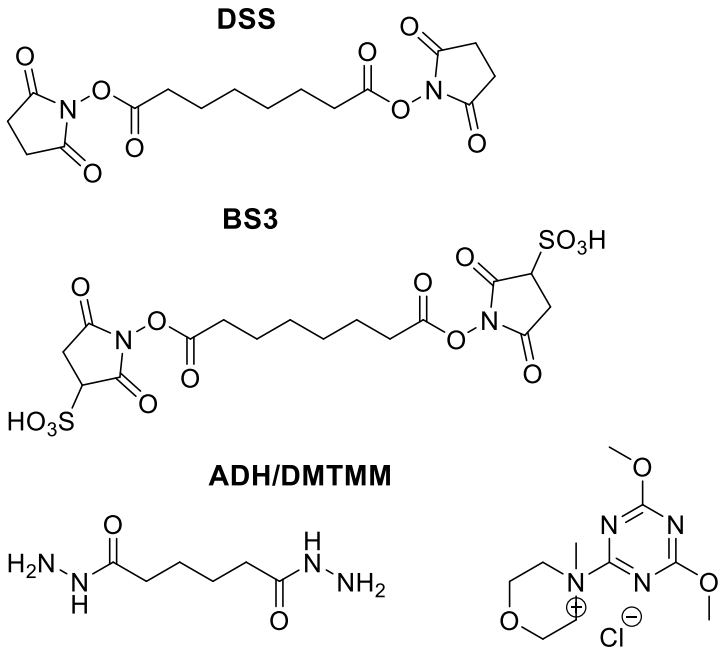


Figure 13: Examples of noncleavable crosslinkers employed in XL-MS. DSS and BS3 are homobifunctional amine-reactive linkers and most often utilized in linking experiments. ADH (in combination with DMTMM) is a homobifunctional acid-reactive linker.

1.6. Enrichment strategies

The latest advances in field push the boundaries of *in vivo* crosslinking. There are many factors hampering an in-depth sight of the protein-protein interaction of a living cell. The first one is cell permeability. The crosslinker must possess the capacity of trespassing the lipid membrane of the living cell without disrupting it. A supplementary restriction is that the studied space is almost solely restricted to the most abundant proteins. In order to get a glimpse of the interactions of the very low concentrated proteins, highly effective enrichment strategies should be employed. Nevertheless, one also must regard the problematic of stoichiometry in crosslinking reactions. Until the linker has managed to reach the proximity of a reactive amino acid residue (after it permeated the lipid membrane), it is most likely partially hydrolysed or completely hydrolysed. Due to all these above-mentioned reasons, it is of vital importance to make use of the enrichment, especially when it comes to in vivo studies.

Affinity Enrichment of Azide-A-DSBSO Cross-Linked Peptides represents a state-of-the-art method and has shown highly promising results.

Before going into a detailed description of the finetuned method, it has to be mentioned (and also credited) that the original idea was using biarylazacyclooctynone (BARAC) conjungated with biotin for the click reaction and the enrichment was done with streptavidin bounded to the beads.[44]

In the first step, DSBSO is dissolved in dry DMSO to the required concentration, added to the sample and incubated at room temperature for 1h. Quenching is performed with TRIS. After this regular linking reaction, the buffer is exchanged to 50 mM HEPES with the help of a Zeba Spin Column. Not only that, but in this step excess linker, N-hydroxysuccinimid side products Mg^{2+} and other salts from the sample are being removed. Additionally, the proteins in the sample are denaturated with sodium deoxycholate, followed by reduction of disulfide bridges with DTT, alkylation of newly formed thiols with IAA and accompanied by enzymatic digestion (benzonase, LysC, tripsin).

In the second main step, the DBCO coupled beads are equilibrated and added to the sample. The DBCO groups and azide groups should be in an excess ratio of 10:1. The click reaction between azide and alkyne is copper free and leads to the formation of 1,2,3-triazole, which is stable. In the literature, it is known as azide alkyne Huisgen cycloaddition and it undergoes through 1,3-dipolar cycloaddition reaction.[45] The reaction takes place at room temperature with gentle agitation for minimum one hour. Alternatively, it can also be incubated at 4°C overnight. After the beads were washed according to the protocol, the acetal group was cleaved in a TFA 2% (v:v) solution for 1h. These are milder conditions in comparison to the original protocol, which used an overnight incubation with 20% formic acid and 20% acetonitrile. After the peptide pairs were hydrolysed at the linker site, the supernatant is transported in new tubes with DMSO in it (a final concentration of 5% v:v). It was observed that the presence of DMSO prior transport of the crosslink solution is beneficial because peptides tend to stick to the Eppendorf tube walls. Another interesting finding was that in-house produced DBCO coupled beads by coupling DBCO with NHS preactivated Sepharose were more compatible to MS as the commercially available DBCO beads, as they did not produce any noticeable background interfering signals.[46]

Overall, the omission of biotin in the working protocol brings a supplementary advantage: the intrinsically biotinylated proteins are not coenriched, and thus not interfering with the method.

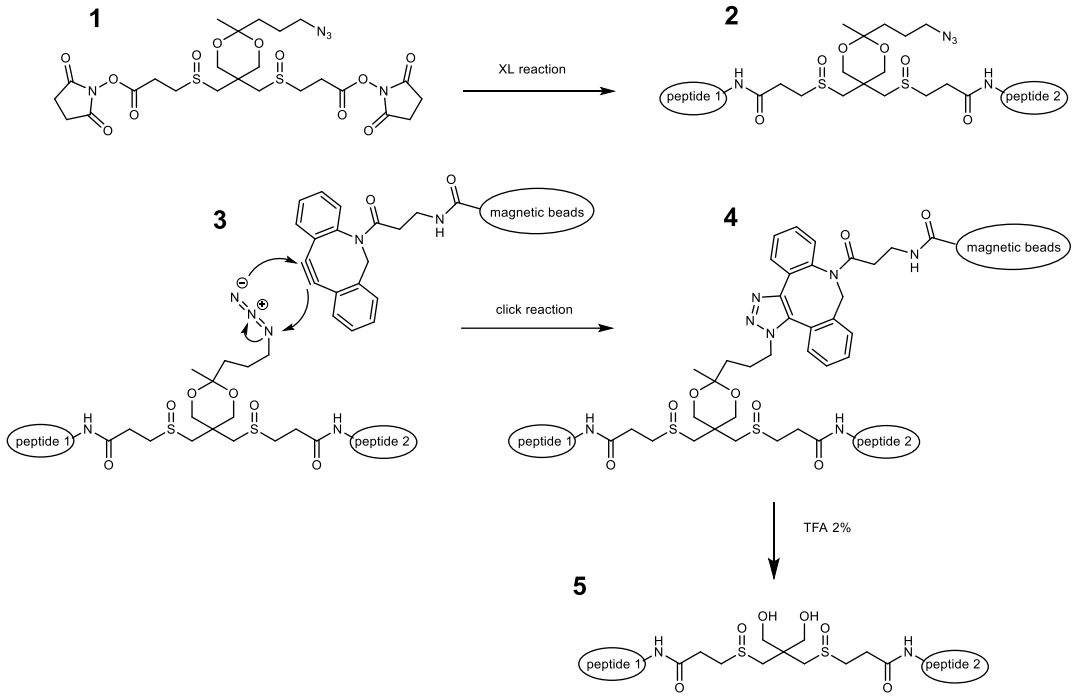


Figure 14: General workflow of crosslinking experiments with DSBSO. The enrichment of the crosslinks happens with DBCO coupled Sepharose beads. Drawn with Chemdraw. Adapted from [46]

1.7. Digestion of proteins

When it comes to proteins and in vivo conditions, complicating factors such as missed cleavages of the digesting enzyme or post-translational modifications lead to a rapid increase in the possible cross-linked peptide pairs. The complexity problem is further complicated by the fact that most crosslinker reagents are designed to bind the amino group of lysines. At the same time, trypsin is the most

commonly used protease that cleaves at the C-terminus of lysines and arginines. However, the enzyme is no longer capable of cleaving lysine when the amino acid is crosslinked.[47]

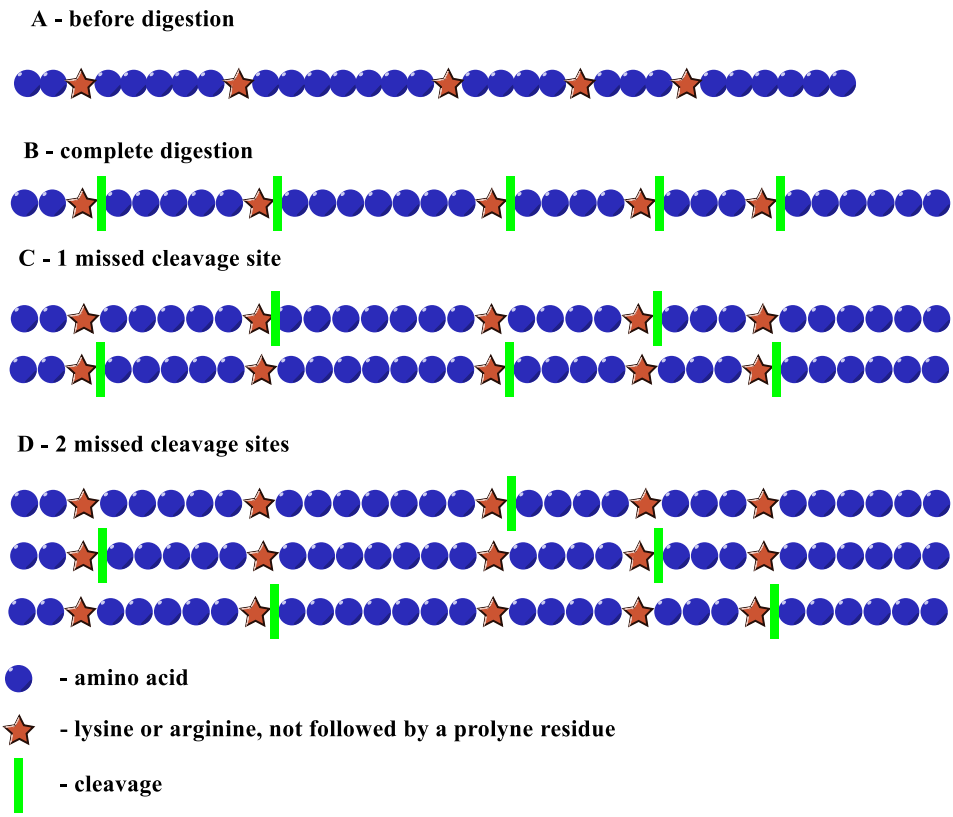


Figure 15: Schematic representation of protein sequence before and after proteolysis and an illustration of the multiplicity of the potential digestion products. (A) shows the protein sequence before digestion. (B) illustrates a fully digested protein and the simplest scenario. (C) However, due to either steric inaccessibility of the enzyme to residue, post-translational modification or prior crosslinking reaction, missing cleavage site my arise. In the case of a repetitive single missed cleavage, there two possible options of partial cleavage, depending on where the first missed cleavage occurs in the sequence. (D) 2 missed cleavage sites are producing 3 possible digestion patterns. Adapted from [47]

Digestion of proteins in the mass spectrometry-based proteomics plays a crucial role in the routine MS- based shotgun proteomics analysis. **Trypsin** has prevailed in this domain, due to it being largely commercially available, highly specific and easy to use. It belongs to the family of Serine proteases and hydrolyses peptides bonds at carboxyl side of K and R. A further upgrade of trypsin is trypsin gold, which has been modified through reductive methylation in order to reduce autolysis. Even though trypsin represents the golden standard, it has its limitations as well. Negatively charged aspartic acid and glutamic acid as well as phosphorylated S and T next to or near the K and R residues hamper the activity of the enzyme in these position and lead to missed cleavages. The popular "Keil rule", which states that trypsin does not cut before proline, must be mentioned here as well, even though it was shown in a study that this rule can be often violated.[48] Furthermore, the protease also manifest a higher cleavage efficiency to R than to K residues.[49]

Through time, it became clear that it may impose certain problems in the proteolysis of some proteins and researchers have found potent alternatives that are able to overcome the arising barriers. Some of the proteins may present regions with a higher density of arginines and lysins or with a very low density. Both of them can be a burden: a too high density could lead to very short peptide sequences. A short peptide sequence is harder to be assigned to a single protein. On the other hand, too long peptide sequences can also be detrimental. A series of alternatives is presented below.[50]

ArgC comes from the Cysteine proteases family and cleaves at the C-terminal of arginine. It is predominantly used to analyse post translational modifications and improve the proteome coverage. Even though it is more ineffective, it also targets the C termini of K residues.[50]

AspN belongs to the family of Metalloproteases and cleaves at N- Terminus of aspartic acid. A noticeable advantage is that it functions over a relatively broad pH range (between 4-9). In the presence of detergents in the solution, the cleavage can also occur at the N-Terminus of the glutamic acid residues. Many missed cleavages appear as result of a lower efficiency in cleaving after E in comparison to aspartic acid. The resulting peptide sequences tend to be longer than the tryptic peptides.[50]

Chymotrypsin is a serine protease and an excellent choice when it comes to digesting transmembrane segments of the membrane proteins. It cleaves at the C-terminal of F, Y, W, L and M, although with different efficiencies. The enzyme possesses a strong complementary power to trypsin in covering the proteome space.[50]

GluC shows a pH dependent behaviour: in a mildly acidic medium (at pH 4), the protease cleaves only at C terminus of glutamic acid. On the other hand, at pH 8 the enzyme targets the C termini of both glutamic acid and aspartic acid.[50]

LysC cleaves the C terminal of K even in 8 M urea solution. It shows an increased level of efficiency and specificity. That makes the protease especially attractive because it allows the digestion of proteins in their denatured form, which boosts the efficiency even further. [50]

LysN, a metalloprotease, has a better resistance to denaturants than trypsin and endures temperatures up to 70°C. It cleaves at the N-terminal of lysine. A limitation of this enzyme is that it sometimes (10% of the cases) cleaves to N termini of A, S and R.[50]

Pepsin, an aspartic protease, cleaves at the C terminus of aromatic amino acids (Y, F, W) and L. Its ability to work at low pH values makes pepsin a perfect digestion enzyme when disulfide bonds determination experiments are performed. Digestion in an acidic medium removes disulfide reshuffling.[50]

1.8. Nomenclature for the peptide crosslinks

Already in the early days of XL-MSMS it was recognised that the creation of a systematic nomenclature for the different types of peptides generated after digestion of cross-linked proteins is necessary. Shilling et al. proposed an accurate and simple system to describe the range of possible crosslinks.[51]

Type 0 refers to a linear peptide, which is modified on single amino acid with a crosslinking reagent that possesses a hydrolysed crosslinking group on the other end. These crosslinks also referred as dead-end modified peptides. They are not able to reveal any kind of information regarding protein-protein interactions or spatial proximity of that amino acid to other protein regions. On the other hand, type 0 crosslinks could provide an insight into the reactivity or accessibility of the residue. They can be compared to the peptide that present post translational modifications like phosphorylation or glycosylation.[51]

Internally crosslinked peptide or **type 1** crosslink is homobifunctional crosslinker reagent that binds 2 amino acids from the same peptide. Type 1 contains also a limited informational value apart from the confirmation of spatial closeness of the two residues, which mostly assumed as they are part of the same peptide sequence.[51]

The **type 2** crosslink is in the majority of cases the most relevant one and describes an interpeptide crosslink which connects two different peptide chains after they underwent proteolytic cleavage.[51]

The α and β nomenclature was also introduced to differentiate between the two chains. The α chain denotes the longer amino acid sequence and the β chain the shorter one. In case both peptides have the same length, the peptide with the higher molecular weight will be marked as alpha. In the unlikely event that both peptide sequences have the same molecular weight, the higher priority has the peptide chain whose first residue in the sequence possesses the higher mass. The nomenclature can extend in case multiple modifications are present.[51]

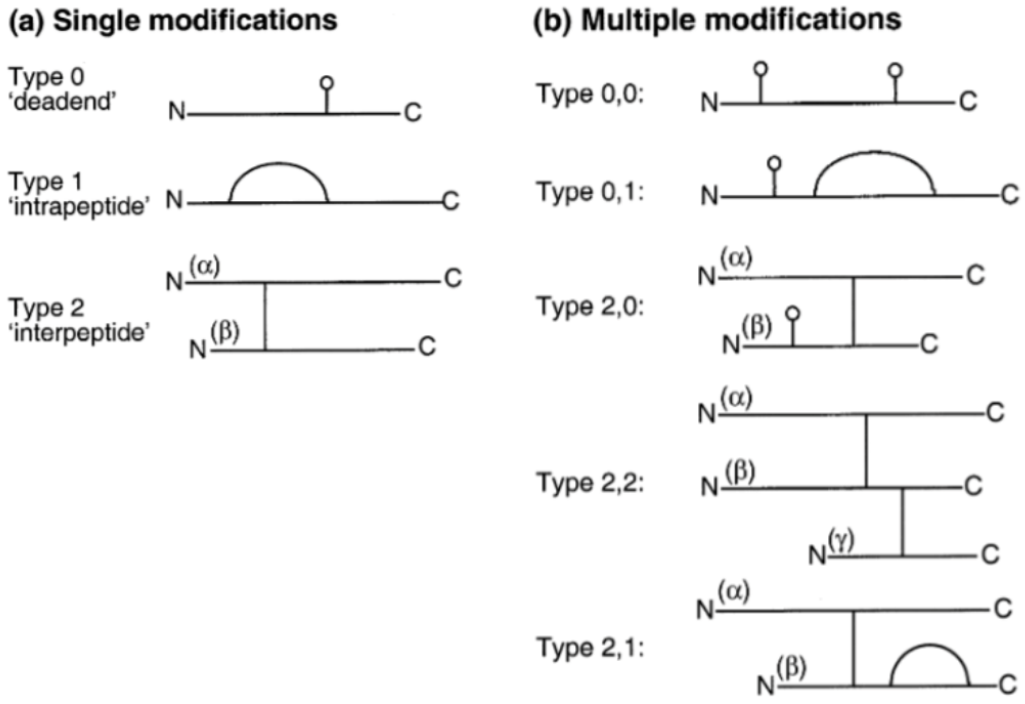


Figure 16: Classification of Crosslinks into type 0, type 1 and type 2. Reprinted from [51]

1.9. Chaperonin GroEL

GroEL is a well-known chaperon. The chaperonin protein helps other proteins to fold into their correct shapes and is composed of 14 smaller monomers placed as two back-to-back stacked circular heptamers.[52]

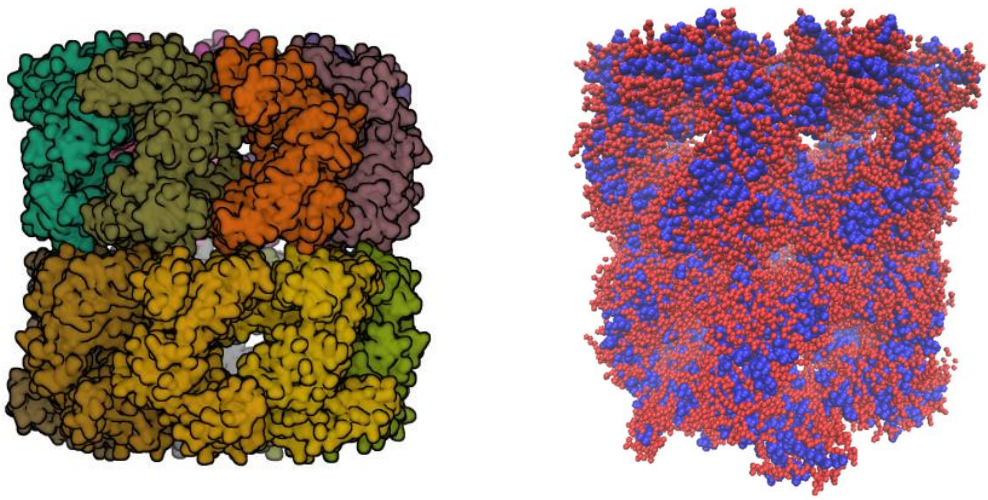


Figure 17: 3D Structure of GroEL. The left image was adapted from PDB[52,53]. The right image was generated with VMD 1.9.3 by using the PDB file of 1KP8. The hydrophobic residues are coloured in blue and displayed with a beta radius of 1.5, while the hydrophilic residues are displayed in red with a beta radius of 1.0. The drawing method chosen is VDW and the material chosen is opaque.

It is easily observable that the tubular complex presents an equilibrated amount of hydrophilic and hydrophobic residues in the outer part as well as in the inner core. This makes the protein not only feasible for solution in water, but also compatible for encapsulation in a micelle structure. Its overall rigidity is the main reason for being selected as model. In XL-MS rigid proteins can be used for benchmarking using 3D structure, but this would be problematic in the case of flexible proteins as their crystal structure does not mirror all the states present in solution.The noticeable inconvenience for choosing this model must be mentioned as well, namely its size.

2. Materials and methods

2.1. Peptide synthesis

The peptides were synthetized through solid phase peptide synthesis by using Fmoc chemistry on a SYRO machine with Tip Synthesis Module (MultiSynTech Gmbh). For each amino acid addition step, a HATU/DIPEA peptide coupling[54] was done. If the peptide sequence ended in the Lysine at C-terminus, the amino group of the Lysine sidechain was protected by an azide group in order to block a possible crosslinking reaction at this position.

N-termini synthetic peptides used were acetyl protected and started with the WGGGGR sequence. In addition, all peptides that were used to test the linkers reactive to the acidic amino acids end in an amide protected RGGGG sequence.

The chosen design of the peptides as well as group allocation are extensively discussed in the results section.

The synthetised peptides were then purified on a C18 Kinetex column (5 µm) and a 30 min gradient. To ensure the quality, all peptides were then measured separately on a 4800 MALDI-TOF/TOF from Applied Biosystems. After the lyophilisation, the peptides were then resolubilized in water. The peptide concentration was measured using a Nanodrop Denovix DS-11 FX+ Spectrophotometer at 280 nm. The respective extinction coefficient of each peptide was theoretically determined by using the ProtParam tool[55]. The volume of the solution was then reduced in a vacuum centrifuge up to the solubility limit and additionally diluted in 50mM HEPES at pH 7.5 to a concentration of 5mM. In case of solubility problems, the samples can be vortexed. Because of the many-steps procedure, each peptide was measured again at 4800 MALDI-TOF/TOF to reconfirm the quality of the samples. The peptides were mixed in groups and kept at -70°C until further use.

For quality control of the synthetised peptides, the absence of a Glycine in the GGGG sequence does not constitute a problem and will be not regarded as an impurity, as it does not affect any aspects of the designed experiments even if it is present in minimal relative amounts. Before being measured in the mass spectrometer, peptides were digested and the GGGG sequence was cut away. In the case of a miscleavage, the peptide would not be recognised by the software anyway as the data base contains the origin-proteins and not the initial undigested peptides.

2.2. Generation of tryptic HEK peptides

Cells were kept on ice during the whole procedure and centrifuges were cooled to 4°C.

Cell pellets were generated by cratching (from dishes) or by trypsinisation (from flasks). Around 10^7 were harvested. The pellet was then washed 1-2 times using PBS. The cells were repelleted again by centrifugation (3 min, 1300 rcf) and the supernatant was removed. Lysis was performed using 10 M urea, 50 mM HCl – (in total 1-2 ml for the whole pellet). Subsequently the cellular membrane was disrupted by sonication for 20 seconds at maximum power. The sonication was repeated three times. The lysate was additionally cleared by a centrifugation at 1400 rcf for 3 minutes. 1M of TRIS was then added to obtain a final concentration of 100 mM and effectively neutralise the solution.

A 0,5M DTT stock solution (diluted in 100mM TRIS) was added until the final concentration of 10mM was achieved. That corresponds to 1:50 in volume. 1 µl of Benzonase was added and the obtained solution was incubated for 1h in the shaker at 37°C.

During the reduction step, a Bradford assay was performed in order to evaluate the total protein concentration. An approximate determination of the total concentration is important for calculating the amount of enzymes necessary later in the procedure.

1M IAA (diluted in 100mM TRIS) was prepared and added until a final concentration of 20 mM IAA. This corresponds to 1:50 in volume. Additionally, the solution was incubated for 30 min in the dark at room temperature. The protein solution was quenched with DTT to a final concentration of 5 mM DTT followed by another incubation of 30 min at room temperature.

The solution was diluted from 10M urea to 6 M urea with 100 mM TRIS. LysC was added (1:200 w:w based on the obtained Bradford results) followed by a 2h incubation at 37°C in the shaker. The solution

was then further diluted to 2.5M urea with 100 mM TRIS and Trypsin Gold was added (1:200) followed by an overnight incubation at 37°C.

The digestion was stopped by adding TFA to a final concentration of 1% (in volume).

A small aliquot of approximately 500ng (based on Bradford calculations) was being measured quantitatively (relative to a known quantity of peptides) on a HPLC for qualitative purposes and yield determination.

The resulting solution was then subjected to a desalting procedure with a Seppack TFA C18 column. The desalting procedure was done in accordance with the instructions provided by the manufacturer. These may vary depending on the manufacturer, size and type of column used.

Finally, another quantity-check of the desalted peptide solution was being done to assure the success and yield of the treatments, followed by aliquoting and storage at -70°C until usage.

2.3. Simplified Bradford procedure

Ten standard solutions with a concentration from 0.5µg/ml up to 10µg/ml of BSA were prepared. For that, the needed amount of a concentrated BSA (2mg/ml) solution was pipetted into each cuvette and then 1ml of Coomassie brilliant blue solution was added. After 5 minutes the absorption of each cuvette was measured. Absorption was measured at a wavelength of 595nm. All measurements were undertaken with a Denovix DS-11 FX+ Spectrophotometer. A calibration line was then created from the measured data.

The protein solution whose concentration must be determined was subjected to the same procedure. A certain amount was pipetted into a cuvette (depending on the presumed concentration) and 1 ml of Coomassie brilliant blue was added. The waiting time before measuring on the spectrophotometer was 5 minutes.

2.4. Optimization of crosslinking-protocol

The aim of this experiment was to determine which concentration is the most adequate and in which proportions the crosslinker must be added to the peptide library in order to obtain the best yields of crosslinking-reaction. Different concentrations and reaction times were analysed and the results were compared. For this experiment, DSSO (cleavable crosslinker) was used and a group of 10 synthetic peptides:

Ac-WGGGGRLDLYITVKGGGISGQAGAIR	P0A7X3
Ac-WGGGGRKAGFVTR	P0A7X3
Ac-WGGGGRKSSAAR	P0A7X3
Ac-WGGGGRQCKANPWQQFAETHNK(N_3)	P0AG67
Ac-WGGGGRANPWQQFAETHNKGDR	P0AG67
Ac-WGGGGRDTLHLEGKELEFK(N_3)	P0AG67
Ac-WGGGGRQLGEDPWVAIAKR	P0AG67
Ac-WGGGGRYPEGTKLTGR	P0AG67
Ac-WGGGGRTDKFIVR	P60422
Ac-WGGGGRYILAPKGLK(N_3)	P60422

2.5µl of each peptide (5mM in 50 mM HEPES, pH 7.5) were mixed and vortexed. The peptide mix solution was then distributed in 5 different Eppendorf tubes.

Sample A: reference sample, no crosslinker and 2.5µl ddH_2O will be added (2.5 h waiting)
Sample B: 0.5µl of 100mM DSSO in DMSO and 2µl ddH_20 will be added once (2.5 h waiting)
Sample C: 0.5µl of 20mM DSSO in DMSO will be added every 30 min, 5 times (2.5 h in total)
Sample D: 0.75µl of 100 mM DSSO in DMSO and 1.75ul dd H_2O will be added once (2.5 h waiting)
Sample E: 0.5µl of 30mM DSSO in DMSO will be added every 30 min, 5 times (2.5 h in total)

The DSSO stock solutions must be kept at 4°C during waiting times. The samples were shaken at RT. 7.5µl of each sample were additionally mixed with 67.5µl ABC 100mmol and shortly vortexed. 75ng Trypsine were added to each sample and let overnight for digestion at 37°C. TCEP was added to a final concentration of 50mM and incubated for 30 min at 56°C. The solution was then acidified with 10% TFA to a final concentration of 1%.

2.5. Sample preparation

Lysine reactive crosslinker-reagents – main peptide library

In the case of the main peptide library, 9.3 mM of cross-linker (DSSO, DSBU, DSBSO, CDI) stock solution was prepared in dry DMSO. Firstly, 1 µl of each peptide group was pipetted in a separate Eppendorf tube. Additionally, 0.5 µl of the crosslinker solution was added in each vial 5 times in 30 min intervals at room temperature. The resulting final volume adds up to 3.5µl in each group. After 2.5h reaction time, the solutions were quenched by pipetting 31.5 µl 100 mM ABC for another 30 minutes at RT. The groups were then polled into a single vial and digested by adding 5 ng Trypsin/group and let overnight at 37°C. The actual quantity of trypsin was calculated by multiplying with the total number of groups pooled. The azide protection groups of the C-Terminus Lysins were then finally reduced to amine groups by adding TCEP to a final concentration of 50 mM and incubated for 30 min at RT. The reduced pooled peptide groups were additionally aliquoted and stored at -70°C until further use.

Lysine reactive crosslinker-reagents – enrichable peptide library

When performing the normal procedure (without HEK peptides spiking) the same steps were followed apart from TCEP reduction.

Aspartic and glutamic acid reactive crosslinker- reagent – acidic peptide library

The quantities (considerably increased) for this procedure were adapted for peptide mixes in order to maximise the number of crosslinks as all previous experiments described in the literature refer just to protein mixtures. These methods are discussed in the results section.

For 1 µl of each peptide group from the acidic peptide library 0.25 µl D/E reactive crosslink (DHSO or ADH 300 mM in 25 mM HEPES at pH 7.5) and 0.25 µl of DMTMM (1.2 M in 25 mM HEPES) were added 5 times in 30 min time intervals. After a total reaction time of 2.5 h, the reaction was quenched by pipetting TFA to a final concentration of 4% (w/v) and incubated for 20 min. Furthermore, the solution was then re-neutralized by adding 50 µl of 1 M TRIS buffer solution at pH 7.5. Additionally, the peptide groups were then mixed and incubated over night at 37°C with the suitable amount of trypsin necessary for digestion (see procedure above).

2.6. GroEL micelle formation and crosslinking protocol improvement

The following protocol was firstly developed on BSA out of financial reasons. It was then tested on GroEL (E. Coli, recombinant from Enzo). The reason for choosing GroEL was to compare the results obtained by our group with the quality of the results obtained in a particularly interesting approach, namely in gel crosslinking. The in-gel-crosslinking topic is covered in the results section as well.
Because the GroEL was acquired in TRIS buffer, the buffer needed to be exchanged with 50mM HEPES at pH 7.5, which does not hinder the crosslinking process. This was done by using a Zeba™ Spin Desalting Column (7K MWCO, 0.5 mL Catalog number: 89882) and the manufacturer instructions were followed.
The fatty acids used were firstly dissolved in dry DMSO. The necessary DMSO volume can vary depending on the fatty acid.
The 5 following solutions were prepared:
Sample A: 5 µl of 1 µg/µl GroEL in 50 mM HEPES
Sample B: 5 µl of 1 µg/µl GroEL in 50 mM HEPES, 10 µg decanoic acid
Sample C: 5 µl of 1 µg/µl GroEL in 50 mM HEPES, 20 µg decanoic acid
Sample D: 5 µl of 1 µg/µl GroEL in 50 mM HEPES, 10 µg dodecanoic acid
Sample E: 5 µl of 1 µg/µl GroEL in 50 mM HEPES, 20 µg dodecanoic acid

After preparations, the solutions were incubated 10 minutes at room temperature. 2 µl of 10mM DSSO was added to each tube and incubated for 2 h at room temperature. Additionally, the reaction was quenched using 1 M TRIS to a final concentration of 50 mM. The fatty acids were then removed with the help Zeba™ Spin Desalting Column 7K MWCO again. The crosslinked protein solution was then reduced with DTT to a final concentration of 10 mM and let to react for 30 min at 56 °C. The newly formed thiol groups were additionally alkylated with IAA to a final concentration of 20 mM and incubated in the dark for 30 min. Furthermore, the reaction mix was quenched with DTT (half of the amount used for the reduction step). Finally, the solution was digested with trypsin (100ng trypsin /5µg protein) overnight at 37°C.

2.7. Enrichment strategies

In order to mimic the complex mixtures present in in vivo conditions and simultaneously still being able to accurately analyse and validate the results, the crosslinked digested peptide pool mix from the enrichable peptide library were combined with 5 to 100 times excess (in mass) of previously generated tryptic HEK peptides. The resulted sample was then exposed to different enrichment strategies presented below.

Size exclusion chromatography (SEC) procedure

Approximately 10 µg of the crosslinked enrichable peptide library and the respective amount of tryptic HEK peptides (depending by which magnitude the sample is spiked) were fractioned on a TSKgel SuperSW2000 column (300 mm × 4.5 mm × 4 µm from Tosoh Bioscience) with a flowing rate of 200 µl/min in 30 % ACN, 0.1% TFA in a 30 min run. The fractions were collected each minute. Additionally, 20 µl of 1 M TRIS was added to each vial followed by an ACN removal in a vacuum centrifuge. TRIS addition can become a critical step for DSBSO before the volume-reduction step depending on what is done with the sample afterwards. If the sample undertakes afterwards an affinity enrichment as well, the maintaining of acetal group is vital, hence avoiding a gradual acidification of the solution by ACN removal under pressure.

Affinity enrichment procedure

The DSBSO crosslinked peptide library (together with varying amounts of linear HEK peptides) can also be enriched by using DBCO bounded to magnetic beads. The NHS-activated Sepharose fast flow (#17-0906-01, from GE Healthcare) was firstly incubated in molar excess (x2) of DBCO amine (#761540, Sigma-Aldrich) for 1h at RT. Additionally, 6 washing steps were performed with 50 mM HEPES at pH 7.5. The DBCO-coupled Sepharose beads were added to the crosslinked sample (10 times molar excess of DBCO groups to the crosslinker) at room temperature and rotated gently for 2h. Because the crosslinked peptides were bounded to the beads in this procedure-state, the beads were washed 3 times with 50 mM HEPES pH 7.5 with 1M NaCl, 3 times with 10 % acetonitrile in water and 3 times in TRIS at pH 7.5. The volume used for each washing step was 5 times the bead volume. For the elution of the peptides, one bead volume of 2% TFA in water (v/v) was added and incubated for 1h at RT.

2.8. List of synthesized peptides and assignation to their respective crosslink group

List of synthesized peptides and assignation to their respective crosslink group

ID number	tryptic sequence (miscleavage at XL site)	synthesized sequence (Ac = acetyl, Am = amide, N3 = azide)	group annotation in main library	group annotation in enrichable library	group annotation in acidic library	annotated protein accession number
1	VALVAKIGENINIR	Ac-WGGGGRVALVAKIGENINIR	1;11	1;9	no	P0A6P1
2	EIAEKMVEGR	Ac-WGGGGREIAEKMVEGR	1;12	1;10	no	P0A6P1
3	KAGNVAADGVIK	Ac-WGGGGRKAGNVAADGVIK(N3)	1;13	no	no	P0A6P1
4	EFIAKLQANPAK	Ac-WGGGGREFIAKLQANPAK(N3)	1;14	no	no	P27302
5	MAALMKQR	Ac-WGGGGRMAALMKQR	1;15	1;11	no	P27302
6	ILKCGFR	Ac-WGGGGRILKCGFR	1;16	1;12	no	P0AG44
7	VKDLPGVR	Ac-WGGGGRVKDLPGVR	1;17	1;13	no	P0A7S3
8	MAHIEKQAGELQEK	Ac-WGGGGRMAHIEKQAGELQEK(N3)	1;18	no	no	P0A7W1
9	QAGELQEKLIAVNR	Ac-WGGGGRQAGELQEKLIAVNR	1;19	1;14	no	P0A7W1
10	EVPAAIQKAMEK	Ac-WGGGGREVPAAIQKAMEK(N3)	1;20	no	no	P0A7W1
11	VKGGFTVELNGIR	Ac-WGGGGRVKGGFTVELNGIR	2;11	1;15	no	P0AG67
12	ITDVEVLKAQFEEER	Ac-WGGGGRITDVEVLKAQFEEER	2;12	1;16	no	P0A6P1
13	TGAGMMDCKK	Ac-WGGGGRTGAGMMDCKK(N3)	2;13	no	no	P0A6P1
14	KAGFVTR	Ac-WGGGGRKAGFVTR	2;14	2;9	no	P0A7X3
15	HKATLLGLGLR	Ac-WGGGGRHKATLLGLGLR	2;15	2;10	no	P0AG51
16	MAKTIK	Ac-WGGGGRMAKTIK(N3)	2;16	no	no	P0AG51
17	VKHPSEIVNVGDEITVK	Ac-WGGGGRVKHPSEIVNVGDEITVK(N3)	2;17	no	no	P0AG67
18	ITLNMGVGEAIADKK	Ac-WGGGGRITLNMGVGEAIADKK(N3)	2;18	no	no	P62399
19	QCKANPWQQFAETHNK	Ac-WGGGGRQCKANPWQQFAETHNK(N3)	2;19	no	no	P0AG67
20	ANPWQQFAETHNKGDR	Ac-WGGGGRANPWQQFAETHNKGDR	2;20	2;11	no	P0AG67
21	AMEKAR	Ac-WGGGGRAMEKAR	3;11	2;12	no	P0A7W1
22	IGVPFVDGGVIKAEVVAHGR	Ac-WGGGGRIGVPFVDGGVIKAEVVAHGR	3;12	2;13	no	P0AG48
23	DTLHLEGKELEFK	Ac-WGGGGRDTLHLEGKELEFK(N3)	3;13	no	no	P0AG67
24	TDKFIVR	Ac-WGGGGRTDKFIVR	3;14	2;14	no	P60422
25	TVGQLLKEHNAEVTGFIR	Ac-WGGGGRTVGQLLKEHNAEVTGFIR	3;15	2;15	no	P0A6P1
26	MAEITASLVKELR	Ac-WGGGGRMAEITASLVKELR	3;16	2;16	no	P0A6P1
27	SGAIKAAK	Ac-WGGGGRSGAIKAAK(N3)	3;17	no	no	P0A6P1
28	VVSMPSTDAFDKQDAAYR	Ac-WGGGGRVVSMPSTDAFDKQDAAYR	3;18	3;9	no	P27302
29	AVEEARAVTDKPSLLMCK	Ac-WGGGGRAVEEARAVTDKPSLLMCK(N3)	3;19	no	no	P27302
30	FAAYAKAYPQEAAEFTR	Ac-WGGGGRFAAYAKAYPQEAAEFTR	3;20	3;10	no	P27302
31	DFLKHNPQNPSWADR	Ac-WGGGGRDFLKHNPQNPSWADR	4;11	3;11	no	P27302
32	EAGQAKESAWNEK	Ac-WGGGGREAGQAKESAWNEK(N3)	4;12	no	no	P27302
33	MKGEMPSDFDAK	Ac-WGGGGRMKGEMPSDFDAK(N3)	4;13	no	no	P27302
34	DIDGHDAASIKR	Ac-WGGGGRDIDGHDAASIKR	4;14	3;12	no	P27302
35	IGVLVAAKGADEELVK	Ac-WGGGGRIGVLVAAKGADEELVK(N3)	4;15	no	no	P0A6P1
36	LQANPAKIASR	Ac-WGGGGRLQANPAKIASR	4;16	3;13	no	P27302

37	VVEPLITLAKTDSVANR	Ac-WGGGGRVVEPLITLAKTDSVANR	4;17	3;14	no	P0AG44
38	HEIIKTTLPK	Ac-WGGGGRHEIIKTTLPK(N3)	4;18	no	no	P0AG44
39	TTLPKAK	Ac-WGGGGRTTLPKAK(N3)	4;19	no	no	P0AG44
40	MAVQQNKPTRSK	Ac-WGGGGRMAVQQNKPTRSK(N3)	4;20	no	no	P0A7N4
41	TSGEKHLR	Ac-WGGGGRTSGEKHLR	5;11	3;15	no	P0A7N4
42	VFQTHSPVVDSISVKR	Ac-WGGGGRVFQTHSPVVDSISVKR	5;12	3;16	1;9	P0A7K6
43	KISNGEGVER	Ac-WGGGGRKISNGEGVER	5;13	4;9	no	P0A7K6
44	AKLYYLR	Ac-WGGGGRAKLYYLR	5;14	4;10	no	P0A7K6
45	LDLYITVKGGGISGQAGAIR	Ac-WGGGGRLDLYITVKGGGISGQAGAIR	5;15	4;11	no	P0A7X3
46	KSSAAR	Ac-WGGGGRKSSAAR	5;16	4;12	no	P0A7X3
47	AFNAKTDSIEK	Ac-WGGGGRAFNAKTDSIEK(N3)	5;17	no	no	P0A7J7
48	TIKITQTR	Ac-WGGGGRTIKITQTR	5;18	4;13	no	P0AG51
49	VAKSNVPALEACPQK	Ac-WGGGGRVAKSNVPALEACPQK(N3)	5;19	no	no	P0A7S3
50	SNVPALEACPQKR	Ac-WGGGGRSNVPALEACPQKR	5;20	4;14	6;12	P0A7S3
51	SKYGVK	Ac-WGGGGRSKYGVK(N3)	6;11	no	no	P0A7S3
52	DEADEKDAIATVNK	Ac-WGGGGRDEADEKDAIATVNK(N3)	6;12	no	no	P0AG67
53	VGFGYGKAR	Ac-WGGGGRVGFGYGKAR	6;13	4;15	no	P0A7W1
54	KLMTEFNYNSVMQVPR	Ac-WGGGGRKLMTEFNYNSVMQVPR	6;14	4;16	no	P62399
55	GLDITITTTAKSDEEGR	Ac-WGGGGRGLDITITTTAKSDEEGR	6;15	5;9	no	P62399
56	QGYPIGCKVTLR	Ac-WGGGGRQGYPIGCKVTLR	6;16	5;10	no	P62399
57	GLSAKSFDGR	Ac-WGGGGRGLSAKSFDGR	6;17	5;11	no	P62399
58	SVAGFKIR	Ac-WGGGGRSVAGFKIR	6;18	5;12	no	P62399
59	MYAVFQSGGKQHR	Ac-WGGGGRMYAVFQSGGKQHR	6;19	5;13	no	P0AG48
60	KQQGHR	Ac-WGGGGRKQQGHR	6;20	5;14	no	P0AG48
61	QLGEDPWVAIAKR	Ac-WGGGGRQLGEDPWVAIAKR	7;11	5;15	no	P0AG67
62	YPEGTKLTGR	Ac-WGGGGRYPEGTKLTGR	7;12	5;16	no	P0AG67
63	FWVESEKR	Ac-WGGGGRFWVESEKR	7;13	6;9	no	P0A7M2
64	SHALNATKR	Ac-WGGGGRSHALNATKR	4;7;14;15	6;10	no	P0A7M2
65	VSAKGMR	Ac-WGGGGRVSAKGMR	7;15	6;11	no	P0A7M2
66	KDIHPK	Ac-WGGGGRKDIHPK(N3)	7;16	no	no	P0A7M9
67	SHPFYTGKLR	Ac-WGGGGRSHPFYTGKLR	7;17	6;12	no	P0A7N1
68	IGSTIKTDR	Ac-WGGGGRIGSTIKTDR	7;18	6;13	no	P0A7N1
69	NFGKHPVTPWGVQTK	Ac-WGGGGRNFGKHPVTPWGVQTK(N3)	7;19	no	no	P60422
70	HPVTPWGVQTKGK	Ac-WGGGGRHPVTPWGVQTKGK(N3)	7;20	no	no	P60422
71	HVVKVVNPELHK	Ac-WGGGGRHVVKVVNPELHK(N3)	8;11	no	no	P60422
72	HIGGGHKQAYR	Ac-WGGGGRHIGGGHKQAYR	8;12	6;14	no	P60422
73	YILAPKGLK	Ac-WGGGGRYILAPKGLK(N3)	8;13	no	no	P60422
74	DLETQSQDGTFDKLTK	Ac-WGGGGRDLETQSQDGTFDKLTK(N3)	8;14	no	no	P0A7V0
75	DMLKAGVHFGHQTR	Ac-WGGGGRDMLKAGVHFGHQTR	8;15	6;15	no	P0A7V0
76	MTDKLTSLR	Ac-WGGGGRMTDKLTSLRGGGG-Am	8;16	6;16	no	P0A870
77	QPHAKGR	Ac-WGGGGRQPHAKGRGGGG-Am	8;17	7;9	no	P25888
78	DYSKYLNIR	Ac-WGGGGRDYSKYLNIRGGGG-Am	8;18	7;10	no	P25888
79	ASFDKANR	Ac-WGGGGRASFDKANRGGGG-Am	8;19	7;11	no	P0A715
80	EGGNEKVILCDR	Ac-WGGGGREGGNEKVILCDRGGGG-Am	8;20	7;12	no	P0A715
81	MTGRELKPHDR	Ac-WGGGGRMTGRELKPHDRGGGG-Am	9;11	7;13	no	P0A9Q1
82	VVNSKEDIR	Ac-WGGGGRVVNSKEDIRGGGG-Am	9;12	7;14	no	P0A836
83	EAEKYANPIPSR	Ac-WGGGGREAEKYANPIPSRGGGG-Am	9;13	7;15	no	P21499
84	SMKQAIYDPENR	Ac-WGGGGRSMKQAIYDPENRGGGG-Am	9;14	7;16	no	P21499
85	VGAATEVEMKEK	Ac-WGGGGRVGAATEVEMKEK(N3)	9;15	no	no	P0A6F5
86	LWDKETLEK	Ac-WGGGGRLWDKETLEK(N3)	9;16	no	no	P0A7D7
87	LSYDTEASIAKAK	Ac-WGGGGRLSYDTEASIAKAK(N3)	9;17	no	no	P0A870
88	QMKAIDDLVK	Ac-WGGGGRQMKAIDDLVK(N3)	9;18	no	no	P0A715
89	ESVLPKAVTAR	Ac-WGGGGRESVLPKAVTAR	9;19	8;9	no	P27302
90	EKLQER	Ac-WGGGGREKLQER	9;20	8;10	no	P0A6F5
91	AQSLKEIK	Ac-WGGGGRAQSLKEIK(N3)	10;11	no	no	P00956
92	GEVKGK	Ac-WGGGGRGEVKGK(N3)	10;12	no	no	P0ABB0
93	QLDHGQKVTELLK	Ac-WGGGGRQLDHGQKVTELLK(N3)	10;13	no	no	P0ABB0

94	GLKVALSK	Ac-WGGGGRGLKVALSK(N3)	10;14	no	no	P00956
95	DVKFGNDAR	Ac-WGGGGRDVKFGNDAR	10;15	8;11	no	P0A6F5
96	DAAAAVGKAVAER	GGGG-AmRDAAAAVGKAVAER	10;16	8;12	no	P0C018
97	KLQLVGVGYR	GGGG-AmRKLQLVGVGYR	10;17	8;13	no	P0AG55
98	VAVIKAVR	GGGG-AmRVAVIKAVR	10;18	8;14	no	P0A7K2
99	MQKQAELYR	Ac-WGGGGRMQKQAELYR	10;19	8;15	no	P0A7D7
100	ATIDGLENMNSPEMVAAKR	Ac-WGGGGRATIDGLENMNSPEMVAAKR	10;20	8;16	no	P0A7W1
101	QALELPR	Ac-WGGGGRQALELPRGGGG-Am	no	no	1;7	P0A850
102	SQAIEGLVK	Ac-WGGGGRSQAIEGLVKGGGG-Am	no	no	1;8	P0A850
103	SELVNVAK	Ac-WGGGGRSELVNVAKGGGG-Am	no	no	1;10	P0A850
104	MTEAMK	Ac-WGGGGRMTEAMKGGGG-Am	no	no	1;11	P37095
105	IDGLGIK	Ac-WGGGGRIDGLGIKGGGG-Am	no	no	1;12	P37095
106	LPLAEFHR	Ac-WGGGGRLPLAEFHRGGGG-Am	no	no	1;13	P37095
107	QTAFMDSMK	Ac-WGGGGRQTAFMDSMKGGGG-Am	no	no	2;7	P37095
108	ITLSTQPADAR	Ac-WGGGGRITLSTQPADARGGGG-Am	no	no	2;8	P37095
109	GITFDSGGYSIK	Ac-WGGGGRGITFDSGGYSIKGGGG-Am	no	no	2;9	P37095
110	EQGYMGLHTVGR	Ac-WGGGGREQGYMGLHTVGRGGGG-Am	no	no	2;10	P37095
111	EAGQAK	Ac-WGGGGREAGQAKGGGG-Am	no	no	2;11	P27302
112	LTAEGVK	Ac-WGGGGRLTAEGVKGGGG-Am	no	no	2;12	P27302
113	QDAAYR	Ac-WGGGGRQDAAYRGGGG-Am	no	no	2;13	P27302
114	ALSMDAVQK	Ac-WGGGGRALSMDAVQKGGGG-Am	no	no	3;7	P27302
115	QNLAQQER	Ac-WGGGGRQNLAQQERGGGG-Am	no	no	3;8	P27302
116	QDGPTALILSR	Ac-WGGGGRQDGPTALILSRGGGG-Am	no	no	3;9	P27302
117	AVTDKPSLLMCK	Ac-WGGGGRAVTDKPSLLMCKGGGG-Am	no	no	3;10	P27302
118	VLDAAVAGK	Ac-WGGGGRVLDAAVAGKGGGG-Am	no	no	3;11	P0A6P1
119	TGAGMMDCK	Ac-WGGGGRTGAGMMDCKGGGG-Am	no	no	3;12	P0A6P1
120	AGNVAADGVIK	Ac-WGGGGRAGNVAADGVIKGGGG-Am	no	no	3;13	P0A6P1
121	MAEITASLVK	Ac-WGGGGRMAEITASLVKGGGG-Am	no	no	4;7	P0A6P1
122	EAAPAK	Ac-WGGGGREAAPAKGGGG-Am	no	no	4;8	P0A9Q7
123	AELGIPK	Ac-WGGGGRAELGIPKGGGG-Am	no	no	4;9	P0A9Q7
124	ALIVTDR	Ac-WGGGGRALIVTDRGGGG-Am	no	no	4;10	P0A9Q7
125	AVQDVILK	Ac-WGGGGRAVQDVILKGGGG-Am	no	no	4;11	P0A9Q7
126	AAALAAADAR	Ac-WGGGGRAAALAAADARGGGG-Am	no	no	4;12	P0A9Q7
127	YPLISELK	Ac-WGGGGRYPLISELKGGGG-Am	no	no	4;13	P0A9Q7
128	YNANDNPTK	Ac-WGGGGRYNANDNPTKGGGG-Am	no	no	5;7	P0A9Q7
129	TGDTISGK	Ac-WGGGGRTGDTISGKGGGG-Am	no	no	5;8	P0AG30
130	MNLTELK	Ac-WGGGGRMNLTELKGGGG-Am	no	no	5;9	P0AG30
131	VFPAIDYNR	Ac-WGGGGRVFPAIDYNRGGGG-Am	no	no	5;10	P0AG30
132	GTGNMELHLSR	Ac-WGGGGRGTGNMELHLSRGGGG-Am	no	no	5;11	P0AG30
133	AGNDANR	Ac-WGGGGRAGNDANRGGGG-Am	no	no	5;12	P00864
134	VLGETIK	Ac-WGGGGRVLGETIKGGGG-Am	no	no	5;13	P00864
135	ATDLFLK	Ac-WGGGGRATDLFLKGGGG-Am	no	no	6;7	P00864
136	MVEVNACLK	Ac-WGGGGRMVEVNACLKGGGG-Am	no	no	6;8	P00864
137	LPVEFVPVR	Ac-WGGGGRLPVEFVPVRGGGG-Am	no	no	6;9	P00864
138	MNEQYSALR	Ac-WGGGGRMNEQYSALRGGGG-Am	no	no	6;10	P00864
139	SLTEIK	Ac-WGGGGRSLTEIKGGGG-Am	no	no	6;11	P0A7Z4
140	HWDQQK	Ac-WGGGGRHWDQQKGGGG-Am	no	no	6;13	P25665
141	AQESYWAGNSTR	Ac-WGGGGRAQESYWAGNSTRGGGG-Am	no	no	7;	P25665

3. Results

3.1. Concentration refinement

Before strategically experimenting on the synthetic peptide libraries under different reaction mediums and with different reagents, there was a necessity to establish an optimal concentration of the linker employed in the reaction, as well as an optimal strategy regarding the interval when the linker should be added. For this, four different methodologies were tested together with an unlinked reference.

Sample Name	Experimental conditions
A	Reference sample, not crosslinked
B	0.5 µl DSSO 100mM in DMSO, 2 µl dd H_2O – added once – 2.5h reaction time
C	0.5 µl DSSO 20mM in DMSO – added 5 times – in 0.5h intervals (2.5h in total)
D	0.75 µl DSSO 100mM in DMSO, 1.75 µl dd H_2O – added once – 2.5h reaction time
E	0.5 µl DSSO 30mM in DMSO – added 5 times – in 0.5h intervals (2.5h in total)

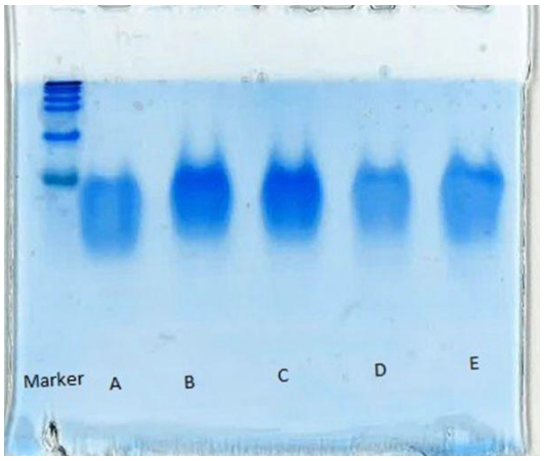


Figure 18: Gel electrophoresis of the samples after the crosslinking reaction. Electrophoresis apparatus settings: running at 150V max at a constant 35 mA for 60 min. followed by 3 times gel-rinsing with ddH_2O, staining overnight at room temperature using Coomassie Blue (MBS-Blue + 2% NaCl supplemented for incubation), destaining with water 3 times, 1 h each.

The gel was allowed to run "just" for 60 minutes. It is clearly observable that the evolution of the PageRuler Prestained Protein Ladder is not complete. The gel was also prepared manually because both synthetic peptides and their crosslinks possess low molecular weights in comparison to proteins. The gel percentage chosen is 20% (percentage of acrylamide/bisacrylamide used in the gel matrix) to achieve smaller pores and better separate smaller compounds than commercially available gels with a lower percentage. This is still not sufficient for compounds that normally do not exceed 5000 Da to be separated efficiently. On the other hand, at this gel concentration, one can observe that the unlinked sample (sample A) has migrated further than the rest of the samples. This slight migration difference is an indicator of successful crosslinking.

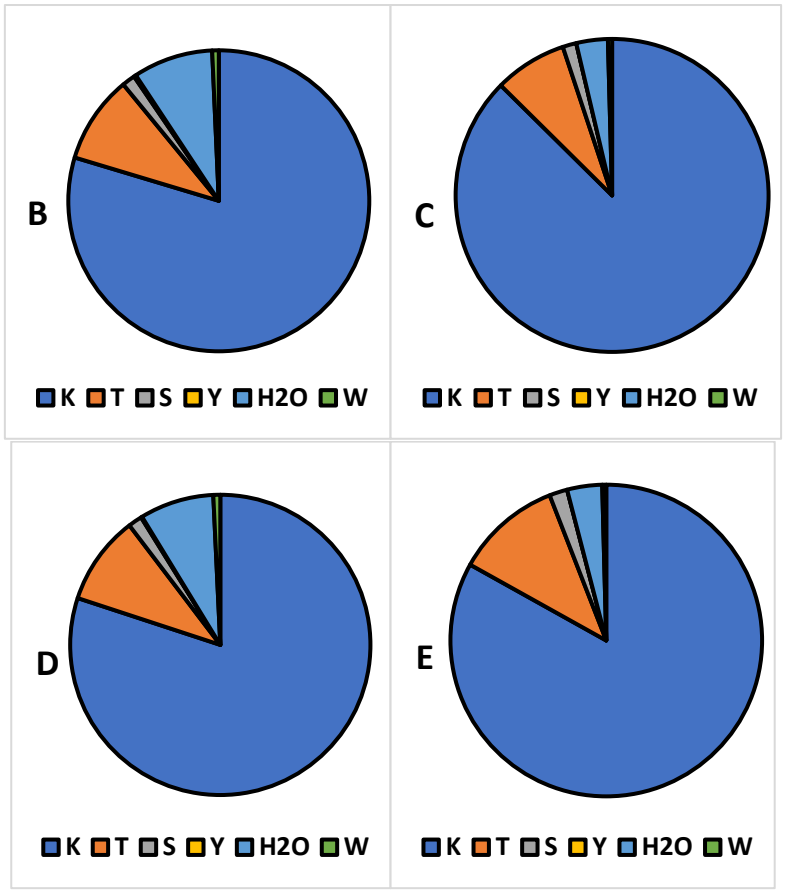


Figure 19: Pie diagram of the reacted residues/compounds on the CSM level depending on different reaction conditions. Both ends were considered individually: In the case of water, one end of the linker reacted with an amino group from a peptide and the other end was hydrolysed resulting in a type 0 crosslink (dead-end). In the case of tryptophan, a crosslinking reaction can take place at the amino group of the N-terminus if the acetyl group is hydrolysed, and the amino group remains unprotected. The analysis was realised with MeroX and the crosslinking sites allowed were KSTY-KSTY, FDR 5%. Type 0 links were not filtered out.

Apart from the lysine residues, which prove to have the highest reactivity, threonine shows a moderate reactivity as well, and is involved in 7-11% of the crosslinks. The difference between threonine and serine, which seems to be substantially more inert, is only a methyl group in the side chain. This methyl could have two roles which influence positively the reaction kinetics: on the one hand it could sterically hinder the hydroxyl group to get involved in hydrogen bridges; on the other hand, the methyl group is also electron withdrawing. It must also be mentioned that there are in total 8 residues of threonine and only 3 residues of serine and 3 residues of tyrosine (assuming equimolar amounts of peptides).

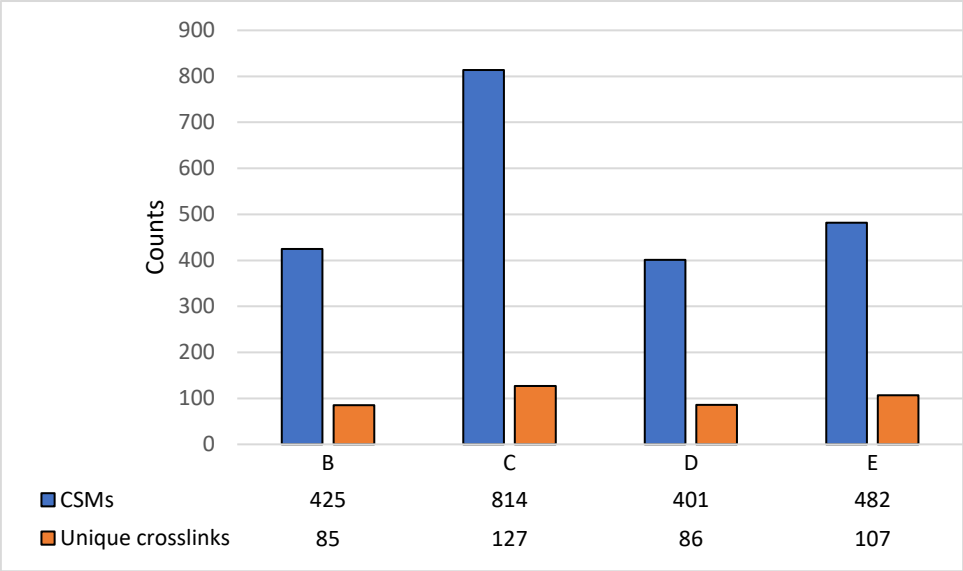


Figure 20: Number of CSMs and unique crosslinks detect by Merox under different reaction conditions, independent on the crosslinking site and crosslink type at a 5% FDR and KSTY-KSTY as possible binding residues.

The highest number of CSMs as well as the biggest diversity of unique crosslinks found was obtained when 0.5 µl of 20 mM DSSO (in DMSO) was added in a 30 minutes interval. Interestingly, when the same procedure (volume and time intervals) is carried out, but with a higher linker concentration (as in sample E), a lower number of links and CSM are being observed. A cause for that would be that too much excess linker just interferes with the measurement, making the XLs unidentifiable for the search-engines. The same explanation could be used when comparing the number of links from the sample IDs B and D. Even though sample D has a higher concentration of linker, a lower number of links is observed. The strategy of adding linker multiple times in shorter time spans proves advantageous in comparison to one add and longer waiting times. The reagent will get hydrolysed after a certain period making it unusable.

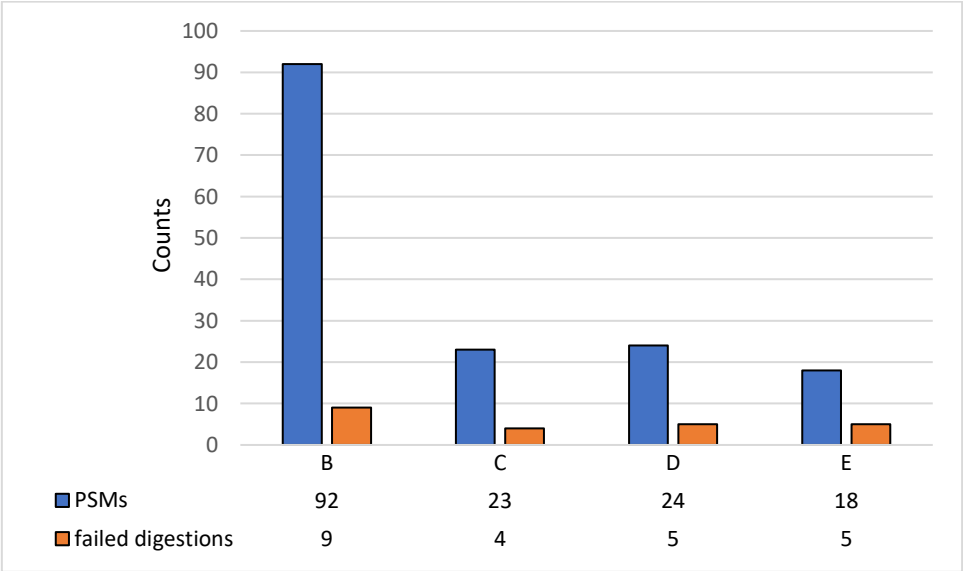


Figure 21: Number of Protein Spectrum Matches (PSMs) and failed digestions in the identified PSMs depending on different experimental conditions analysed with MS Amanda[56] in a Proteome Discoverer workflow. The shown number of peptides have passed an additional filtering with the condition that the score >150. Through the expression "failed digestion" it is meant that the "WGGGGR" was not cleaved from synthetic peptide. Only ions with a charge >+2 were considered.

All the raw files (from experimental conditions A-E) were analysed simultaneously in a Proteome Discoverer workflow where MS Amanda was utilized as peptide-search engine. Initially, 769 peptides

were discovered, but after an additional filtering process (score >150 which from experience has proven to be necessary and good practice in order to be able to ensure a high quality of the results) 651 remained. 494 out of 651 were traceable to the experimental conditions from A. 67 out of 651 peptides were still presenting "WGGGGR" sequence denoting a failed digestion of the trypsin and 10 PSMs have been detected with an azide group ($-N_3$), proving a failed reduction to the amino-group.

3.2. Comparison of Synthetic libraries and study design

Table 1: Comparison of firstly proposed peptide library for benchmarking crosslinking workflows with the state-of-the-art peptide library workflow.

	Beveridge R. et al.[57]	**Matzinger M., Vasiu A. et al.**[58]
Peptide library support	**small peptide library with peptides from S. pyogenes Cas9 designed to work with Lys reactive linkers.**	**enlarged peptide library comprising peptides from 38 E.coli ribosomal proteins and compatible to various linker chemistries**
Group structuring	**Simple group structure (95 peptides,12 groups, 426 allowed XLs**	**Complex group structure (presence of each peptide in 2 groups) see Figure 22**
Analysed crosslinkers	**DSSO, DSBU, DSS**	**DSBU, DSSO, CDI, ADH/DMTMM, DHSO/DMTMM, DSBSO**
Crosslink search engines analysed	**Xi, MeroX, XlinkX 2, Stravox, pLink 2, Kojak**	**MeroX, XlinkX 2, Annika, pLink 2, (later MaxLynx and xiSearch added)**
Software package	**no software package published**	**FDR-check tool, physicochemical investigations, Venn Diagrams incorporated in IMP-X-FDR user interface**

Even though Beveridge's work is remarkable and brought new aspects to light when it comes to FDR credibility, in our work many improvements are being made. Her work was a milestone in the Crosslinking Mass Spectrometry domain because it was the first time when synthetic peptide library with a specific design of both groups and peptide sequence (which restricted the number of possible results) was implemented. In the work of Beveridge R et al.[57], 95 synthetic peptides were synthetised. They are all originating from one single protein sequence, namely S. pyogenes Cas 9. By choosing peptide sequences from only one protein, the crosslinks search engines will interpret all possible links as intralinks. They were then divided into 12 groups and 425 crosslinks were practically possible, when the groups were separately crosslinked. Theoretically, the software will interpret 4560 possible crosslinks (if all peptides sequence would have been mixed). Additionally, the data base against the crosslinks were searched from included sequence of S. pyogenes Cas 9 and the sequences of 10 other proteins from CRAPome. In our work, 141 peptides were synthetised, and three different peptide libraries were created. The main peptide library contained 100 peptides which were cleverly separated into 20 groups, in such a way that a crosslink cannot occur in two different groups (expect for the crosslinks between two peptides with the same sequence). Therefore, every peptide sequence was present in two different groups. Because of the aspiration to reach a golden standard, after the quantification of peptides and realisation of experiments, a supplementary quality control was

performed again. By examining every single peptide in MALDI-TOF, traces of peptide 64 (Ac-WGGGGRSHALNATKR) were found in peptide 35 (Ac-WGGGGRIGVLVAAKGADEELVK(N3)). As a result, peptide 64 was included in all the groups where peptide 35 was present for the FDR-check. In total, 1018 crosslinks can be reached in reality and 5050 are possible if all peptides were mixed in one group. For the enrichable library, 64 peptide sequences are separated into 16 groups and 512 crosslinks are potentially possible. In the case of the acidic library, 43 peptides were divided into 13 groups and 280 crosslinks are made possible.

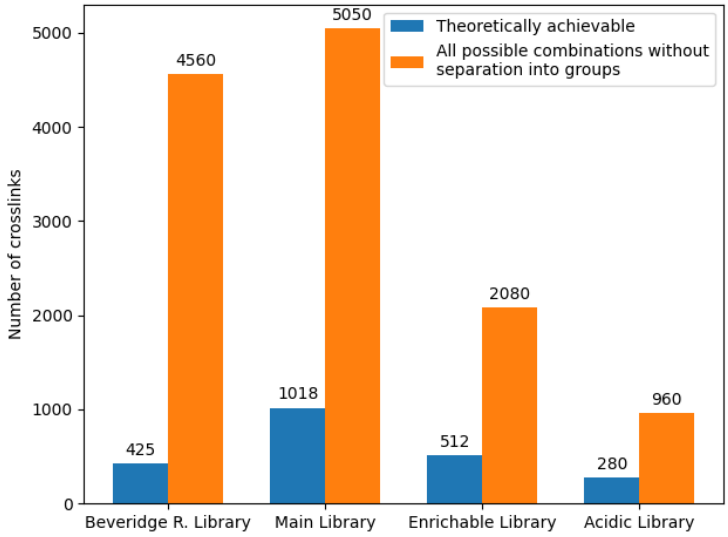


Figure 22: Comparison of peptide libraries' potential crosslinks

The crosslinker choice for benchmarking also differs. In the first peptide library DSSO, DSBU and DSS were utilized. These options seem very viable choices and of interest since DSS is the most utilized crosslinker. Because we had a variety of libraries at disposal, enrichable heterotrifunctional crosslinkers like DSBSO or acid-reactive linkers such as ADH/DMTMM and DHSO/DMTMM could also be studied. CDI was also included in the analysis thanks to its rather demanding steric conditions.

In Beveridge's work, the focus point regarding XL-search engines was on software that are mostly specialized in uncleavable crosslinkers. In our work, we opted for newer engines that can either accept cleavable crosslinkers or are specialized in cleavable crosslinkers.

The design of the single peptides also varies in some aspects:

In the precedent work, a biotin was covalently attached to a constant standard sequence "- YGGGGR -", which was then bounded to the N-terminal of the peptide sequence. In the inner part of the sequence (other than the N-terminus and C-terminus), a lysine residue is always present and allows the crosslinking reaction. At the C-terminus, there could either be an arginine or a lysine. In case there is a lysine at the C-terminus, it is present as an epsilon-azido-L-lysine in order to hinder the reaction of the amino group with the crosslinker.

The analyses were realized on a customised shotgun database incorporating 171 ribosomal proteins from E. coli at a FDR of 1%. In the case of Beveridge library, two types of fasta files were used for analyses: one with Cas9 and 10 other proteins from "Crapome" and one with Cas9 and an extended range of "Crapome"- proteins.

In our main peptide library, we renounced at the use of biotin. Not only that its presence can have an impact (beneficial or detrimental) on how the crosslinking works, which we cannot quantify, but it also interferes in the MS-analysis and must be removed prior to the analysis with streptavidin beads. The N-terminus of the peptide is covalently bounded to the "Ac-WGGGGR- "sequence. The

tryptophan's presence serves for quantifications purposes and its amino group is acetylated in order to block it from reacting with the crosslinker. The length of the following peptide sequence is minimum 5 and there is one lysine inside of it. The peptide sequence ends either with an arginine or with an epsilon-azido-L-lysine. By the simple replacing of the biotin with an acetylated tryptophan, the biological conditions and the actual structure of a protein are mimicked better.

The enrichable peptide library is a subset of the main one, where only those sequences which end with an arginine were chosen. The peptides ending in epsilon-azido-L-lysine cannot be used because they would react with the DBCO coupled Sepharose beads, which are intended to be used for enriching the crosslinks out of a highly complex biological system.

The structure of the peptides in the so called "acidic library" was adapted again for the intended scope. The N-terminus is connected to a "Ac-WGGGGR-" sequence and the C terminus (either an arginine or a lysine) is followed by "-GGGG-Amide". Instead of a carboxyl group in the last glycine, there is an amide for hampering a reaction with the acid sensitive crosslinker. The peptide must contain only one glutamic acid or aspartic acid.

Table 2: Representation of peptide design across different synthetic peptide libraries

Beveridge's Peptide Library	**Biotin-YGGGR-XXXKXXXK(N3)** **Biotin-YGGGR-XXXKXXXR**
Main Peptide Library	**Ac-WGGGGR-XXXKXXXK(N3)** **Ac-WGGGGR-XXXKXXXR**
Enrichable Peptide Library	**Ac-WGGGGR-XXXKXXXR**
Acidic Peptide Library	**Ac-WGGGGR-XXXD/EXXXRGGGG-Amide** **Ac-WGGGGR-XXXD/EXXXKGGGG-Amide**

As depicted in the figure below, DSSO and DSBU do show relatively low average FDR values of 2.7% and 2.8%, which can be considered good, but still above the expected value of 1%. While it is true that DSSO and DSBU present different fragmentation patterns and fragmentation energies, DSSO unveils ~12% more correct crosslinks than DSBU on the main peptide library, when using MS Annika as search engine. Of note MS Annika might be biased toward the analysis of DSSO crosslinked peptides as it was optimized using data from that crosslinker[59]. All other algorithms tested show better outcomes for DSBUI. This would be rather expected, as the spacer length in the case of DSBU is 12.5 Å compared to only 10.3 Å for DSSO. As there are no real known interactions between the synthetic peptides, the arm

length plays a decisive role. There is still some room for improvement in the XL-MS workflow as only ~60% of all possible XLs (for DSSO) are being found.

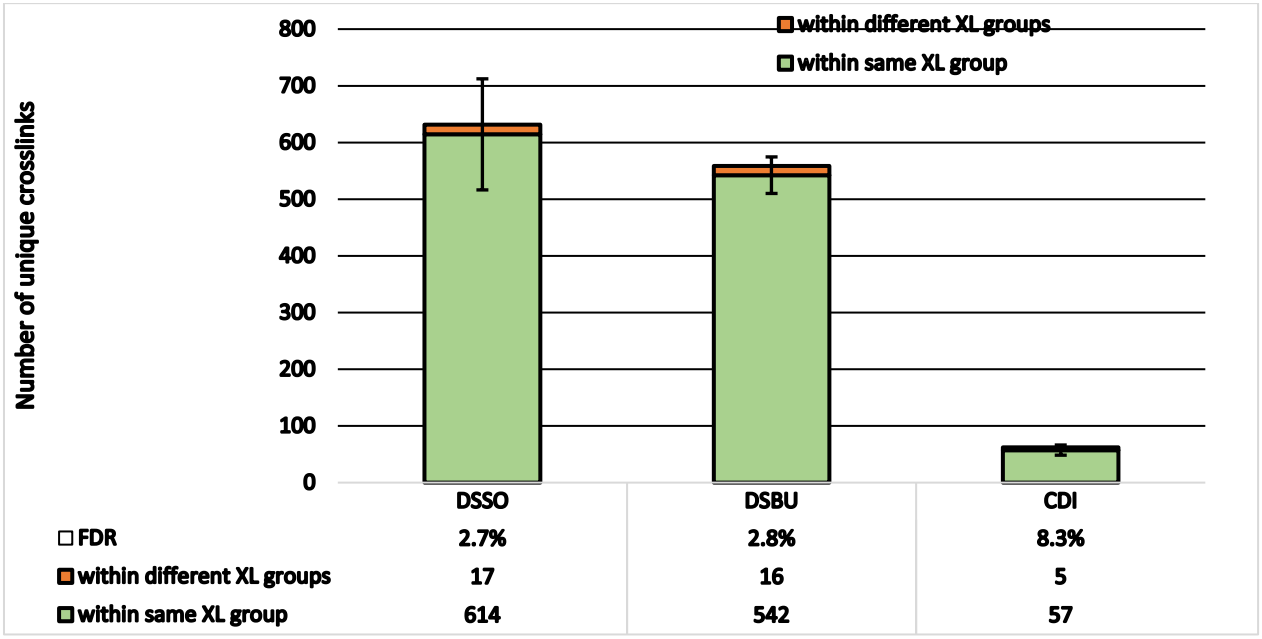


Figure 23: Average unique crosslinks identified utilizing different crosslinkers on the main peptide library. A stepped HCD MS2 acquisition strategy was used, and the raw data was interpreted with MS Annika integrated in a Proteome Discoverer workflow. The sample size is n=3 and the samples were measured on different days. The set FDR value is 1% and the search was conducted against database of ribosomal proteins from E. coli.

The interesting result is coming from 1,1'-carbonyldiimidazole: its spacer length is truly short (2.6 Å) and designed for interactions of higher resolution, but once one end of the linker reacted, the other will follow. Even though it is firstly hydrolysed, it is still able to react with an amino group from lysine/N-terminus or an hydroxy group from threonine, tyrosine or serine. Nevertheless, the CDI developer have reported a ratio of a 43% K–K to circa 57% K–S/T/Y crosslinks and no dead-end links.[60] With this advantage of not producing relatively information-poor dead-end links, the low number of XLs and high FDRs could also be due to the fact that the search was done exclusively for K-K and not for K-KTSY.

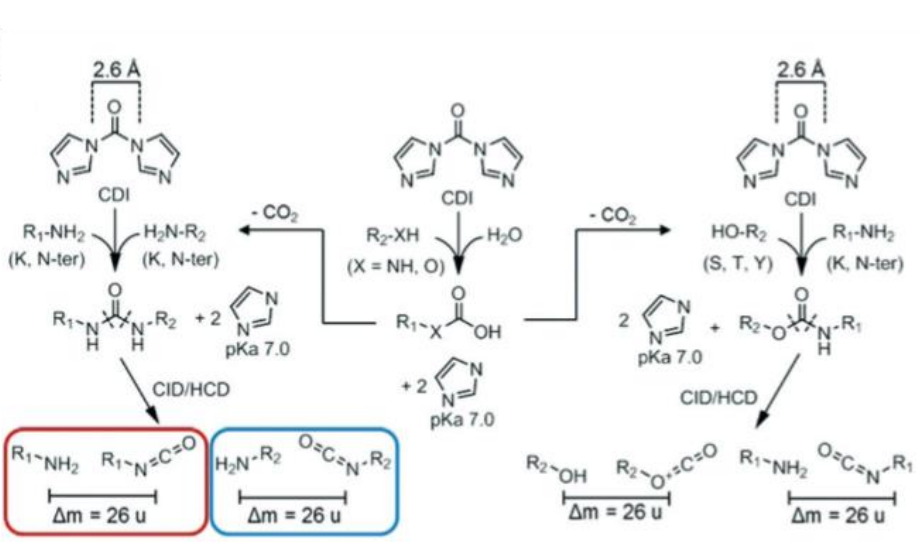


Figure 24: Schematic representation of CDI's reactivity towards lysine residues and towards residues containing a hydroxyl group. Figure adapted from [60]

The overlap between data coming from various reagents is high which is expected as the same linker chemistry was utilized.

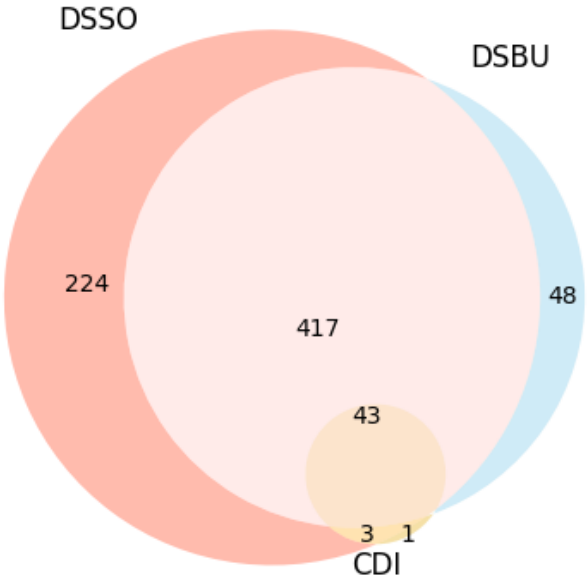


Figure 25: Overlap of the correctly identified crosslinks from a replicate of each reagent used on the main library. All the raw data was measured and analysed under the same conditions: stepped HCD, FDR=1%, ribosomal proteins from E. coli., MS Annika as XL search engine

For the enrichable peptide library, DSSO and DSBSO were investigated. DSSO shows a slightly higher average number of links, but also a slightly higher false discovery rate. The fact that, in this case, 373 correct XLs out of 512 possible (~73% in case of DSSO) and 350 correct XLs out of 512 (~68% for DSBSO) has stimulated a careful reflexion regarding the reaction yields. In the case of DSSO experiment with the main library only around 60% of all possible links were detected. The crosslinking protocol is the same as for the main library except for two aspects: the enrichable peptide set is a subset of the main

peptide collection which excludes the peptide sequences that have an azide-protected lysine at the C-terminus a TCEP reduction is employed as well for the main library to finally reduce the azide group. It is known that the reduction of the azide group with the TCEP does not occur completely, and a certain loss of data is present through the unrecognition links because of their different mass. Another argument, which has not been considered until now, is the loss of links through the reduction of the sulfoxide group to a sulfide. There is enough evidence that the reaction takes place under certain conditions with extremely high yields[61], which are partly present also in our system. For example, dibenzyl sulfoxide is 98% converted with 1.1 molar equivalent of TCEP in a dioxan solution under reflux for 1h. It is imaginable, that the reaction takes places in our conditions as well even though in considerably lower conversion yields as RT conditions are applied. If the sulfoxide group is reduced by TCEP in an already formed crosslink, the crosslinks becomes "invisible" and cannot be detected anymore. Even though the difference in total yields between the two libraries are only about 10%, many workflows do utilise TCEP as part of the experiment. Further research should be conducted to validate the hypothesis and assess the magnitude of the bottleneck in the case of this type of cleavable linkers.

The factor of sterical hindrance of the DSBSO-enrichable functionality can be ignored in this synthetical systems, as no real interaction takes place. Not just the absolute numbers of crosslinks differ only non-significantly, but the overlap of the identified links is also extremely high, both linkers covering almost the same crosslink-space. The overlap is comparable to the one obtained between different replicates of the same sample.

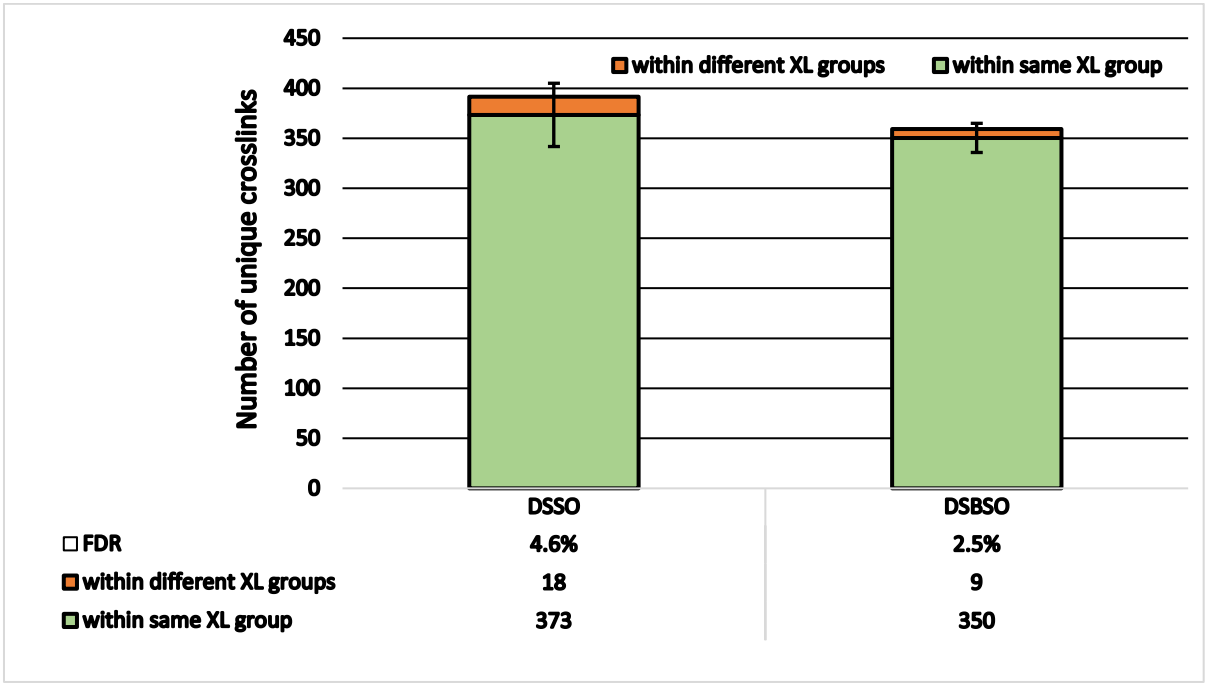


Figure 26: Average unique crosslinks identified utilizing DSSO and DSBSO crosslinkers on the enrichable peptide library without an actual enriching step. A stepped HCD MS2 acquisition strategy was used, and the raw data was interpreted with MS Annika integrated in a Proteome Discoverer workflow. The sample size is n=3 and the samples were measured on different days. The set FDR value is 1% and the search was conducted against database of ribosomal proteins from E. coli.

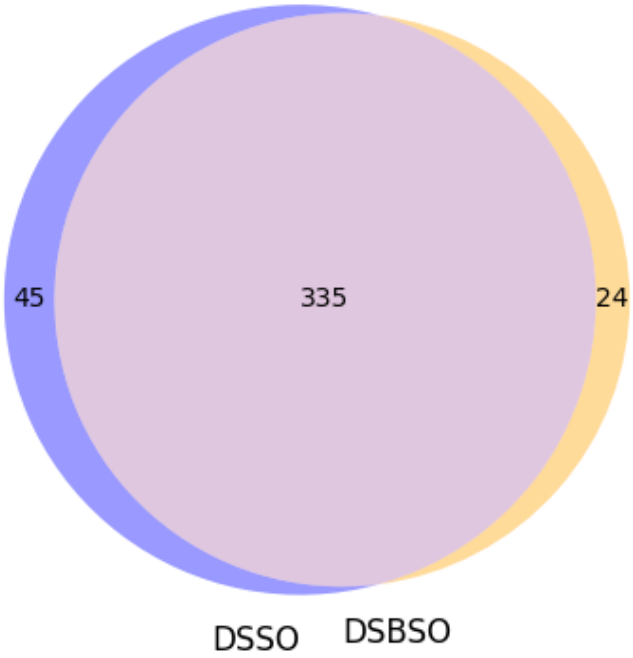


Figure 27: Overlap of the correctly identified crosslinks from a replicate of each reagent utilized on the enrichable peptide library. All the raw data files were measured and analysed by applying the same conditions: stepped HCD, FDR=1%, ribosomal proteins from E. coli., MS Annika as XL search engine

4-methyl morpholinium chloride (the coupling reagent DMTMM) is being used together with ADH and DHSO in the linking experiments of the acidic library, and also true zero-length linkers could theoretically result from this, where one K residue is coupled to a D/E. In the library there are two peptides that do have a lysine on the inner part of the sequence (ID=50 and ID=117). The lysine residues that would become the C-terminal after digestion are not considered because a link at these positions would impede the digestion and removal of the -GGGG-Am sequence in the first place. No evidence of zero length linkers was found in the data set analysed.

The hydrazine linkers shown in the figure below also present a low reactivity compared to the one targeting primarily lysines. In the experiment where the groups were separately linked, only 71 correct XLs were identified on average out of the total 280 (corresponding to ~25%), with a mean FDR of 4.9%. When it comes to the groups being mixed and linked together, the algorithms are not so strongly challenged and deliver a FDR of 2.5%, but still a low yield of only 182 correct XLs out of 960 possible (approximately 19%).

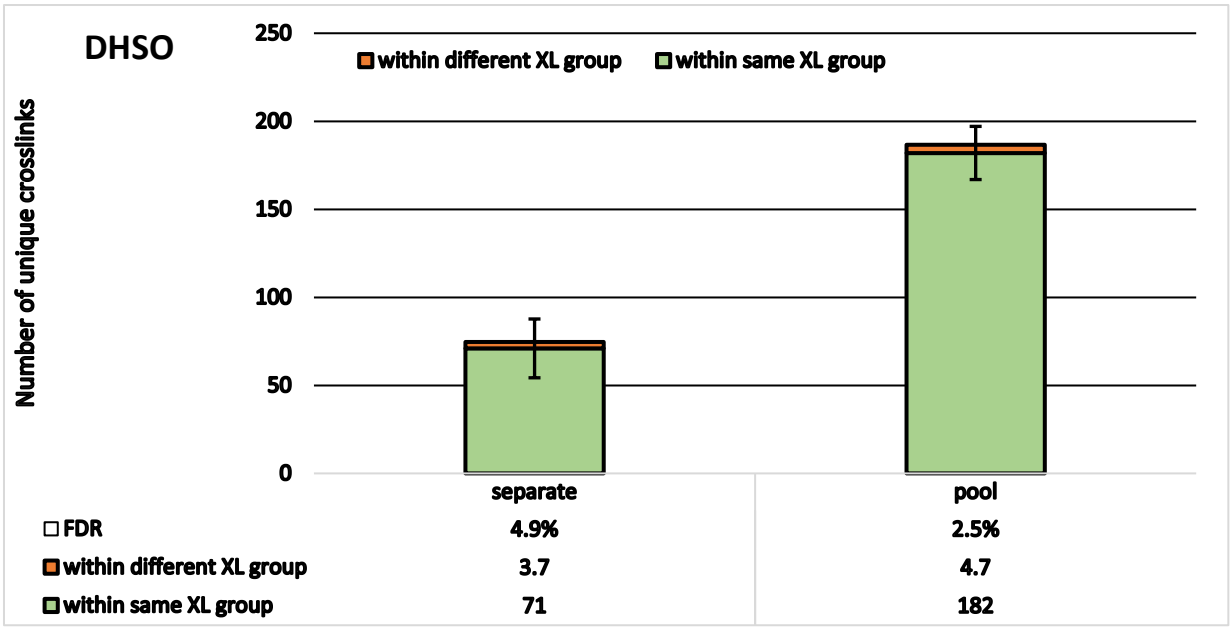


Figure 28: Average number of unique XLs of the acidic library linked in 2 hypostases: the conventional way where the groups are separately linked and when the group are mixed into one pool prior to crosslinking reaction. The data was acquired using stepped HCD MS2 and MS Annika integrated in Proteome Discoverer with an estimated FDR value at 1%. The links were searched against a database containing ribosomal proteins from E. Coli. The error bars rely on the standard deviations from the average values.

3.3. Crosslinking with different Crosslinkers and different XL-search engines

			total number				correct number				false number				FDR			
			MeroX	Annika	XlinkX	pLink	MeroX	Annika	XlinkX	pLink	MeroX	Annika	XlinkX	pLink	MeroX	Annika	XlinkX	pLink
main library	DSSO	replicate 1	638	581	463	420	593	563	447	400	45	18	16	20	0	3%	3%	5%
		replicate 2	612	599	510	385	580	583	482	373	32	16	28	12	0	3%	5%	3%
		replicate 3	724	715	605	564	688	697	579	541	36	18	26	23	0	3%	4%	4%
		average	658	632	526	456	620	614	503	438	38	17	23	18	0	3%	4%	4%
		standard deviation	59	73	72	95	59	72	68	90	7	1	6	6	-	-	-	
	DSBU	replicate 1	804	594	590	412	752	572	528	398	52	22	62	14	0	4%	11%	3%
		replicate 2	746	561	550	428	697	547	512	412	49	14	38	16	0	2%	7%	4%
		replicate 3	751	520	546	455	695	508	512	434	56	12	34	21	0	2%	6%	5%
		average	767	558	562	432	715	542	517	415	52	16	45	17	0	3%	8%	4%
		standard deviation	32	37	24	22	32	32	9	18	4	5	15	4	-	-	-	
	CDI	replicate 1	86	68	44	39	84	62	40	37	2	6	4	2	0	9%	9%	5%
		replicate 2	83	65	44	47	77	63	40	45	6	2	4	2	0	3%	9%	4%
		replicate 3	73	54	37	43	67	47	35	39	6	7	2	4	0	13%	5%	9%
		average	81	62	42	43	76	57	38	40	5	5	3	3	0	8%	8%	6%
		standard deviation	7	7	4	4	9	9	3	4	2	3	1	1	-	-	-	
enrichable sublibrary	DSSO	replicate 1	426	400	392	393	391	383	367	362	35	17	25	31	0	4%	6%	8%
		replicate 2	435	423	384	405	398	399	358	370	37	24	26	35	0	6%	7%	9%
		replicate 3	392	352	323	293	354	338	302	276	38	14	21	17	0	4%	7%	6%
		average	418	392	366	364	381	373	342	336	37	18	24	28	0	5%	7%	8%
		standard deviation	23	36	38	61	24	32	35	52	2	5	3	9	-	-	-	
	DSBSO	replicate 1	409	370	363	387	375	362	334	350	34	8	29	37	0	2%	8%	10%
		replicate 2	407	361	347	382	382	355	323	349	25	6	24	33	0	2%	7%	9%
		replicate 3	378	347	299	291	361	334	276	271	17	13	23	20	0	4%	8%	7%
		average	398	359	336	353	373	350	311	323	25	9	25	30	0	3%	8%	8%
		standard deviation	17	12	33	54	11	15	31	45	9	4	3	9	-	-	-	
acidic library	ADH, XL in separate groups	replicate 1	26		22	111	21		21	104	5		1	7	0		5%	6%
		replicate 2	36		18	88	29		17	82	7		1	6	0		6%	7%
		replicate 3	32		53	102	26		44	91	6		9	11	0		17%	11%
		average	31		31	100	25		27	92	6		4	8	0		12%	8%
		standard deviation	5		19	12	4		15	11	1		5	3	-	-	-	
	ADH, XL in pooled group	replicate 1	46		85	220	32		81	217	14		4	3	0		5%	1%
		replicate 2	37		2	226	31		0	224	6		2	2	0		100%	1%
		replicate 3	44		3	209	34		1	202	10		2	7	0		67%	3%
		average	42		30	218	32		27	214	10		3	4	0		9%	2%
		standard deviation	5		48	9	2		46	11	4		1	3	-	-	-	
	DHSO, XL in separate groups	replicate 1	76	94	89	68	74	90	74	59	2	4	15	9	0	4%	17%	13%
		replicate 2	92	65	79	68	86	62	66	63	6	3	13	5	0	5%	16%	7%
		replicate 3	71	65	65	76	67	61	57	61	4	4	8	15	0	6%	12%	20%
		average	80	75	78	71	76	71	66	61	4	4	12	10	0	5%	15%	14%
		standard deviation	11	17	12	5	10	16	9	2	2	1	4	5	-	-	-	
	DHSO, XL in pooled group	replicate 1	141	180	179	84	132	178	171	83	9	2	8	1	0	1%	4%	1%
		replicate 2	132	204	152	145	126	195	144	142	6	9	8	3	0	4%	5%	2%
		replicate 3	93	176	157	161	90	173	146	154	3	3	11	7	0	2%	7%	4%
		average	122	187	163	130	116	182	154	126	6	5	9	4	0	3%	6%	3%
		standard deviation	26	15	14	41	23	12	15	38	3	4	2	3	-	-	-	

3.4. Enrichment strategies

Apart from the measurements that can be successfully done and offering good results with DSSO and DSBSO on the enrichable peptide library without performing any additional enrichment step, there was the need of developing a functional workflow for the peptide libraries which unveils the power of the third functionality of the DSBSO linker. The procedure starts with the repeated addition of DSBSO in DMSO every 30 minutes (5 times in total). After the linking reaction is complete, every group is quenched with ABC to a final concentration of 100 mM. The rationality of the ABC-addition is to ensure that, at the moment of mixing, no XL reaction can take place anymore. The groups are then mixed and digested. At this point, the options were tried to reduce the sample complexity: size exclusion chromatography, affinity enrichment and size exclusion chromatography (SEC) coupled with affinity enrichment (AE). The reason why SEC was coupled with AE was to eliminate the perturbative effects of the DSBSO in excess and to test if two methods combined deliver even better results.

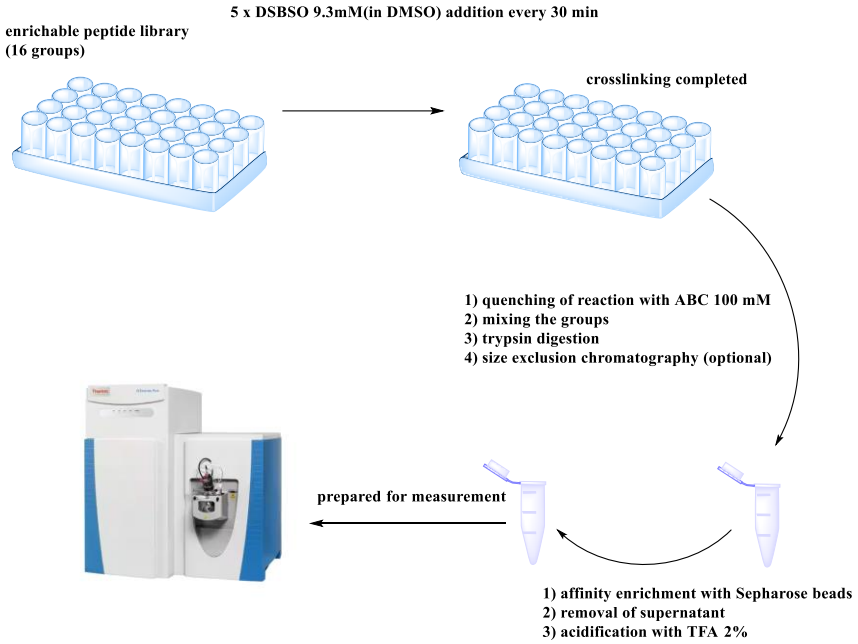


Figure 29: Enrichment workflow viable for the enrichable peptide library. Drawn in ChemDraw 19.1.The image of mass spectrometer was adapted from [62]

As the SEC result reveals, there are three peaks between minutes 12 and 16 with a low intensity, which are not exceeding 500 mAU. Being the first coming peaks, they must represent the crosslinks, which are the heaviest molecular species in the sample. They are followed by a relatively intense peak (~1300 mAU) between the minutes 16-17. At the end (denoting a light molecular species), a peak presents a high intensity of over 2300 mAU. By measuring DSBSO reagent alone without any sample and by the same SEC procedure, it could be concluded that the peak is produced by the linker: the position, intensity and shape of the peak are the same.

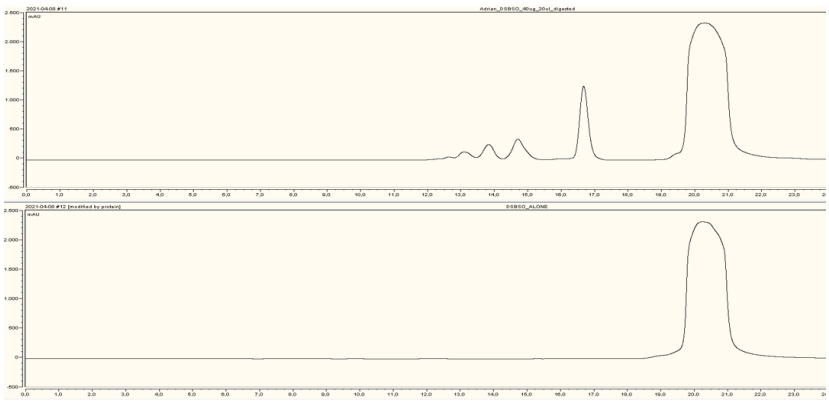


Figure 30: The upper part shows the chromatogram of the size exclusion chromatography realised on 20 µl (equivalent to 40 µg sample) of digested sample from the enrichable library- DSBSO experiment. The lower part of the figure represents a size exclusion chromatography run under the same conditions (flow rate etc.) of the crosslinker reagent (DSBSO) alone. On the x-axis there is the time represented in minutes in a 30 min run and on the y-axis there is the absorbance shown in mili-absorbance units.

In order to determine the nature of the peak produced at min 16-17, HPLC-MS/MS run was taken. The main peak at MS1 and MS2 level with a mass of 631.29m/z indubitably indicates the sequence Ac-WGGGGR, which has this molecular weight. This peak can be seen in every HPLC-MS/MS run with the peptide libraries and is a strong and necessary indicator for a succesful digestion.

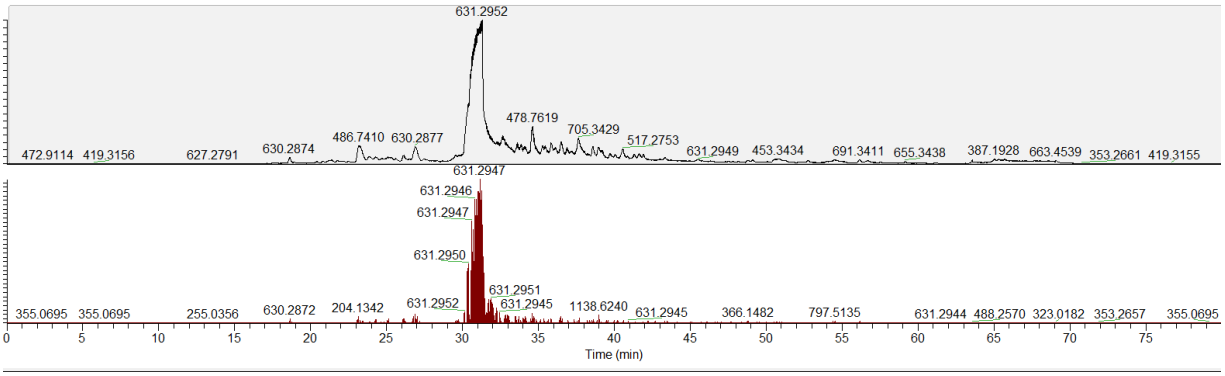
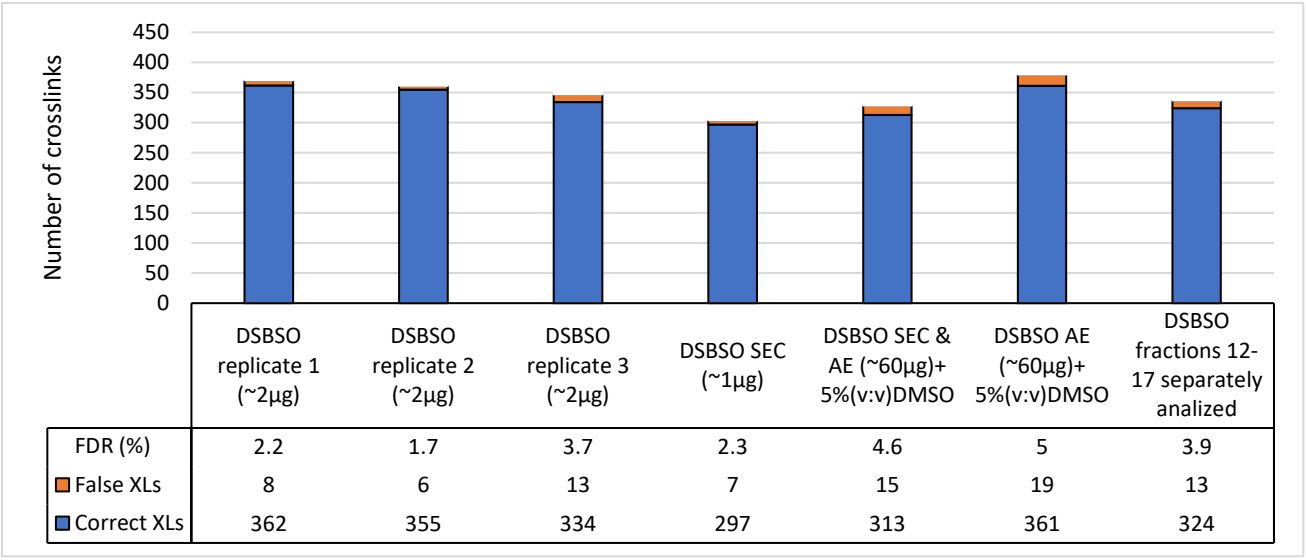


Figure 31: HPLC-MS/MS analysis of the fraction from the minute 16 to the minute 17 of the digested sample of enrichable library-DSBSO experiment. Upper part corresponds to the MS level and the lower part corresponds to the MS2 level.

The number of links obtained in all three replicates of the enrichable library (16 groups) with DSBSO does not differ much showing a high reproducibility. Almost 70% of all allowed links were achieved and the real FDR values stay relatively low. The SEC treatment, when the sample is not spiked, does not lead to better results, but quite on contrary to a slight decrease of the links found most probably due to unavoidable sample loss. The numbers do slightly improve when the fractions collected from the SEC are analysed separately, but the costs of five HPLC MS/MS runs of 140 minutes makes the incremental growth unworthy.

Affinity enrichment alone performs not only better than SEC, but also better than AE coupled with SEC. The concern over the enrichment of unreacted DSBSO in surplus (detected in SEC chromatogram) with the Sepharose beads proved to be unsubstantiated. AE presents 15% more XLs than SEC-AE treatment.

The Eppendorf tubes were “rinsed” with DMSO in the last step before measurement in the case of SEC-AE and AE, which has shown from experience a slight improvement in the numbers identified.



	DSBSO replicate 1 (~2µg)	DSBSO replicate 2 (~2µg)	DSBSO replicate 3 (~2µg)	DSBSO SEC (~1µg)	DSBSO SEC & AE (~60µg)+ 5%(v:v)DMSO	DSBSO AE (~60µg)+ 5%(v:v)DMSO	DSBSO fractions 12-17 separately analized
FDR (%)	2.2	1.7	3.7	2.3	4.6	5	3.9
False XLs	8	6	13	7	15	19	13
Correct XLs	362	355	334	297	313	361	324

Figure 32:Number of crosslinks obtained from the enrichable library (16 groups) linked with DSBSO and consequently purified through different enrichment strategies without any HEK spiking prior to enrichment procedures. The samples were analysed with the MS Annika at an FDR of 1% against a database of ribosomal protein from E. coli identified in a shotgun analysis run. Crosslinks within the same group are represented in blue, crosslinks from different groups are shown in orange.

Most of the time there is a lot of qualitative information present in the chromatogram of the run even before any kind of bioinformatic analysis is being executed. On the MS1 level, the chromatograms from AE and SEC-AE exhibit almost identical fingerprint profiles. With a fairly different fingerprint profile, the same sample where no enrichment treatment was done, the typical signal for Ac-WGGGGR (631,28 m/z) is dominant.

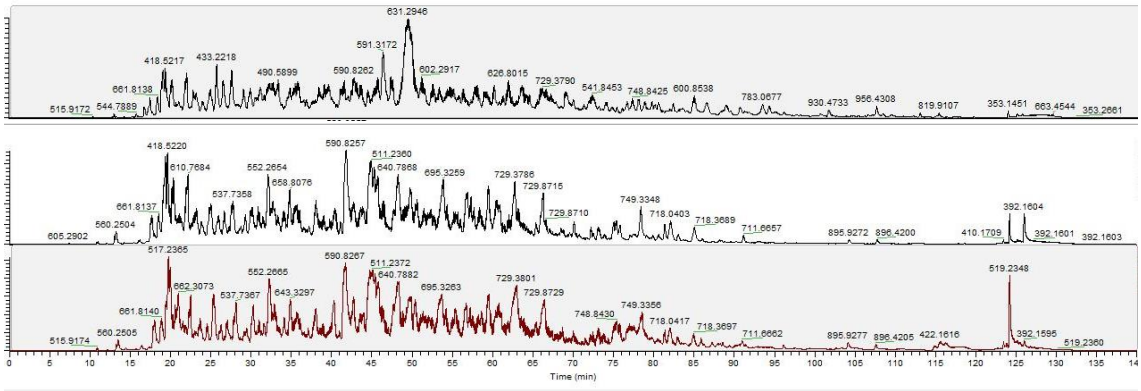


Figure 33: the chromatograms in a 140 min HPLC-MS/MS run on the MS1 level from a) the upper image shows the enrichable library-DSBSO experiment without any enrichment procedure b)the middle image represents the enrichable library-DSBSO experiment after an affinity enrichment step c) the lower images shows the results from enrichable library-DSBSO after a size-exclusion-chromatography coupled with affinity enrichment

An enrichment treatment has a better justification in the case of a sample with a higher complexity. Additional to the carried measurements, the sample was spiked in two different mass ratios: 1:10 and 1:100. The spiked solutions were analysed before and after the enrichment. Expectedly, the 1:10 sample shows a loss of 49% in the found crosslinks while the 1:100 spiked sample shows a 97% loss. Interestingly, the real FDR is 0%. After the enrichment, both spiked samples deliver even better results than the normal replicates (not spiked and not enriched). Notably, neither the quality nor the quantity of the results differ in the 1:10, respectively 1:100 enriched sample proving the utility of this extra step added in the workflow when the complexity of the mix requires it.

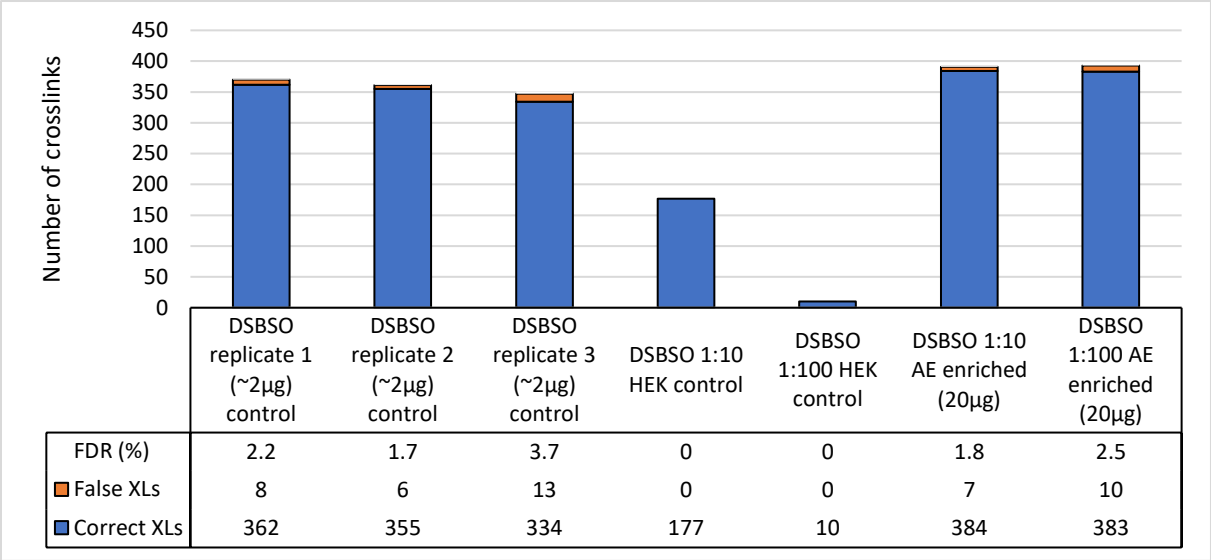


Figure 34: Enrichable library linked with DSBSO, spiked with peptides from HEK lysate in different proportions and additionally affinity enriched. The samples were analysed with the MS Annika at an FDR of 1% against a database of ribosomal protein from E. coli identified in a shotgun analysis run. Crosslinks within the same group are represented in blue, crosslinks from different groups are shown in orange.

3.5. Adaptation of crosslinking protocol from Leitner et al.

In the recent years there has been a lot of effort from Leitner et al. to cover further the loops in the crosslinking field through carboxyl-group specific crosslinking. In such a study, the authors have fine-tuned the number of links one can achieve with the help of a hydrazide linker by systematic ratio-variation of the PDH and DMTMM. It was empirically shown that a 2-4 times molar excess of the DMTMM is favouring the incorporation of the dihydrazide over the zero-length link formation.[63]

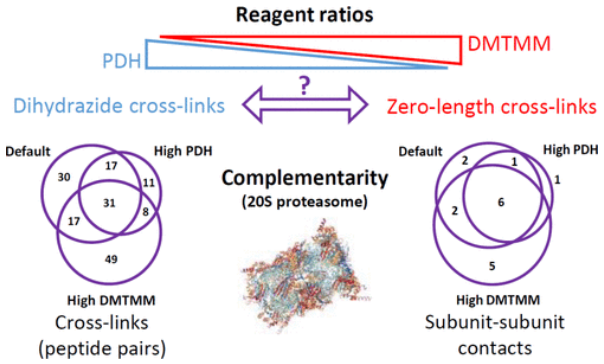


Figure 35: reagent ratios variation to maximize both dihydrazide crosslinks and zero-length crosslinks or so-called "DMTMM crosslinks". Figure adapted from [63]

That was the first time in the Mechtler group that an acidic targeting linker was used. The method had to be firstly tested and adapted subsequently for the use in context of acidic peptide library.

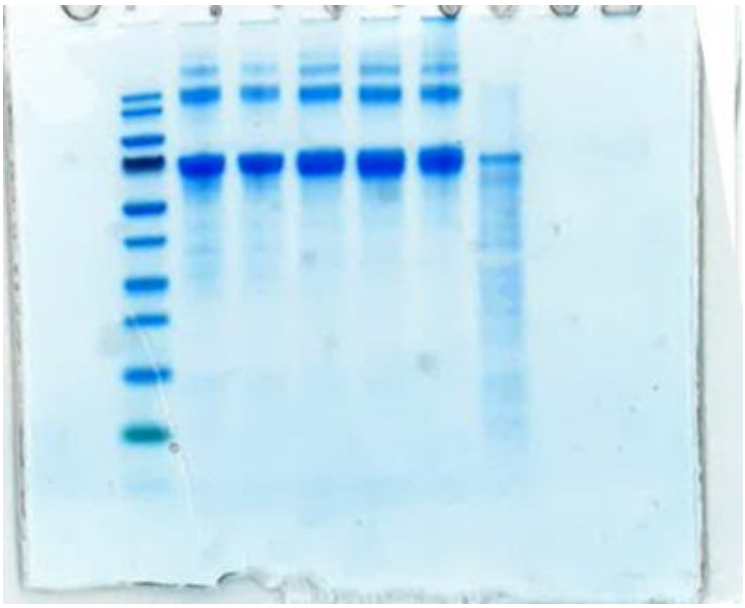


Figure 36: Gel electrophoresis; PageRuler prestained Protein Ladder stored at -20°C with running settings of 150 V max, and 35 mA for 60 min. The gel was then rinsed 3 times with double distilled water and stained overnight using MBSBlue + 2% NaCl supplemented for the overnight incubation)

1)	Protein ladder
2)	BSA A) (1,44mg/ml DMTMM; 1mg/ml ADH)
3)	BSA B) (4,316mg/ml DMTMM; 3 mg/ml ADH)
4)	BSA C) (8,636mg/ml DMTMM; 6 mg/ml ADH
5)	BSA D) (12mg/ml DMTMM; 8,3mg/ml ADH)
6)	BSA E) (14,4mg/ml DMTMM; 10mg/ml ADH)
7)	BSA pure

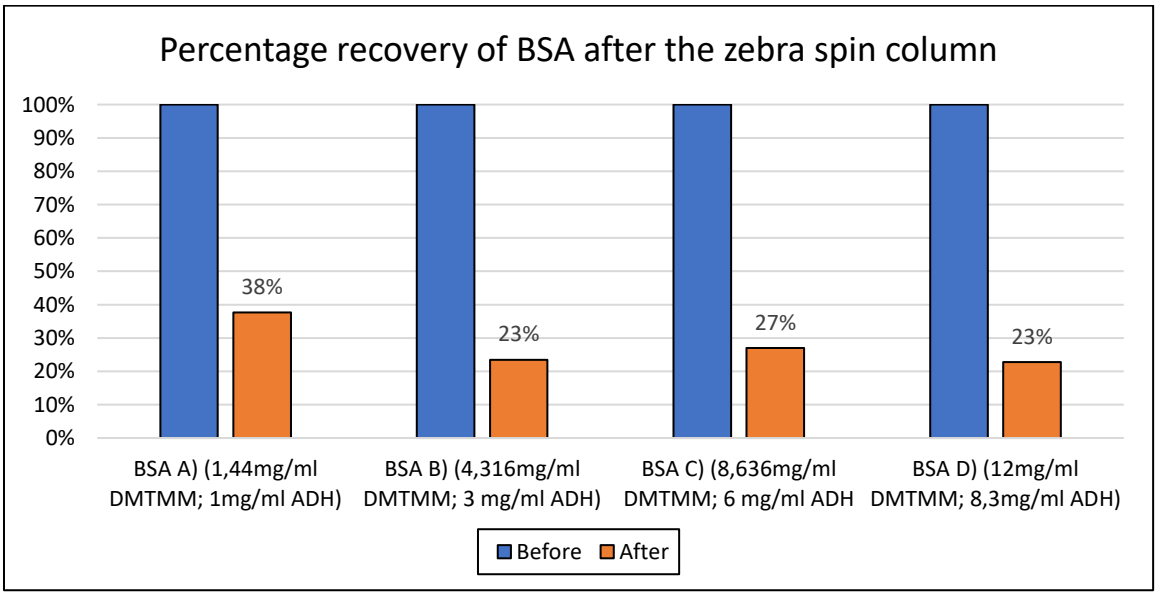


*Figure 37: Percentage of recovery of the BSA protein. The same volume was measured before and after the desalting column in a HPLC with a UV-VIS detector with a 214 nm channel. The protein total recovery was calculated as ratio between the total areas measured before and after in mAU*min (mili-absorbance units*minute).*

In the gel electrophoresis, it is visible that the experimental conditions from “BSA E)” are not feasible, as clear signs of over-crosslinking are shown. There is a relatively preeminent band, which did not migrate at all suggesting a high order aggregate has formed and many BSA proteins have been bound together. The pocket filled only with BSA (the 7th pocket from left to right) is exhibiting signs of smearing likely due to either protein fragmentation or insufficient purification. The BSA was kept in all samples to a concentration of 2 mg/ml.

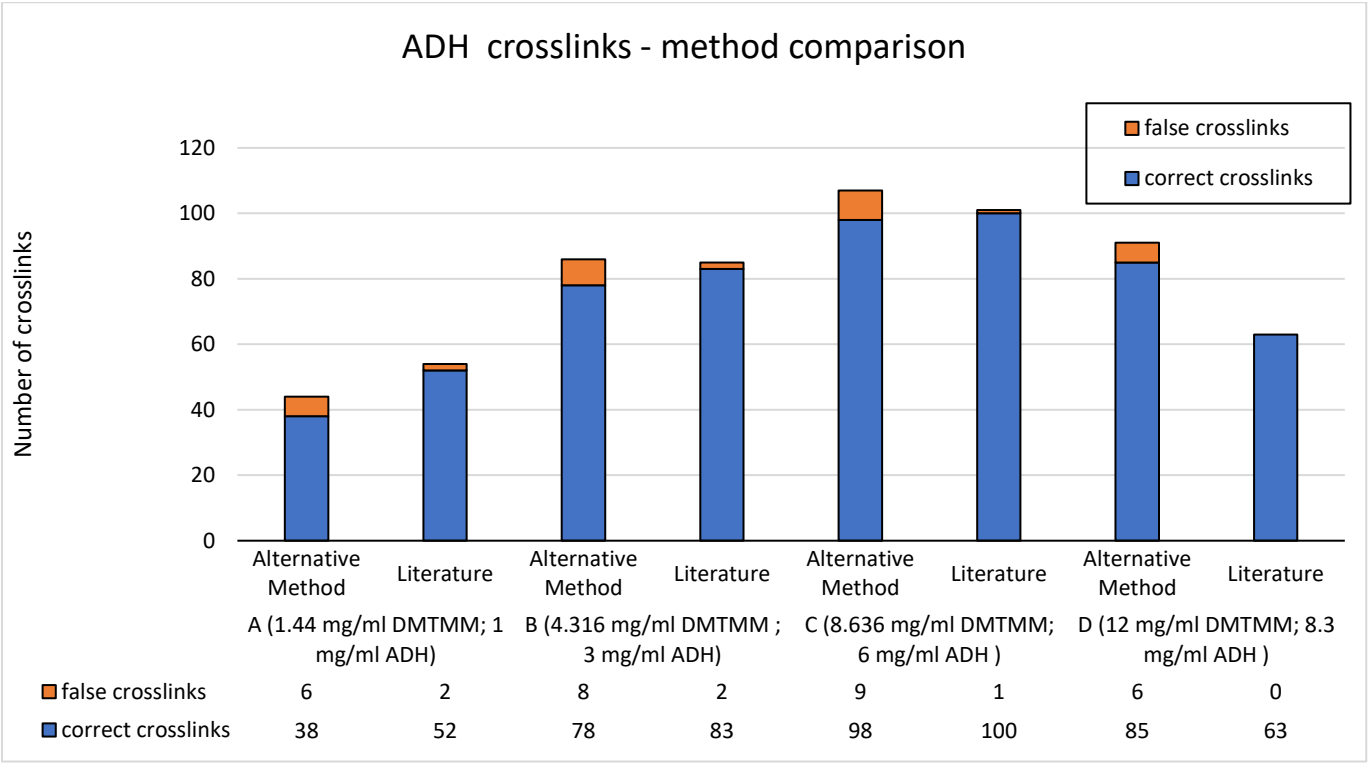
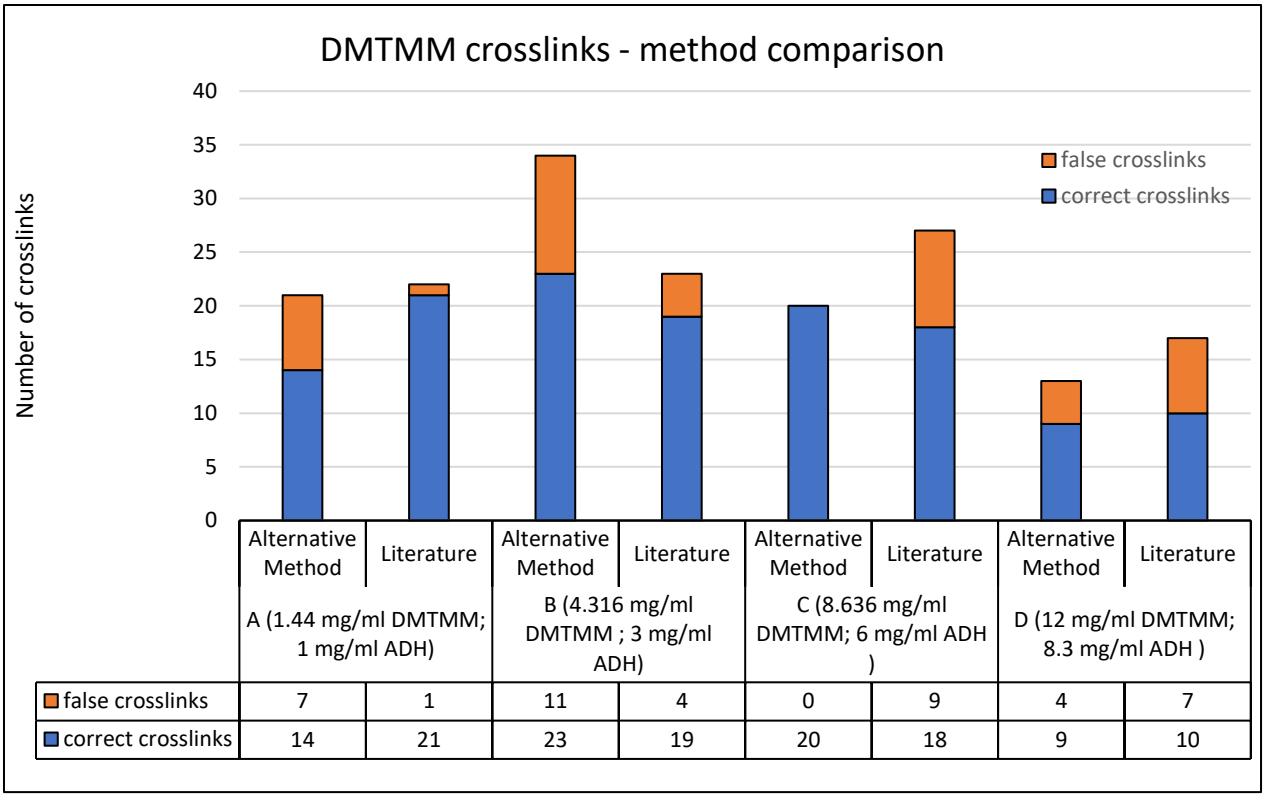


Figure 38:Number of ADH crosslinks depending on the Literature vs Alternative method which skips the Zebra spin desalting column; The concentration of BSA was kept constant and different crosslinker concentration were being utilised. The analysis was done with Merox at an FDR level of 5% against a fasta file with contained BSA and contaminants. Of note: the chemical modification M->m (from iodoacetamide) was not considered.



	Alternative Method	Literature	Alternative Method	Literature	Alternative Method	Literature	Alternative Method	Literature
	A (1.44 mg/ml DMTMM; 1 mg/ml ADH)		B (4.316 mg/ml DMTMM ; 3 mg/ml ADH)		C (8.636 mg/ml DMTMM; 6 mg/ml ADH)		D (12 mg/ml DMTMM; 8.3 mg/ml ADH)	
false crosslinks	7	1	11	4	0	9	4	7
correct crosslinks	14	21	23	19	20	18	9	10

Figure 39: Number of DMTMM crosslinks depending on the Literature vs Alternative method which skips the Zebra spin desalting column; The concentration of BSA was kept constant and different crosslinker concentration were being utilised. The analysis was done with Merox at an FDR level of 5% against a fasta file with contained BSA and contaminants. Of note: the chemical modification M->m (from iodoacetamide) was not considered.

The expressions “false crosslinks” and “true crosslinks” in the case of BSA are just indicatively used and to simplify things. The crosslinks between contaminant proteins with themselves or with BSA are in fact wrong, but there is no way of knowing if the intralinks of BSA are true as in the peptide library. A more reality-reflective phrasing would be “minimum false crosslinks” and “maximum correct crosslinks”.

While it is well understood why the Zebra spin desalting column would be beneficial (removing salt is gentle and protective for the MS machinery), the results were not strongly influenced by removal of this step (neither the XL number nor the FDR). Not only that, but the protein loss that the experimenter can expect is up to 75%, which for small amounts of sample can be very problematic.

In the case of hydrazide-based links, the number of XLs increases until the molar ratio between BSA/DMTMM/ADH of 166.25:1:1, where nearly 100 crosslinks are detected (a surprisingly good number for BSA). It must be mentioned that this ratio is in great excess compared to a normal crosslinking protocol of a protein, but necessary because of the lower yields. The approximate FDR values (calculated based on the inexistant contaminants) are all under 10%, which is acceptable for a set value of 5%.

In the case of zero-length links, the number of obtained links does not appear to be correlated to the different concentrations utilized nor to the method utilized. The range of found links varies from 13 to 34 (relatively low numbers), but the FDR varies strongly from 0% to 33% making the results unreliable.

3.6. GroEL Protein. Crosslinking Mass Spectrometry in a lipidic medium

In a very eye-catching workflow presented in a study[64] a new approach of XL-MS is being described, where the authors are performing in gel crosslinking within a blue native polyacrylamide. This was shown to reduce non-specific crosslinks otherwise artificially formed within in solution XL-Ms workflows. The native PAGE keeps the secondary, tertiary, and quaternary structure intact and does not alter the structure of the proteins/ protein complexes. In the following step, the bands of interests are cut, depending on which complex is of interest, and moved to an Eppendorf tube where a crosslinking buffer and crosslinker is being added. After the reaction, digestion is being performed and the resulted mix is measured through the usual HPLC-MS/MS procedure. Through this workflow, distance constraints are introduced, the sample complexity can be dramatically reduced (allowing the research to concentrate only on the proteins of interest) and the overlength (possibly random or false crosslinks) are being hampered.

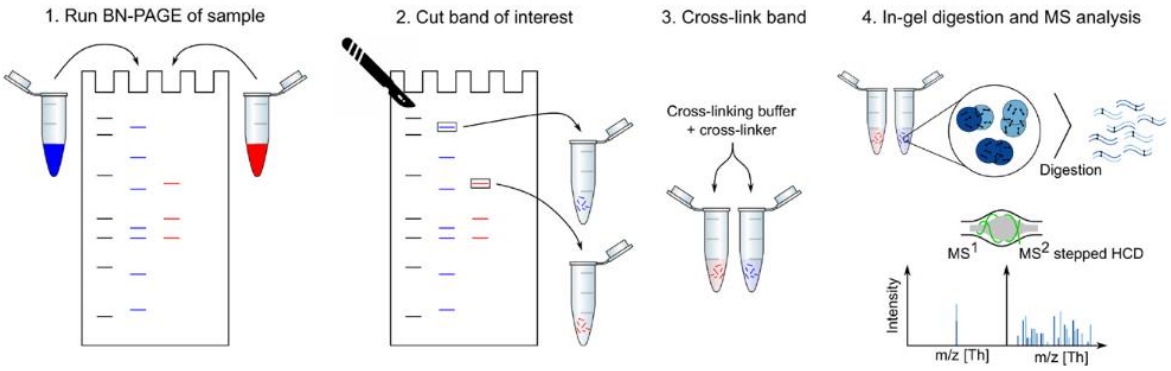


Figure 40: Schematic representation of the IGX-MS workflow. Figure adapted from "Selective cross-linking of coinciding protein assemblies by in-gel cross-linking mass spectrometry"[65]

One of the issues tackled is important: the hindrance of non-specific overlength crosslinks. The proteins or protein assemblies can form higher-order aggregates in solution which generate non-representative crosslink interactions. In the gel the proteins cannot roam freely as in the solution. The authors also show that, while the traditional in-solution XL-experiment with GroEL and DSS linker resulted in more crosslinks than the in-gel XL variant, virtually all extra links discovered were having very high Cα-Cα distances between the bounded lysines residues (above 25Å).[64]

Before presenting our new approach, some special parameters of the proteins in the solution must be discussed. The minimum radius a protein can have is dependent on its molecular weight (assuming it would take a spherical form, so the values are rather conservative). These values are represented in the table below. However, the average distance between two proteins is not only dependent on its size, but also on its concentration in the solution. The values of the average distance between proteins are provided below.

Protein M (kDa)	5	10	20	50	100	200	500
R_{min} (nm)	1.1	1.42	1.78	2.4	3.05	3.84	5.21

Figure 41: Rmin for proteins of different mass. Table adapted from [66]

Concentration	1M	1mM	1µM	1nm
Distance between molecules (nm)	1.18	11.8	118	1180

Figure 42: Distance between molecules as function of concentration. Table adapted from [66]

Having these values at disposal, it is easy to conclude that random interactions do occur and proteins in a solution, through their movement, will be in each other's proximity without having a biologically relevant interaction. Most notably, if two monomers of the same protein come close and are then bounded by the linker, the resulting crosslink will be interpreted by the XL-engines as an intralink, resulting in apparent overlength XLs of high spectrum quality (since they are really existing, but artificially formed), even though it is an intermolecular-link.

The very presence of crosslinks in the synthetical peptide libraries is an indubitable argument that peptides and hence likely also proteins regularly come in each other's proximity without the need of any biologically relevant interaction.

To overcome these impediments, we proposed a mechanistic model which involves the formation of micelles prior to the crosslinking. This would have several significant advantages:

- The proteins and protein assemblies would get isolated from each other by being caged in a lipidic layer. The lipidic should be permeable for the crosslinker but increase the minimum achievable distance between two different biomolecules.
- Structural analysis of membrane proteins could become accessible through an XL-MS protocol. The scientific community still has problems to elucidate biologically relevant conformation (to better understand the function and reaction mechanism) of a multitude of membrane proteins through conventional methods. By providing a lipidic medium, they would not only become soluble in solution, but their conformations would not be affected.
- Compared to IGX-MS, samples of higher complexity could be simultaneously studied without the need of cutting only the band of interest.
- Just limited information can be offered with the normal techniques. Crystallization for X-ray is very difficult, and the actual structure is determined by the membrane. For the NMR these proteins are usually too large, and the membrane must be taken into consideration as well. Through the electron microscopy the position of helices can be observed, but the resolution is too low. [67,68]

The concept is presented in the figure below and it is composed of 4 main steps. In the first one, the micelles (or protein cage) would form. Emulsion can occur. Additionally, the crosslinking reaction can take place and it is followed by the removal of the lipids (alternative: labialization of the polymer and accompanied by size exclusion) through size-exclusion. Afterwards, the normal crosslinking steps which are composed of digestion, HLPC-MS/MS run, and raw data processing can take place.

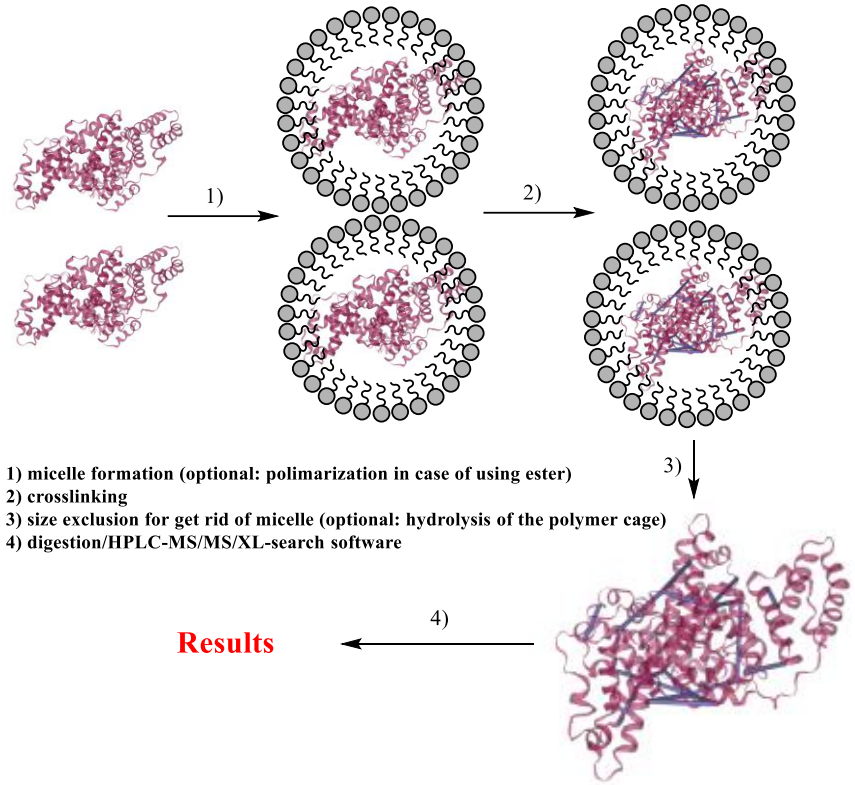


Figure 43: Schematic representation of protein encapsulation followed by XL-MS procedure.

In the initial phase, we wanted to create a non-ionic detergent, which has non-denaturating properties, low critical micelle concentration (CMC – above this concentration the detergent molecules self-assembly into micelle structures)[69–71] and relatively low aggregation number (property of a micelle which defines the number of monomers present in its structure)[69,70,72]. As hydrophilic head, glycerol dimethacrylate (a mixture of 1,2- and 1,3- form stabilized with MEHQ) was chosen. The decanoic acid would be esterified with the free hydroxy group to form not only a detergent capable of forming micelle like structures, but also a monomer that could in situ polymerize around any protein to cage it.

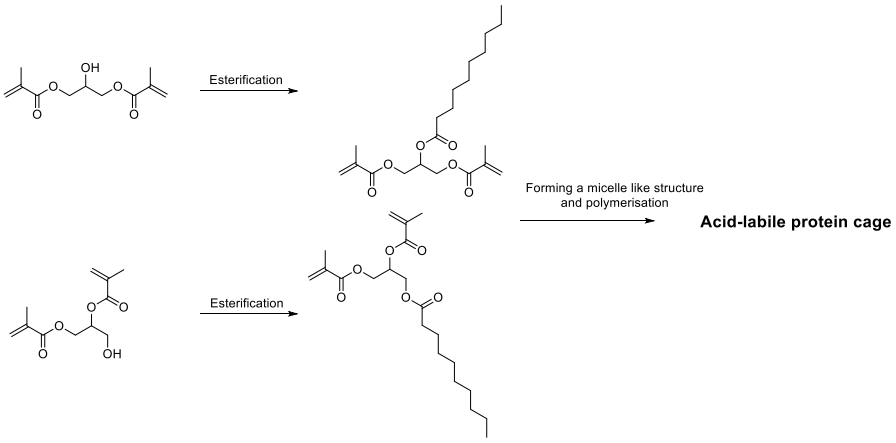


Figure 44: Initial plan, not accomplished of detergent-monomer formation.

Before approaching to carry out the esterification, stability tests on glycerol dimetacrylate were realised with the help of thin layer chromatography. Because the results were not assessable, a MALDI analysis was then employed.

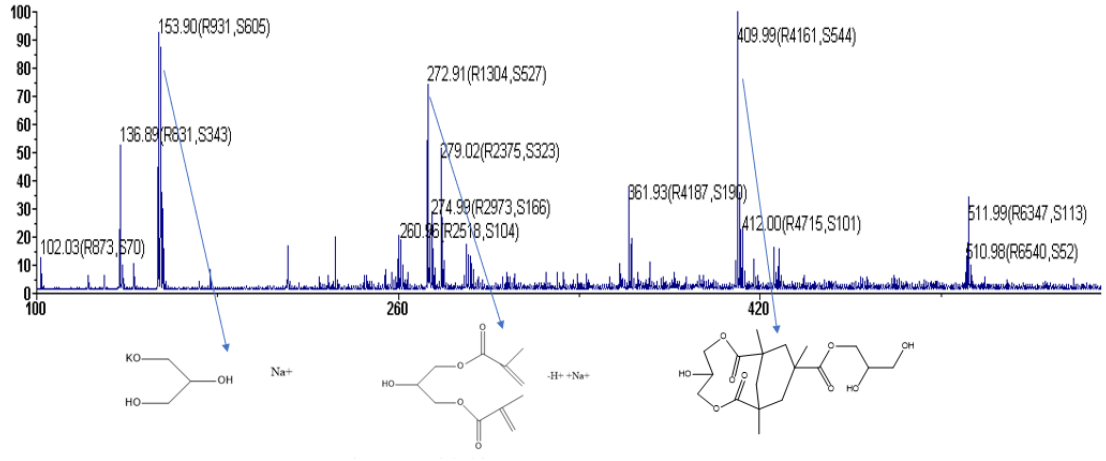


Figure 45: MALDI analysis on alcohol stability in CH_2Cl_2 after 135 min; on the x-axis there is m/z represented and, on the y-axis, the relative intensity of the signals. There are numerous signals to be seen, of which some correspond to the molecular mass of glycerol accompanied by potassium and natrium ions, 2-hydroxypropane-1,3-diyl bis(2-methylacrylate) with a hydrogen ion interchanged by a natrium ion and 2,3-dihydroxypropyl 5-hydroxy-1,9,11-trimethyl-2,8-dioxo-3,7-dioxabicyclo[7.3.1]tridecane-11-carboxylate where a hydrogen atom was interchange with a natrium atom. The structures were not confirmed with H-NMR.

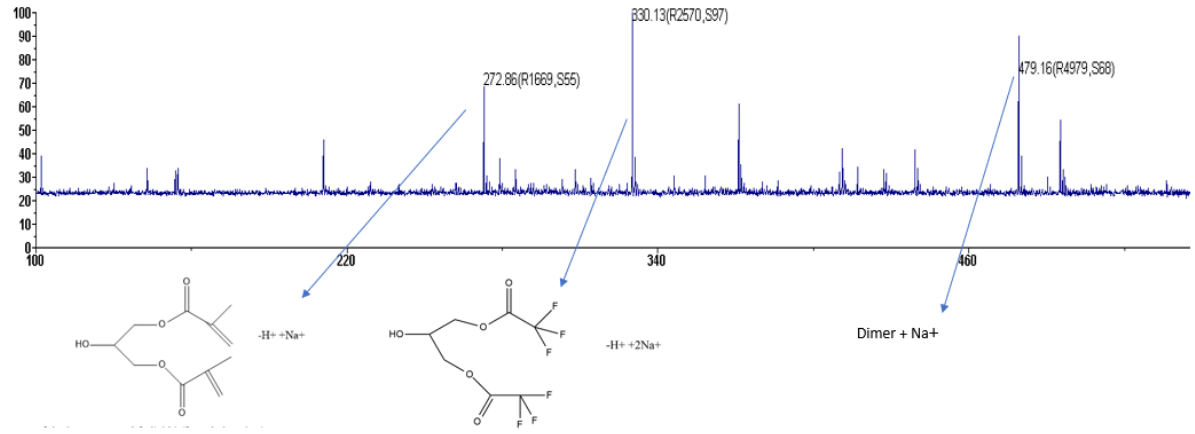


Figure 46: MALDI analysis on alcohol stability in a TFA 1% solution and CH_2Cl_2; on the x-axis the m/z is represented and on the y-axis the relative intensity of the signals. There are numerous signals to be seen, of which some correspond to the molecular mass of glycerol dimetacrylate, 2-hydroxypropane-1,3-diyl bis(2,2,2-trifluoroacetate) $-H^+ +2Na^+$ and a glycerol dimetacrylate dimer $+Na^+$. The structures were not confirmed with H-NMR.

Strong evidence of the instability of the glycerol dimetacrylate in dichloromethane was found. Not only the presence of signals which indicate the existence of glycerol, but also signals from compounds of higher molecular weight suggest a broader chemical landscape. Additionally, in the MALDI spectrum of the glycerol dimetacrylate, after contact with trifluoroacetic acid, a multitude of unexpected peaks are found. Having acquired this new information, it was decided to renounce at the esterification reaction and the synthesis of detergent monomer. Instead of it, decanoic acid and dodecanoic acid were tested alone.

Due to financial reasons, the methods were firstly tested and improved on BSA (Bovine Serum Albumin) as test protein and subsequently on Chaperonin GroEL from Escherichia Coli.

ID	A	B	C	D	E	F1	F2
Protein mass	BSA(1,5ug)	BSA(1,5ug)	BSA(1,5ug)	BSA(1,5ug)	BSA(1,5ug)	BSA(1,5ug)	BSA(3,0ug)
Linker	DSSO	DSSO	DSSO	DSSO	DSSO	No linker	No linker
Fatty acids	No fatty acids	x4 (m:m) decanoic acid	x5 (m:m) decanoic acid	x4 (m:m) decanoic acid	x5 (m:m) decanoic acid	No fatty acids	No fatty acids
Salt concentration of the solution	No salt	No salt	No salt	0.135M NaCl	0.135M NaCl	No salt	No salt

In the first try, only decanoic acid was utilized to build the micelles; additionally, the effect of natrium chloride was analysed in this context. After the crosslinking reaction took place and the salts and fatty acids were removed through size exclusion procedure, a gel electrophoresis was carried out.

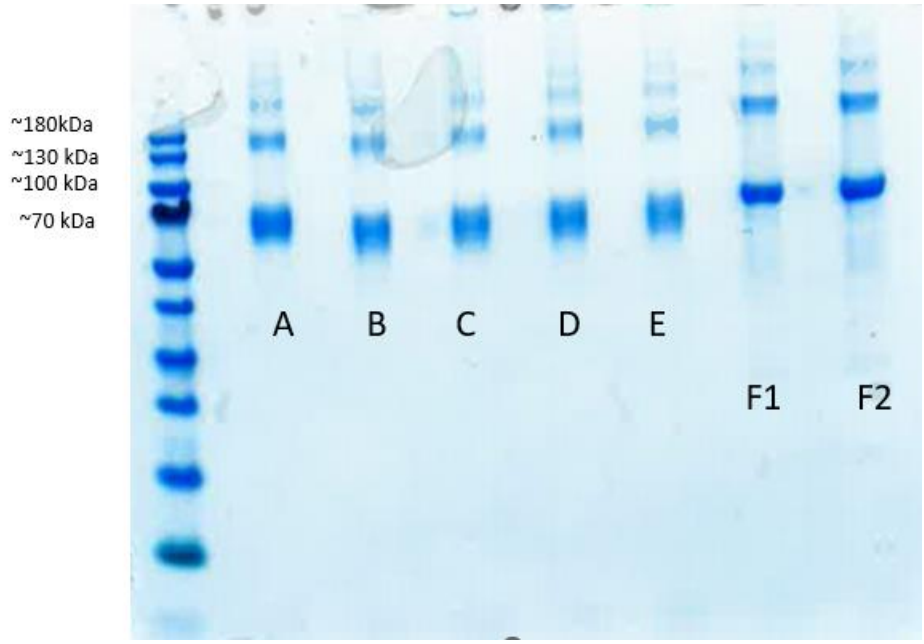


Figure 47: Gel Electrophoresis on the first BSA-crosslinking experiment with DSSO in the presence of fatty acids. Electrophoresis was realised on a polyacrylamide gel and was then stained Coomassie Blue for the visualization of the protein bands. The first lane on the left side represents a commercially available protein ladder. The respective approximate molecular weights for landmark points of interests are on the left side.

The BSA has an approximate molecular weight of 66.5 kDa. One can see in the non-crosslinked samples of BSA (F1 and F2) that higher order molecular aggregates are being formed. There are signals above 130kDa and higher, so one can assume that BSA dimers and trimers could self-assembly. Also, the band intensity of BSA polymers is higher in the case of F1 and F2 replicates than in the rest of the bands. The spots with low migration distances from the replicate where no fatty acids were involved appear to be slightly more intense than in the samples where decanoic acid was employed. From experience, one would normally expect approximately 80-100 crosslinks in the BSA protein when DSSO is used as linker. In the above-described experiment, a very low number of links was detected.

ID	A	B	C	D	E
Nr. of XLs	8	22	30	21	27

The XL-search engine was Merox, used in the Rise mode with an FDR value of 5% where K-K link were targeted and a fasta file of BSA with contaminants was utilised as search data base. The obtained number of links was strongly below expectations. Apart from this, the replicates where fatty acids were employed produced better outcomes. The experiment was then repeated, and the expected number of links was obtained.

For the GroEL protein, it was decided to quit salty conditions (0.135 M NaCl) and benchmark instead decanoic and dodecanoic acid in two different mass ratios each. The individual steps are explained in detail in the methods chapter. In the table below there is a list of the tested replicates in the GroEL experiment with the corresponding technical parameters.

ID	Description	Total Unique XLs	TIC MS1	XLs score cut-off = 50	XLs score cut-off = 50 and d(Cα-Cα)≤35Å
A	**5µg GroEL**	**112**	**9.52E9**	**90**	**62**
B	**5µg GroEL+10µg decanoic acid**	**150**	**8.49E9**	**116**	**83**
C	**5µg GroEL+20µg decanoic acid**	**207**	**9.39E9**	**170**	**109**
D	**5µg GroEL+10µg dodecanoic acid**	**237**	**1.12E10**	**206**	**136**
E	**5µg GroEL+20µg dodecanoic acid**	**208**	**8.45E9**	**183**	**117**

The analysis was done in Merox by using the Rise mode. Only K-K crosslinks were counted, and the FDR was set to 5%. DSSO was used as a crosslinker, and the crosslinks were searched against a data base containing the sequence for GroEL and contaminant proteins. The results obtained by Merox were saved in csv files, which were then uploaded in xiVIEW[73]. The data was then further processed by setting a score cut-off of minimum 50 and a maximum distance between the Cα atoms of the lysine residues of 35 Å to ensure the high quality of the obtained crosslinks. The determination of the distance between the linked residues was possible by importing the crystal structure of (GroEL-KMgATP)14 at 2 Å.[74]

TIC (short for total ion current) in a chromatogram refers to the total ion current integrated over the retention time. This term is widely used in chromatographic method coupled with mass spectrometry[75] and gives a sense of how much (quantitatively) material is in the analysed sample. The ions have different charges so it cannot be seen as strictly directly proportional to the injected mass. In the 5 analysed cases, the TIC values do not vary much and are not correlated to the number of crosslinks found.

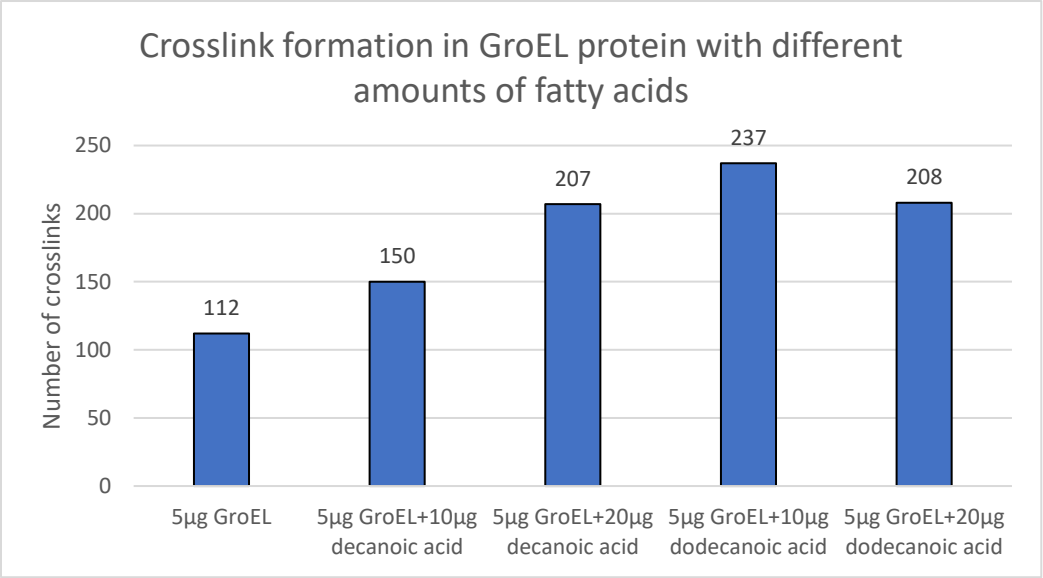


Figure 48: Bar chart of the crosslinks obtained by analysing crosslinking experiments under two different fatty acids in different ratios compared to the fatty acids' free medium.

For the same order of magnitude regarding the TIC, one obtains significantly better results in the number of links detected when decanoic acid and dodecanoic acid are being added to the sample. Not only that, but dodecanoic acid seems to have an even bigger positive influence on the crosslink formation than decanoic acid.

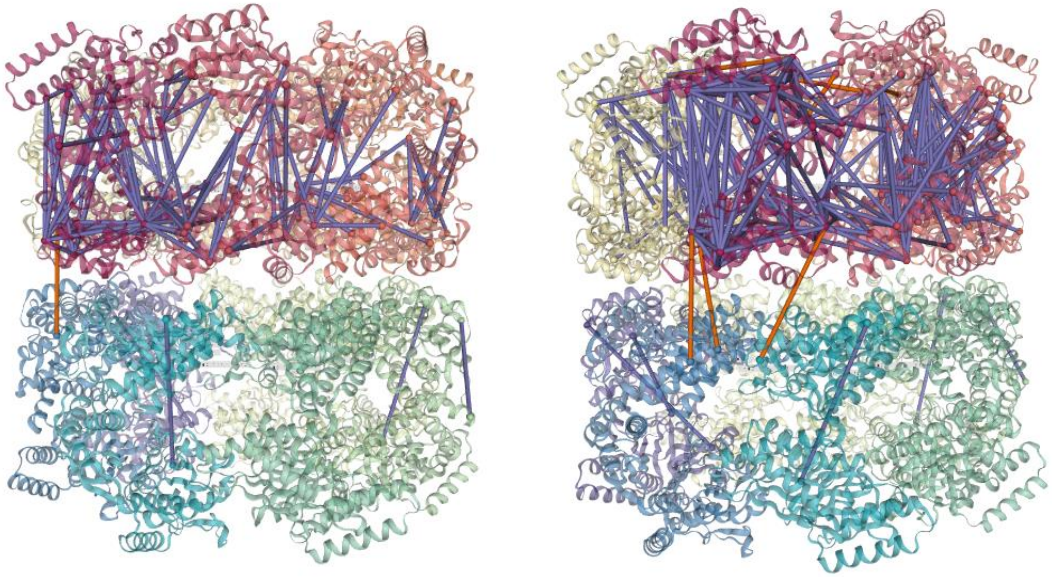


Figure 49: Quaternary structure of the GroEL 14-mer and the crosslinking topologies displayed in the crystal structures of the protein. On the left side of the image – ID = A (no fatty acids); on the right side of the image - ID = D (5 µg GroEL + 10 µg dodecanoic acid). The pdb file (1KP8) from the Structural basis for GroEL-assisted protein folding from the crystal structure of (GroEL-KMgATP)14 at 2.0 Å was loaded into xiVIEW. The structure covers approximately 525 out of 548 amino acids (96%).

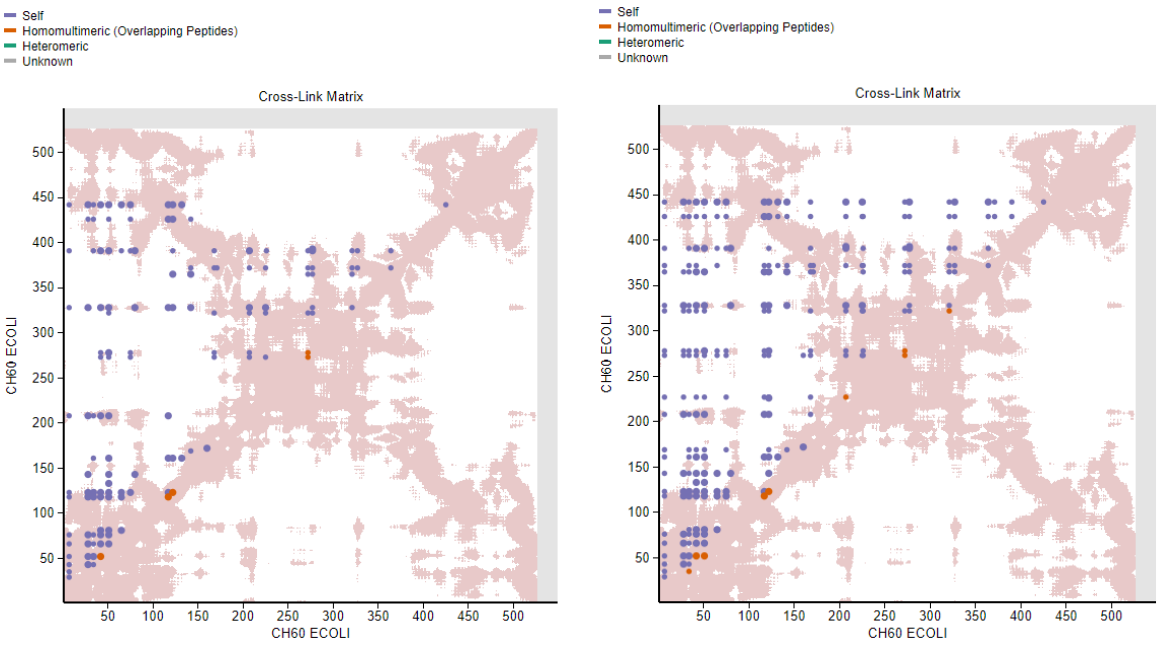


Figure 50: Crosslink Matrix comparison of two different experiment conditions. On the left ID=A where no fatty acids were added; on the right side of the image 10 µg dodecanoic acid were added (ID=D).

By comparing the crosslink matrices, it is visible that under normal experimental conditions, where no fatty acids are present, some regions are being negatively discriminated and no crosslinks are observable. In contrast, the high number of crosslinks acquired is uniformly distributed. This result can be determined by an enhanced reactivity of the crosslinker or an enhanced GroEL flexibility conferred by the fatty acids. The presence of an increased number of homomultimeric crosslinks in the fatty acids conditions is not necessarily dismantling the hypothesis of protein encapsulation/ micelle formation because the protein naturally forms a 14-mer.

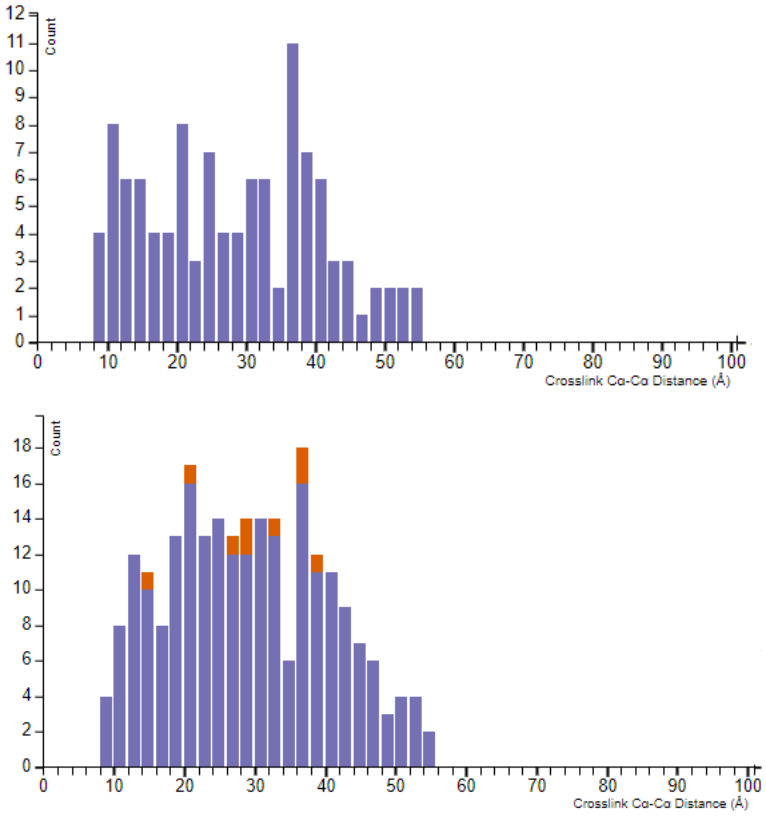


Figure 51: Crosslink Count dependent on the crosslink Cα- Cα measured in Å. The violet colour denotes intralinks and orange represents homomultimeric links (overlapping peptides). The upper part of the figure shows the results from the GroEL protein linked with DSSO without any fatty acid medium (experimental conditions from "A"). The lower part of the figure shows the distribution from the 5µg GroEL + 10 µg dodecanoic acid-replicate (experimental conditions from D)

In the figure above, where conditions A and D are compared, more XLs are found but the encapsulation hypothesis can not be confirmed as also longer links can be detected as well.

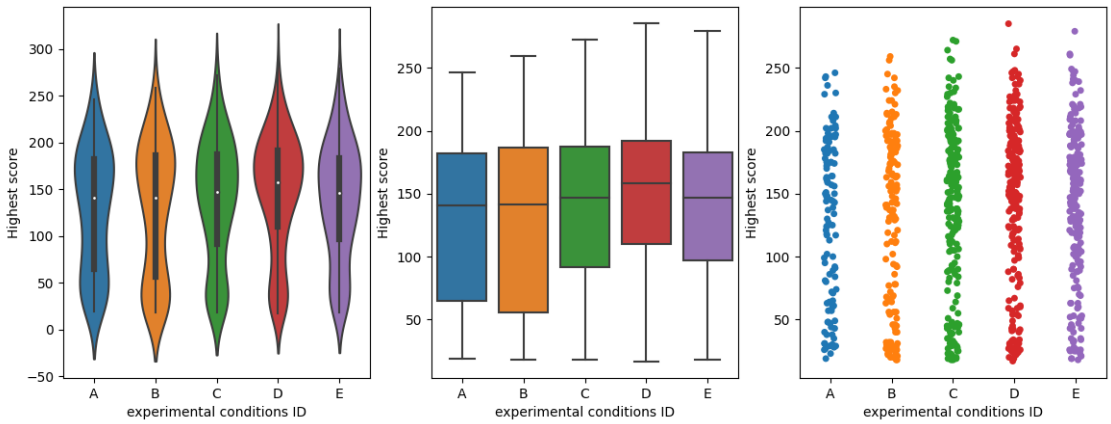


Figure 52: Violin plot, boxplot and strip plot of the highest CSM score obtained per crosslink.

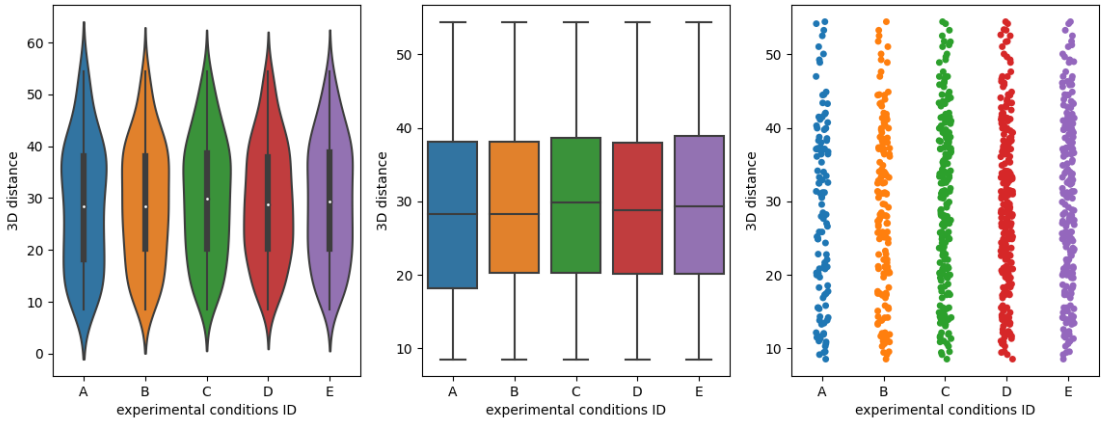


Figure 53: Violin plot, boxplot, and strip plot of the 3D distance between Cα-Cα measured in Å.

Multiple plot types of the same data are represented above. Violin plots display the distribution of data, while boxplot confer a perspective when it comes to the general tendencies of data. At last, the strip plot emphasizes on specific patterns and single points such as possible outliers. The median score of the obtained crosslinks in a lipidic medium either stays approximately the same or even improves. There are also no significant changes regarding the median tridimensional distance of the linked residues. It can be concluded that, for this data set, not only the quality but also the quantity of the results become better by adding fatty acids. The best results by far are obtained in the experimental conditions from “D", where the sample and a double amount in mass of dodecanoic acid is tested.

4. Conclusion and Outlooks

A synthetic peptide library is an important tool in scientific research and comprises a diverse mixture of peptide sequences that offer a unique opportunity to study and comprehend physicochemical parameters, chemical mechanisms, and their respective reactivities. Unlike the analysis of individual proteins or peptides or relying just on computational methods that may not properly mirror the complexity of the biological world, peptide libraries enable a more comprehensive understanding of biological systems. While living cells present immense complexity, making it challenging to show causality, synthetic peptide libraries strike a balance by offering a controlled, but representative medium that facilitates the identification of underlying factors in the analysed processes. The mere physical existence of the 3 new synthetic peptide libraries is an immense financial and timewise effort which requires highly qualified scientific personnel. It will allow the Mechtler group (and maybe other groups as well) to further conduct research where the libraries are implemented. The collection of peptides is a realistic representation of the E. coli ribosomal complex digestion by incorporating sequences from 38 proteins originating from this organelle. But more importantly, all the raw data collected among with the specific analyses done and interpretations are published and freely available for access. Other researchers can make use of it for benchmarking their crosslink search engines.

DSSO, DSBU, CDI, DSBSO, ADH and DHSOs capabilities were benchmarked. MS Annika proved to identify the most XLs when using DSSO, but DSBU proved slightly more successful in all other algorithms. By probing multiple reaction conditions, an optimal concentration as well as reaction times were validated to obtain the best outcomes. Beyond that, it was also validated that digestions are prone to fail in some cases, as well as the reductions with TCEP. It should be analysed if and in which proportion sulfoxide-linkers are reduced by TCEP or other reducing agents that are employed to reduce the disulphide bridges or the azide groups.

Furthermore, in future experiments with CDI, a search with the setting K-KTSY ought to be considered as well instead of resuming to the low number of higher resolution (in the context of protein structure determination- the level of detail achieved) links obtained.

The affinity enrichment method proved to be more efficacious than the size-exclusion method in our experiments. Not only that, but AE-XL managed to unveil ~70% of all discoverable XLs, even when the sample was spiked into a non-crosslinked HEK proteome at a ratio of 1:100. The obtained result demonstrates the utility of the trifunctional linkers.

A workflow for linking D/E residues was shortened and enhanced by eliminating an unnecessary desalting step. This protocol change not only prevented sample loss, but also led to an increased total number of links achieved. There is a necessity to further improve the yields of D/E targeting linkers due to their prevalence in proteins and to cover the regions where lysine could be missing.

Further research should be realised in the domain of crosslinking with fatty acids, and it could be the chance for XL-MS to gain as much popularity as cryo-EM, for example by managing to produce consistent reliable results in the case of membrane protein-structure, dynamics, and functioning mechanisms (which have shown to be more demanding). Additional confidence in obtained XLs can be gained by crosslinking in the context of a micelle which adds value to protein structure modelling and interaction surface modelling from in vitro constituted complexes. The approach does not influence resolution which is always dependent on the linker length and hence still lower compared to cryo EM. It is worth mentioning that there are multitude of advantages of XL-MS over cryo EM as well (e.g. no purified protein needed, applicable in native conditions reflecting the reality). Our work with GroEL protein proved to be an interesting starting point of creating an improved protocol to obtain higher quality links with a reduced number of non-specific interactions.

5. List of abbreviations

ABC- Ammonium bicarbonate
HEK – human embryonic kidney 293
rcf – relative centrifugal force
PBS – phosphate buffered saline
TRIS - tris(hydroxymethyl)aminomethane
DTT - Dithiothreitol
IAA – 2-Iodoacetamide
LysC – endoproteinase which cleaves peptide bonds at the carboxyl side of lysine[76]
TFA - Trifluoroacetic acid
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)
TCEP - tris(2-carboxyethyl)phosphine
HEPES - 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
DMTMM- 4-(4,6-dimethoxy-1,3,5-triazin-2-yl)-4-methyl-morpholinium chloride
DBCO- Dibenzocyclooctin
IGX – in gel crosslinking

6. References

1. Griffiths, J. A Brief History of Mass Spectrometry. doi:10.1021/ac8013065.
2. The Nobel Prize in Chemistry 2002 - NobelPrize.org. https://www.nobelprize.org/prizes/chemistry/2002/summary/.
3. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nat. 2003 4226928* **422**, 198–207 (2003).
4. Banerjee, S. & Mazumdar, S. Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. *Int. J. Anal. Chem.* **2012**, 1–40 (2012).
5. Banerjee, S. & Mazumdar, S. Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. *Int. J. Anal. Chem.* **2012**, 40 (2012).
6. Kim, J. Sample preparation for matrix-assisted laser desorption/ionization mass spectrometry. *Mass Spectrom. Lett.* **6**, 27–30 (2015).
7. Jaskolla, T. W. & Karas, M. Compelling evidence for Lucky Survivor and gas phase protonation: the unified MALDI analyte protonation mechanism. *J. Am. Soc. Mass Spectrom.* **22**, 976–988 (2011).
8. Yang, H.-J., Lee, A., Lee, M.-K., Kim, W. & Kim, J. Detection of Small Neutral Carbohydrates Using Various Supporting Materials in Laser Desorption/Ionization Mass Spectrometry. *Carbohydr. Detect. Using Var. LDI-MS Methods Bull. Korean Chem. Soc* **31**, 35 (2010).
9. Makarov, A. Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Anal. Chem.* **72**, 1156–1162 (2000).
10. Michalski, A. *et al.* Mass spectrometry-based proteomics using Q exactive, a high-performance benchtop quadrupole orbitrap mass spectrometer. *Mol. Cell. Proteomics* **10**, (2011).
11. Construction details of the Q Exactive. This instrument is based on the... | Download Scientific Diagram. https://www.researchgate.net/figure/Construction-details-of-the-Q-Exactive-This-instrument-is-based-on-the-Exactive-platform_fig2_51192900.
12. Kaufmann, A. & Bromirski, M. Selecting the best Q Exactive Orbitrap mass spectrometer scan mode for your application. *Thermo Fish. Broch.*
13. 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).
14. 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.* (2020) doi:10.1021/acs.jproteome.0c00583.
15. Miteva, Y. V, Budayeva, H. G. & Cristea, I. M. Proteomics-based methods for discovery, quantification, and validation of protein-protein interactions. *Anal Chem* **85**, 749–768 (2013).
16. O'Reilly, F. J. & Rappsilber, J. Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. *Nat. Struct. Mol. Biol.* **25**, 1000–1008 (2018).
17. Mohr, A. Quantum Computing in Complexity Theory and Theory of Computation. *Carbondale, IL* 1–6 (2014).

18. Kao, A. *et al.* Development of a Novel Cross-linking Strategy for Fast and Accurate Identification of Cross-linked Peptides of Protein Complexes* □ S. *Mol. Cell. Proteomics* **10**, M110.002170 (2011).

19. Müller, M. Q., Dreiocker, F., Ihling, C. H., Schäfer, M. & Sinz, A. Cleavable cross-linker for protein structure analysis: Reliable identification of cross-linking products by tandem MS. *Anal. Chem.* **82**, 6958–6968 (2010).

20. Sinz, A. Divide and conquer: cleavable cross-linkers to study protein conformation and protein-protein interactions. *Anal. Bioanal. Chem.* doi:10.1007/s00216-016-9941-x.

21. Bennett, K. L. *et al.* Chemical cross-linking with thiol-cleavable reagents combined with differential mass spectrometric peptide mapping-A novel approach to assess intermolecular protein contacts. *Protein Sci.* **9**, 1503–1518 (2000).

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

23. Gardner, M. W., Vasicek, L. A., Shabbir, S., Anslyn, E. V. & Brodbelt, J. S. Chromogenic cross-linker for the characterization of protein structure by infrared multiphoton dissociation mass spectrometry. *Anal. Chem.* **80**, 4807–4819 (2008).

24. Gardner, M. W. & Brodbelt, J. S. Preferential cleavage of N-N hydrazone bonds for sequencing bis-arylhydrazone conjugated peptides by electron transfer dissociation. *Anal. Chem.* **82**, 5751–5759 (2010).

25. Dreiocker, F., Müller, M. Q., Sinz, A. & Schäfer, M. Collision-induced dissociative chemical cross-linking reagent for protein structure characterization: Applied Edman chemistry in the gas phase. *J. Mass Spectrom.* **45**, 178–189 (2010).

26. Müller, M. Q., Dreiocker, F., Ihling, C. H., Schäfer, M. & Sinz, A. Fragmentation behavior of a thiourea-based reagent for protein structure analysis by collision-induced dissociative chemical cross-linking. *J. Mass Spectrom.* **45**, 880–891 (2010).

27. Müller, M. Q., Dreiocker, F., Ihling, C. H., Schäfer, M. & Sinz, A. Cleavable cross-linker for protein structure analysis: Reliable identification of cross-linking products by tandem MS. *Anal. Chem.* **82**, 6958–6968 (2010).

28. Arlt, C. *et al.* Integrated Workflow for Structural Proteomics Studies Based on Cross-Linking/Mass Spectrometry with an MS/MS Cleavable Cross-Linker. *Anal. Chem.* **88**, 7930–7937 (2016).

29. Kao, A. *et al.* Development of a Novel Cross-linking Strategy for Fast and Accurate Identification of Cross-linked Peptides of Protein Complexes. *Mol. Cell. Proteomics* **10**, M110.002170 (2011).

30. Bigi, F., Maggi, R. & Sartori, G. Selected syntheses of ureas through phosgenesubstitutes. *Green Chem.* **2**, 140–148 (2000).

31. Padiya, K. J. *et al.* Unprecedented 'in water' imidazole carbonylation: Paradigm shift for preparation of urea and carbamate. *Org. Lett.* **14**, 2814–2817 (2012).

32. Dachs, K. & Schwartz, E. Pyrrolidon, Capryllactam und Laurinlactam als neue Grundstoffe für Polyamidfasern. *Angew. Chemie* **74**, 540–545 (1962).

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

34. Iacobucci, C. & Sinz, A. To Be or Not to Be? Five Guidelines to Avoid Misassignments in Cross-Linking/Mass Spectrometry. *Anal. Chem.* **89**, 7832–7835 (2017).

35. 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).

36. Zhang, H. *et al.* Identification of protein-protein interactions and topologies in living cells with chemical cross-linking and mass spectrometry. *Mol. Cell. Proteomics* **8**, 409–420 (2009).

37. Weisbrod, C. R. *et al.* In vivo protein interaction network identified with a novel real-time cross-linked peptide identification strategy. *J. Proteome Res.* **12**, 1569–1579 (2013).

38. Chowdhury, S. M., Munske, G. R., Tang, X. & Bruce, J. E. Collisionally activated dissociation and electron capture dissociation of several mass spectrometry-identifiable chemical cross-linkers. *Anal. Chem.* **78**, 8183–8193 (2006).

39. 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 (2005).

40. Kaake, R. M. *et al.* A new in vivo cross-linking mass spectrometry platform to define protein-protein interactions in living cells. *Mol. Cell. Proteomics* **13**, 3533–3543 (2014).

41. Steigenberger, B., Albanese, P., Heck, A. J. R. & Scheltema, R. A. To cleave or not to cleave in XL-MS? *J. Am. Soc. Mass Spectrom.* **31**, 196–206 (2020).

42. Partis, M. D., Griffiths, D. G., Roberts, G. C. & Beechey, R. B. Cross-linking of protein by ω-maleimido alkanoyl N-hydroxysuccinimido esters. *J. Protein Chem.* **2**, 263–277 (1983).

43. 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).

44. Kaake, R. M. *et al.* A new in vivo cross-linking mass spectrometry platform to define protein-protein interactions in living cells. *Mol. Cell. Proteomics* **13**, 3533–3543 (2014).

45. Lauria, A. *et al.* 1,2,3-Triazole in Heterocyclic Compounds, Endowed with Biological Activity, through 1,3-Dipolar Cycloadditions. *European J. Org. Chem.* **2014**, 3289–3306 (2014).

46. Matzinger, M., Kandioller, W., Doppler, P., Heiss, E. H. & Mechtler, K. Fast and Highly Efficient Affinity Enrichment of Azide-A-DSBSO Cross-Linked Peptides. *Cite This J. Proteome Res* (2020) doi:10.1021/acs.jproteome.0c00003.

47. Holding, A. N. XL-MS: Protein cross-linking coupled with mass spectrometry. *Methods* **89**, 54–63 (2015).

48. Rodriguez, J., Gupta, N., Smith, R. D. & Pevzner, P. A. Does trypsin cut before proline? *J. Proteome Res.* **7**, 300–305 (2008).

49. Corporation, P. Sequencing Grade Modified Trypsin, Frozen, Product Information 9PIV5113. (1998).

50. Giansanti, P., Tsiatsiani, L., Low, T. Y. & Heck, A. J. R. Six alternative proteases for mass spectrometry–based proteomics beyond trypsin. *Nat. Protoc. 2016 115* **11**, 993–1006 (2016).

51. Schilling, B., Row, R. H., Gibson, B. W., Guo, X. & Young, M. M. MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides. *J. Am. Soc. Mass Spectrom.* **14**, 834–850 (2003).

52. Wang, J. & Boisvert, D. C. Structural basis for GroEL-assisted protein folding from the crystal structure of (GroEL-KMgATP)14 at 2.0 Å resolution. *J. Mol. Biol.* **327**, 843–855 (2003).

53. RCSB PDB - 1KP8: Structural Basis for GroEL-assisted Protein Folding from the Crystal Structure of (GroEL-KMgATP)14 at 2.0 A Resolution. https://www.rcsb.org/structure/1kp8.

54. Miranda, L. P. & Alewood, P. F. Accelerated chemical synthesis of peptides and small proteins. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 1181 (1999).

55. 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.

56. Dorfer, V. *et al.* MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *J. Proteome Res.* **13**, 3679–3684 (2014).

57. 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).

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

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

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

61. Faucher, A. M. & Grand-Maître, C. Tris(2-Carboxyethyl)phosphine (TCEP) for the reduction of sulfoxides, sulfonylchlorides, N-oxides, and azides. *Synth. Commun.* **33**, 3503–3511 (2003).

62. Q Exactive™ BioPharma Platform. https://www.thermofisher.com/order/catalog/product/0726055.

63. 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).

64. Hevler, J. F. *et al.* Selective cross-linking of coinciding protein assemblies by in-gel cross-linking mass spectrometry. *EMBO J.* **40**, (2021).

65. Hevler, J. F. *et al.* Selective cross-linking of coinciding protein assemblies by in-gel cross-linking mass spectrometry. *EMBO J.* **40**, (2021).

66. Erickson, H. P. Size and Shape of Protein Molecules at the Nanometer Level Determined by Sedimentation, Gel Filtration, and Electron Microscopy. *Biol. Proced. Online* **11**, 32 (2009).

67. Carpenter, E. P., Beis, K., Cameron, A. D. & Iwata, S. Overcoming the challenges of membrane protein crystallography. *Curr. Opin. Struct. Biol.* **18**, 581 (2008).

68. Puthenveetil, R., Christenson, E. T. & Vinogradova, O. New Horizons in Structural Biology of Membrane Proteins: Experimental Evaluation of the Role of Conformational Dynamics and Intrinsic Flexibility. *Membr. 2022, Vol. 12, Page 227* **12**, 227 (2022).

69. Seymour, K. G. Surfactants and Interfacial Phenomena. *J. AOAC Int.* **62**, 700–700 (1979).

70. Helenius, A., McCaslin, D. R., Fries, E. & Tanford, C. Properties of detergents. *Methods Enzymol.* **56**, 734–749 (1979).

71. Gaines, G. L. Critical micelle concentrations of aqueous surfactant systems. *J. Colloid Interface Sci.* **38**, 671–672 (1972).

72. Neugebauer, J. M. [18] Detergents: An overview. *Methods Enzymol.* **182**, 239–253 (1990).

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. Wang, J. & Boisvert, D. C. Structural basis for GroEL-assisted protein folding from the crystal structure of (GroEL-KMgATP)14 at 2.0 Å resolution. *J. Mol. Biol.* **327**, 843–855 (2003).

75. Murray, K. K. *et al.* Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013). *Pure Appl. Chem.* **85**, 1515–1609 (2013).

76. Endoproteinase LysC | NEB. https://international.neb.com/products/p8109-endoproteinase-lysc#Product Information.

7. List of Figures