Improved Cross-Linking Mass Spectrometry Algorithms for Probing Protein Structures and Interactions

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

M. Sc. Eugen Netz

aus Omsk / Russland

Tübingen
2023

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	25.09.2023
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Oliver Kohlbacher
2. Berichterstatter:	Prof. Dr. Andrei Lupas

Erklärung

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:

Improved Cross-Linking Mass Spectrometry Algorithms for Probing Protein Structures and Interactions

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich erkläre, dass die Richtlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 25.5.2000) beachtet wurden. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Abstract

Proteins are the most active molecules in living bodies. They catalyze chemical reactions, provide structural support for cells and allow organisms to move. Their function is intrinsically linked to their folded structure. Resolving the structures of proteins and protein complexes is crucial for our understanding of basic biological processes and diseases. Cross-Linking Mass Spectrometry (XL-MS) is a method to gain structural insights into protein complexes. The field of XL-MS data analysis software is not yet as established as many other methods in proteomics. XL-MS analysis software has significant room for improvement in terms of sensitivity, efficiency and standardization of file formats and workflows to facilitate interoperability and reproducibility.

In this thesis we present a new XL-MS search engine, OpenPepXL. We develop an algorithm that scores all candidate cross-linked peptide pairs and is efficient enough to be used on a standard desktop PC for most applications. OpenPepXL supports the standardized XL-MS identification file format defined as a part of the MzIdentML 1.2 specifications that were developed in collaboration with the Proteomics Standards Initiative.

We benchmark OpenPepXL against other state-of-the-art XL-MS identification tools on multiple datasets that allow cross-link validation through structures or other means. We show that our exhaustive approach, although not the quickest one, is superior in sensitivity to other tools. We suggest this is due to some tools improving their processing time by discarding too many candidates in early steps of the data analysis.

We apply XL-MS analysis with OpenPepXL to multiple protein complexes related to meiosis and the type III secretion system. The first project involved several proteins with unknown structures, some of which are expected to be at least partially intrinsically disordered and therefore difficult to investigate using most traditional structural research methods. Unfortunately, we could not find cross-links between the interaction sites we were interested in the most, but we were able to identify many others in these complexes and gained some structural insights. In the second project we used the photo-cross-linking amino acid pBpa to test very specific hypotheses about

interactions within the type III secretion system. We were not able to gain any new structural information yet. However, we could confirm that this is a viable approach. It is possible to identify cross-links between a pBpa residue incorporated into a protein sequence and a residue it cross-links to on a residue level resolution.

Zusammenfassung

Proteine sind die aktivsten Moleküle in Lebewesen. Sie katalysieren chemische Reaktionen, bilden das Zytoskelett jeder Zelle und ermöglichen es jedem Organismus sich zu bewegen. Ihre Funktion ist untrennbar mit ihrer gefalteten Struktur verbunden. Die Aufklärung der Strukturen von Proteinen und Proteinkomplexen ist entscheidend für unser Verständnis grundlegender biologischer Prozesse und Krankheiten.

Cross-Linking Mass Spectrometry (XL-MS) verbindet die chemische oder UV-induzierte Proteinquervernetzung mit Massenspektrometrie, um strukturelle Einsichten in Proteinkomplexe zu erlangen. Das Forschungsgebiet der XL-MS Datenanalyse Software ist noch nicht soweit etabliert, wie viele andere Methoden in der Proteomik. XL-MS Software haben einen erheblichen Raum für Verbesserung bezüglich Sensitivität, Effizienz und Standardisierung von Dateiformaten und Workflows, um Kompatibilität und Reproduzierbarkeit untereinander zu unterstützen.

In dieser Arbeit präsentieren wir ein neues XL-MS Identifikationsprogramm, OpenPepXL. Wir entwickeln einen Algorithmus, der jeden möglichen Cross-Link in Betracht zieht und trotzdem effizient genug ist, um für die meisten Anwendungen auf einem normalen Desktop PC verwendbar zu sein. OpenPepXL unterstützt das Standardisierte XL-MS Identifikationsdateiformat, das als Teil der MzIdentML 1.2 Spezifikation definiert und in Kollaboration mit der Proteomics Standards Initiative entwickelt wurde.

Wir benchmarken OpenPepXL auf mehreren strukturell oder auf andere Weise validierbaren Datensätzen gegen andere XL-MS Identifikationsprogramme auf dem neuesten Stand der Technik. Wir zeigen, dass unser erschöpfender Ansatz vielleicht nicht der schnellste, aber jedoch sensitiver ist als die anderen. Wir legen nahe, dass der Grund dafür ist, dass manche andere Programme ihre Laufzeit verbessern, indem sie zu viele Kandidaten in frühen Schritten der Datenprozessierung verwerfen.

Wir wenden eine XL-MS Analyse mit OpenPepXL auf mehrere Proteinkomplexe in engem Zusammenhang mit Meiose und dem Typ-III-Sekretionssystem an. Im ersten Projekt haben wir uns mit mehreren Proteinen unbekannter Struktur auseinandergesetzt.

Manche von ihnen waren zumindest teilweise unstrukturiert und deshalb schwierig zu erforschen mit den meisten traditionellen Methoden zur Strukturbestimmung. Unglücklicherweise konnten wir keine Cross-Links für die Interaktionen finden, in die wir am meisten interessiert waren. Jedoch wurden viele andere Cross-Links in diesen Proteinkomplexen gefunden und wir haben einige strukturelle Einsichten gewonnen. Im zweiten Projekt verwendeten wir die Photo-Cross-Linking Aminosäure pBpa, um sehr spezifische Hypothesen über Interaktionen innerhalb des Typ-III-Sekretionssystems zu testen. Wir konnten noch keine neuen strukturellen Einsichten gewinnen. Jedoch konnten wir die Realisierbarkeit dieses Ansatzes bestätigen. Es ist möglich eine Aminosäure zu identifizieren, zu der eine in ein Protein eingebaute pBpa Aminosäure einen Cross-Link gebildet hat.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to Prof. Oliver Kohlbacher for giving me this opportunity and for his continuous guidance, support, trust and patience during the last few years.

I thank all the other previous and current members of the Applied Bioinformatics group at the University of Tübingen for the abundant discussions about data processing and coding, as well as the less dry conversations we had on many occasions, and for creating a productive and enjoyable working environment. I especially thank Jens Krüger, who through his lectures ignited my interest in protein structures long before I started on this journey and then supervised my Master thesis, which was in a way where this whole mess started.

I also thank the OpenMS team, both here in Tübingen and elsewhere in the world, for being such a great community to be a part of. Many thanks to Timo Sachsenberg, without whom my introduction to OpenMS and C++ coding would have been much more bumpy, and Mathias Walzer who introduced me to the beautiful world of XML file format specifications and controlled vocabularies.

I thank Prof. Andrei Lupas for his Christmas parties and also his entire Protein Evolution group, from whom I learned a lot about protein sequences, structures and evolution, often with a beer bottle in hand on the lofty terrace of the MPI for Developmental Biology. I also wholeheartedly thank the few members of our small Biomolecular Interactions group at the MPI, Tjeerd Dijkstra and Lukas Zimmermann with whom I explored multiple cities inside and outside of Europe, and Hadeer Elhabashy for all the conversations about protein structure elucidation.

Last but not least, I am deeply grateful to my extended family. I want to thank my parents Olga and Sergei who always supported my career path, even though they still don't quite understand what it is I do. I also want to thank my cousins Eduard, Eduard and Max and my brother Alex, who are also my best friends, for all the times we relaxed on a poolside, beach, balcony, while grilling shashlik, or on a couch while playing video games. I largely attribute my mental health to them.

Contents

Chapter 1

Introduction

Motivation

Proteins are the molecular machinery of life. They play important roles as enzymes catalyzing the biochemical reactions of life, as the skeletal structure for cells, and as nanomachines enabling single cells as well as larger organisms to move. The folded structure of a protein is essential for its function. Many necessary functions can not be fulfilled by single proteins and require them to interact and form large complexes. Many diseases can be attributed to the misfolding of proteins or the failed assembly of protein complexes. Understanding how they fold, interact and function is a major goal of the life sciences. It is crucial for our understanding of cellular functions and diseases, as well as for designing effective drugs, gene therapies, food crops, and processes for industrial production using biosynthesis.

Most proteins are made of 20 canonical amino acids that differ in their physical and chemical properties. They interact with each other and their environment in complex ways to fold into functional domains and multi-protein complexes. A general algorithm to directly predict the structure of a protein complex of arbitrary size from its sequence does not exist to date. Sometimes the folding process is modulated by outside factors such as the crowded intracellular environment, or even explicitly guided by a class of proteins called chaperones. So far it has proved prohibitively complex to simulate the folding of large proteins and protein complexes into their native structure. Several experimental and computational approaches are being used to tackle this problem.

Elucidating Protein Structures

Experimental methods have been around for more than a hundred years, starting with X-Ray Crystallography (XRC)[1] in the 1910s. At first, it was only applicable to small molecules, but the technology improved and later the other methods Nuclear Magnetic Resonance (NMR) spectroscopy,[2] and most recently atomic resolution Cryogenic Electron Microscopy (Cryo-EM) were developed. To date, XRC has contributed the highest number of published structures, but Cryo-EM[3] is becoming more popular with an increasing pace.

All these experimental methods can determine a protein structure with a resolution high enough to assign 3D coordinates to every atom, but they each have limitations. Most of these limitations relate to the size of the studied protein complexes. The larger the target protein complex is, the harder it is to crystallize it for XRC or to assign NMR signals correctly. They are also usually time- and resource-intensive and therefore low throughput. An example of a computational approach are methods that model protein and protein complex structures by using already known experimental structures as templates. There are many proteins that are closely related to each other. This makes it possible to model several proteins computationally after one example from the group has been modeled using experimental approaches. Another computational method making use of evolutionary relationships are Evolutionary Couplings (ECs).[4] They can be inferred from an alignment of homologous sequences and do not require a template. ECs report direct contacts between residues similarly to NMR spectroscopy, but usually the number of contacts that can be confidently identified is less than by using NMR. However, new protein interactions can also be discovered by detecting ECs between proteins. The latest trend is using the popular deep learning approach. This method requires a database of sequences and their structures to train models to represent the relationship between sequence and structure. These trained networks can then predict the structures of completely new proteins without a template. During the 13th and 14th CASP competitions, Deep Mind's AlphaFold proved that this approach is the most effective purely computational method to model proteins so far.[5] However, at the time of writing this thesis, modeling of protein interactions with AlphaFold is limited to dimers.[6] All of these computational methods generally have a higher throughput than the experimental methods, but they are also not seen as reliable enough to be trusted without any experimental confirmation or other corroborating data.

Most of the methods used to study the structure of single proteins can also be applied to protein complexes to some extent. However, there are a few additional methods that specifically look for interactions between proteins. Two commonly used experimental methods are the yeast two-hybrid approach[7] and affinity purification.[8] Both of these methods require the fusion of tags into the targeted proteins. Yeast two-hybrid uses green fluorescent protein that releases a light signal for the readout, while for affinity purification the interacting proteins are usually identified using mass spectrometry. These methods can detect whether proteins are interacting in some way, but they do not provide more specific structural information about the interactions.

Cross-Linking Mass Spectrometry

Cross-Linking Mass Spectrometry (XL-MS) has proven to be a useful tool in studying the structures and interactions of proteins.[9–12] The experiment usually involves chemically inducing covalent bonds between protein residue side chains in a sample and digesting the proteins with proteases. The resulting mixture of cross-linked and linear peptides is then enriched for cross-link containing species in some way, and the sample is analyzed with tandem mass spectrometry. After that, the spectra from the mass spectrometry analysis are searched for cross-linked peptides. This method at the same time allows the detection of novel interactions and provides distance constraints between residues. Although XL-MS provides structurally relevant information, it has an inherently low resolution. However, one benefit of XL-MS is that ultimately the computational problem can be reduced to searching for pairs of cross-linked peptides. It does not matter whether these peptides are from the same protein or not and how large the protein complex physically is. The computational complexity increases with the number of considered target proteins, but that only starts to become a problem with more than a hundred proteins. XL-MS can even be done proteome-wide to look for protein interactions between any proteins in a cell type.[13–15]

Integrative Structural Modeling

Every approach has its own domains of applicability and limitations. Combining data from different approaches can lead to better results than any of them could achieve on their own. Cryo-EM data on large protein complexes tend to have a very high resolution in the hydrophobic core of the complex and a much lower resolution closer to the surface.[16,17] XL-MS data can therefore provide useful additional information in exactly those regions where Cryo-EM data can be difficult to interpret. The same is true for ECs. They provide information about directly interacting residues within the interface of the interaction, where cross-links can not reach. Combining these methods can give a more comprehensive view of a protein complex than any method can provide alone. Today it has become common to build structures of large complexes by fitting XRC or NMR structures into Cryo-EM density maps with the help of interaction data from XL-MS and evolutionary coupling.[18–21] XL-MS and evolutionary coupling are both methods that in principle can be applied proteome-wide and therefore complement each other very well. This combination of complementary methods, experimental or computational, new and modern or long-established, can outperform any one of these methods and makes each one of them even more relevant today than they ever were. This means advances in each method can translate into improvements for many projects researching protein complexes.

XL-MS Algorithms

Although XL-MS has matured as a very useful method, there are opportunities for improvement at every step of the workflow. In many XL-MS experiments the concentration of non-cross-linked peptides is far higher than cross-linked pairs. Cross-linked peptides have lower intensities and are consequently not always selected for fragmentation in data-dependent acquisition. Therefore a few cross-links have to be identified in a large heap of spectra. This is one of the issues that make the computational identification of cross-linked peptide pairs challenging compared to normal peptides.[11,13,22] The search space for XL-MS is also much larger. Because of the pairing of unrelated peptides, the search space for each spectrum is squared compared to linear peptide identification. This leads to high computational complexity and therefore long runtimes. Many of the modern tools rely on heuristics and other shortcuts to reduce the runtime, but this can lead to a loss of sensitivity or specificity.

Protein Complexes of Interest

As mentioned previously, there are various types of protein complexes that are difficult to research with the traditional experimental methods. The double-strand break protein Mer2 is critical in the process of meiotic recombination during meiosis and interacts with multiple other proteins and complexes to achieve that function.[23,24] Mer2 itself and multiple of its interactors have substantial disordered regions. Although these proteins serve such a critical biological function, the structures of some of these proteins and complexes are not known.

The type III secretion system is an overwhelmingly complex and important piece of cellular architecture that is encapsuled in both the inner and outer membranes.[25] It is a major factor in the pathogenicity of many bacteria, like *Salmonella*. Its enormous scale and its hydrophobic regions embedded in the membranes make it difficult to study its assembly and the interactions between components.

Thesis Outline

This thesis mainly describes algorithmic advancements that were implemented in OpenPepXL, an efficient open-source software for the identification of cross-linked peptides in fragment mass spectra. It is based on an exhaustive exploration of the full search space of cross-linked peptide pairs in order to achieve a high sensitivity. In this thesis, OpenPepXL and its algorithms are described. Then its performance is benchmarked against several other commonly used tools for the identification of cross-linkes on a set of diverse XL-MS experiments. It is shown that a thorough search for cross-linked peptide pairs without significant shortcuts can be done with reasonable runtimes and pays off with a higher sensitivity without a loss in specificity. At the end XL-MS analysis is applied to several protein complexes that were difficult to study with other experimental methods. Multiple protein complexes that are involved in the control of DNA structure during meiosis were cross-linked with different cross-linkers. Most of these proteins are very flexible and have functional disordered regions. Additionally protein complexes from the type III secretion system were cross-linked by incorporating the unnatural photo-cross-linking amino acid pBpa into the protein sequence. We show that we can identify and localize pBpa cross-links, making this a viable approach to test very specific hypotheses about protein interactions.

Chapter 2

Background

2.1 Cross-Linking Mass Spectrometry

2.1.1 Mass Spectrometry based Proteomics

Liquid Chromatography coupled to Mass Spectrometry (LC-MS) is currently the high throughput method of choice for the identification and quantification of proteins in biological samples. There are many different methodologies under the umbrella of MS based proteomics with varying ways of preparing the samples and a multitude of instrument types and settings. This background section will cover the basics of bottom-up proteomics with data dependent acquisition, insofar as they are required to understand the procedures and parameters of the computational analysis of XL-MS data.

2.1.2 Sample Preparation

The sample preparation for bottom-up proteomics usually starts with cell lysis and the depletion of unwanted material, like membranes and sometimes also unwanted types of proteins. In many XL-MS experiments a specific protein complex is the target for the analysis and this target complex can be pulled out or purified in a number of ways. The proteins are then cross-linked and denatured. The cystein residues are alkylated using iodoacetamide to disable disulfide bonds. The cysteins are therefore almost always irreversibly modified with a modification called carbamidomethylation that has to be considered during the data analysis. All this leads to a loss of the tertiary structure of the proteins and makes most residues accessible to enzymes. Enzymes are used to digest the proteins into short peptides. Trypsin is used for this most of the time, because peptides cleaved with it will always have a lysine or arginine at the end to ensure at least one positive charge. In some cases the sample is then additionally enriched for specific molecule types, for example peptides with specific modifications or for cross-linked peptides. To enrich for cross-linked peptides usually size exclusion chromatography or cation exchange chromatography[26] are applied to make use of the fact that cross-linked peptide pairs have larger sizes and higher charges than linear peptides. The sample is then desalted and cleaned up in a way to optimize the solution for LC-MS.

Chemical Cross-Linking

XL-MS relies on covalent bonds between amino acid side chains. To capture structurally relevant information, the step of inducing these bonds has to be completed before the proteins are denatured. With some more hydrophobic cross-linking reagents like formaldehyde[27] it is possible to do this before cell lysis, but usually it is done after the proteins of interest were purified. There are several different types of cross-linking reagents. So called zero-length cross-linkers activate amino acid side chains so that they can react with each other directly. This does not add any additional spacing between the amino acids beyond the lengths of the side chains themselves.[28] The more commonly used chemical cross-linkers are reagents with two reactive sites connected by a spacer. The spacer can be a simple carbon chain or a CID-cleavable (or MS-cleavable) chemical group. Disuccinimidyl suberate (DSS) is one of the most commonly used cross-linkers of the non-cleavable type. It has a spacer of 8 carbon atoms with N-hydroxysuccinimide (NHS) esters on both ends. The stretched out spacer has a length of 11.4 Å after the linking reaction and the NHS esters react with primary amines (see Fig. 2.1).

Primary amines are found on each unmodified N-terminus and lysine side chains. Lysine is not a rare amino acid in proteins. Because it is usually positively charged, it is mostly found on the surface of a protein. The reaction of primary amines with NHS esters is also the most efficient cross-linking reaction that targets a specific residue type found so far. These factors have made lysines the favorite target for cross-linking and they are targeted by many types of cross-linkers. DSS is hydrophobic and has a very similar water soluble counterpart in bis(sulfosuccinimidyl)suberate (BS3). BS3 has sulfo groups on its NHS esters giving the molecule a charge. Otherwise the reactivity and spacer length of DSS and BS3 is the same. Both of these cross-linkers also exist in variants with shorter and longer spacers. For example, DSG and BS2G are the hydrophobic and water soluble cross-linkers with a spacer made of 5 atoms and a length of 7.7 Å (see Fig. 2.2).

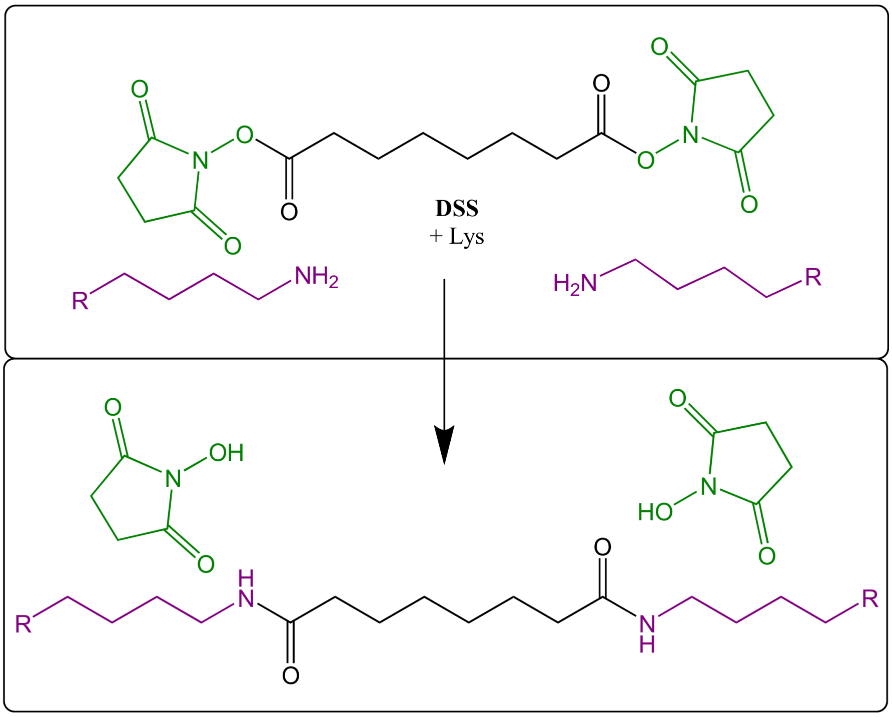


Figure 2.1: The topological structure of DSS and its linking reaction with two lysine side chains. The spacer is black, the linked residue side chains are purple and moieties of the cross-linker that are left as additional reaction products are shown in green.

There are also cross-linkers that target other amino acids. The carboxylic acids of aspartate and glutamate can be activated using DMTMM to allow certain cross-linkers like pimelic acid dihydrazide (PDH) to link these side chains (see Fig. 2.3). Using multiple types of cross-linkers can provide complementary information about protein structures by targeting different residue pairs and distances.

Another class of cross-linking reagents are CID-cleavable cross-linkers. They contain bonds, that have about the same strength as peptide backbone bonds and can be broken at similar collision energies using CID or HCD fragmentation. The most commonly used cross-linkers of this class are DSBU[29] and DSSO,[30] both very similar to DSS in spacer length and reactivity. These linkers are designed with two such breakable bonds around a stable central moiety. Upon fragmentation this central moiety will stay with one of the two peptides. This means for each of the two peptides there will be fragments containing one full peptide mass with a remnant of the cleaved cross-linker with and without this central moiety. To detect these, one just has to look for pairs of peaks with a mass difference equivalent to the mass of this central moiety. The advantage of these cross-linkers is that from these pairs of peaks the masses of the two full peptides can be derived, which significantly reduces the search space for

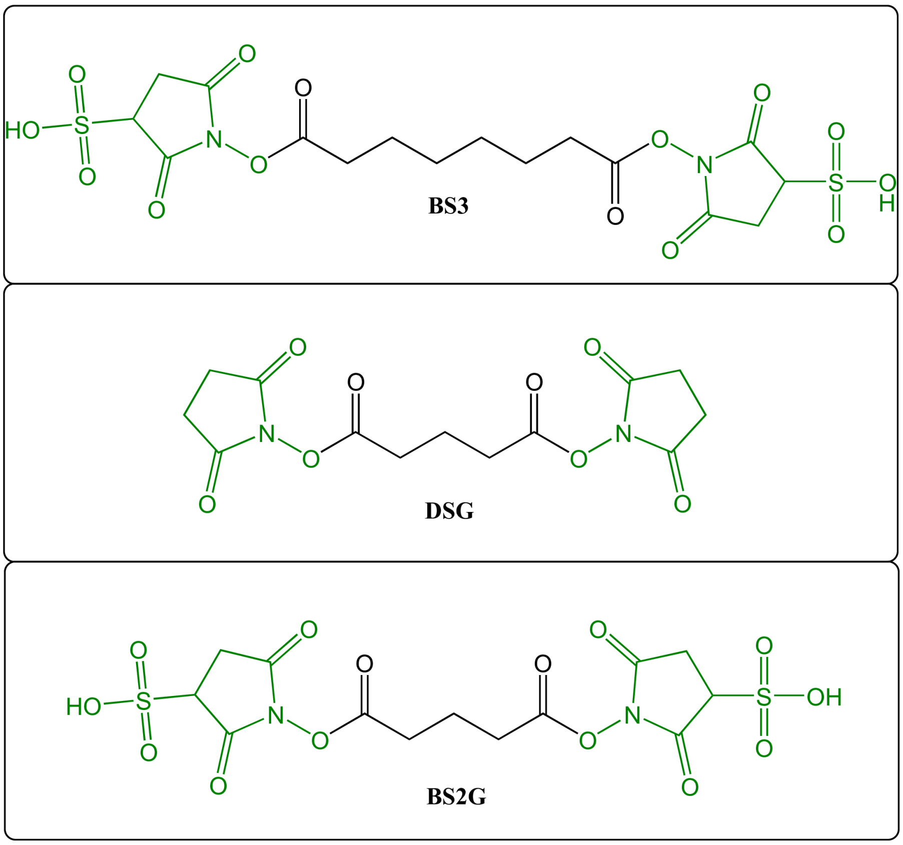


Figure 2.2: The topological structures of BS3, DSG and BS2G. Their linking reactions are equivalent to DSS (see Fig. 2.1). The spacer is black and moieties of the cross-linker that are left as additional reaction products are shown in green.

cross-link identification software and enables the analysis of proteome-wide samples. The structure of DSBU is shown in Fig. 2.4.

In a typical XL-MS experiment, the cross-linker solution is mixed with the protein sample and incubated for a set period of time. The optimal temperature, pH level and time period for this incubation differs between cross-linkers. The cross-linking reaction is then quenched, for example by adding tris-buffered saline in the case of lysine-reactive cross-linkers. Tris(hydroxymethyl)aminomethane or tris has its own primary amine and triggers reactions with the remaining cross-linking reagent. After this the sample preparation can continue with the denaturation and digestion of the proteins without the risk of inducing structurally irrelevant cross-links.

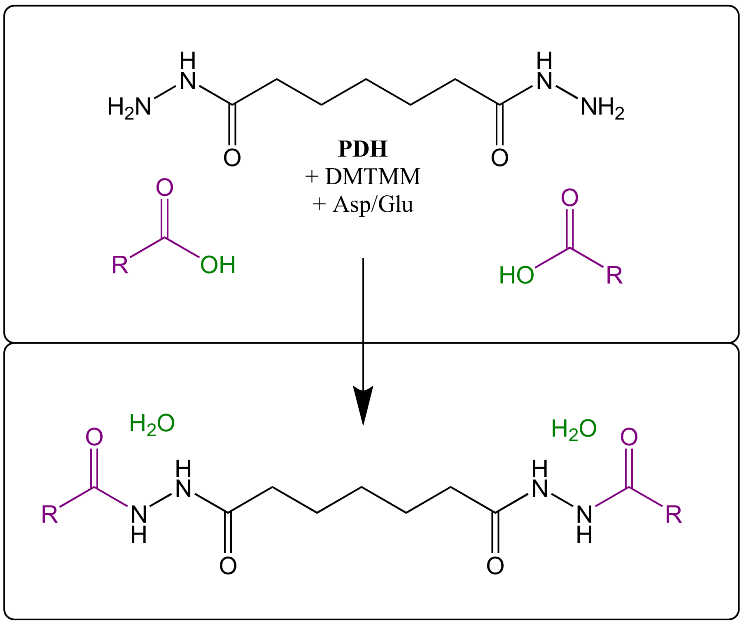


Figure 2.3: The topological structure of PDH and its linking reaction with aspartate or glutamate side chains. The spacer is black, the linked residue side chains are purple and parts of the side chains that are left as additional reaction products are shown in green.

Labeling

Labeling in proteomics is usually used to get a differential quantification of multiple samples. Reagents that introduce heavier isotopes can be fed to living cell cultures, so that they metabolize and use these labeled building blocks for the translation of new proteins. Chemical or stable isotope labels can also be applied *in vitro* to proteins or the digested peptides. Most often the labeled and label-free samples are mixed and analyzed together. The mass difference is used to identify compounds from the different samples and the intensity is used to quantify the differences in protein concentrations. However, labeled chemical cross-linkers have an additional use of aiding the identification of cross-linked peptide pairs. In this case an equimolar mixture of the labeled and unlabeled cross-linker is used and the resulting pairs of masses in both the MS1 and MS2 spectra are used to reduce the search space and to denoise spectra.

2.1.3 Mass Spectrometry Data

A raw mass spectrum acquired by a mass spectrometry instrument can be represented as a 2D line graph. The X-axis shows the mass-to-charge-ratio. It uses the unit

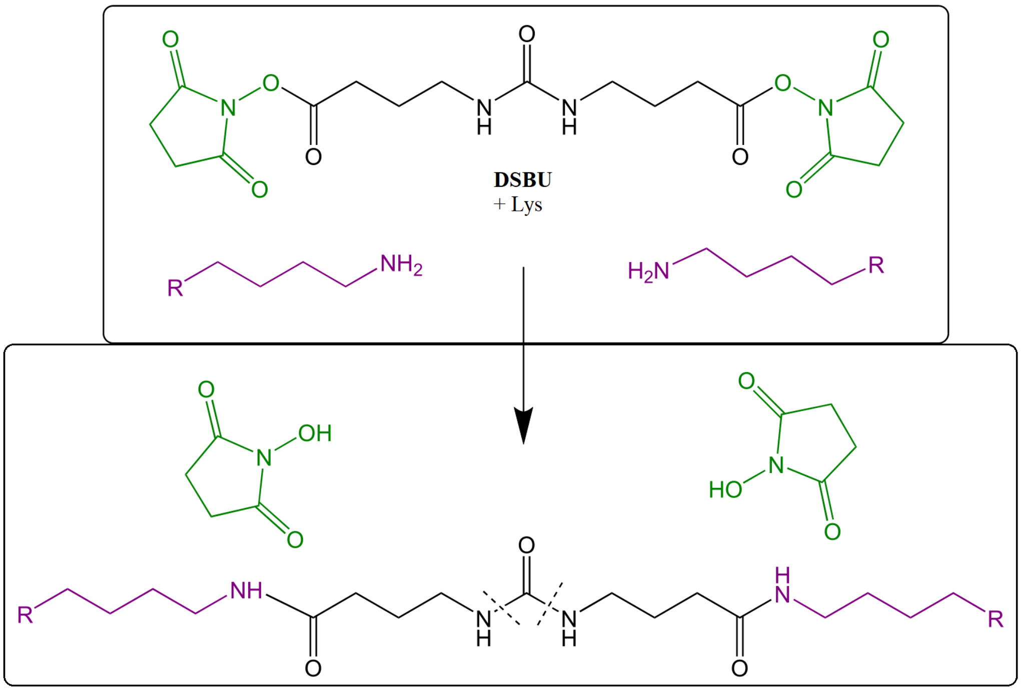


Figure 2.4: The topological structure of DSBU and its linking reaction with two lysine side chains. The spacer is black, the linked residue side chains are purple and moieties of the cross-linker that are left as additional reaction products are shown in green. The dashed lines show the two symmetrical molecule bonds that are CID-cleavable.

$mass/charge$ that is abbreviated as m/z and sometimes also referred to as Dalton (Da), especially in the context of tolerances. The Y-axis shows the intensity. It is proportional to the amount of ions that were detected with any specific m/z. Due to the nature of most experimental measurements, the resulting graph for any detected ion has a bell shaped curve. In a process called peak-picking or centroiding these curves are transformed into one point that represents the m/z and intensity at the position of the highest intensity of that curve. This data reduction step turns the bell-curve for each ion into a pair of two numbers that sufficiently characterizes that ion for most applications. However, for comparisons of m/z-values from different spectra or against theoretical masses, a tolerance always has to be considered. With modern high-resolution orbitrap mass analyzers and the most common resolution settings a relative tolerance in the range between ±5 and ±20 ppm is usually applied. However, the second most common type of mass analyzer still used in proteomics today is the ion trap with a lower accuracy that for older instruments requires absolute tolerances of up to ±0.2 Da. Most molecules result in multiple peaks because of different isotope compositions. After the monoisotopic peak consisting of the most common isotopes for each element, the most intense additional peaks are usually caused by ^{13}C.

In tandem MS or MS/MS two different types of spectra are recorded. At first an MS1 spectrum is acquired from the intact ions as they enter the mass spectrometer from the LC column. A few of the highest intensity peaks are then selected for further analysis. Starting with the highest intensity peak, the instrument starts to filter out most of the ions and only allows ions in a narrow m/z range around the selected peak to go through. These ions are then fragmented, most often by collision-induced dissociation and an MS2 or fragment spectrum is recorded. An example spectrum is shown in Fig. 2.7. The instrument also stores additional information about the MS2 spectrum in meta-information data fields. The isolation window is analyzed to determine the monoisotopic mass and charge of the selected MS1 peak that is called the precursor of this MS2 spectrum. This analysis is done by the instrument during the data acquisition and therefore has to be done very quickly. The algorithms used for this are not always accurate and can misidentify the monoisotopic peak by mistakenly selecting the second or third isotopic peak instead. This is more likely to happen for larger ions where the monoisotopic peak can have a much lower intensity than the second or third peak. After MS2 spectra for the 10 to 20 highest intensity peaks are recorded, their m/z-values are stored in an exclusion list for some time to avoid recording MS2 spectra of the same precursors for the next few MS1 spectra. Then the next MS1 spectrum is recorded and the process is repeated.

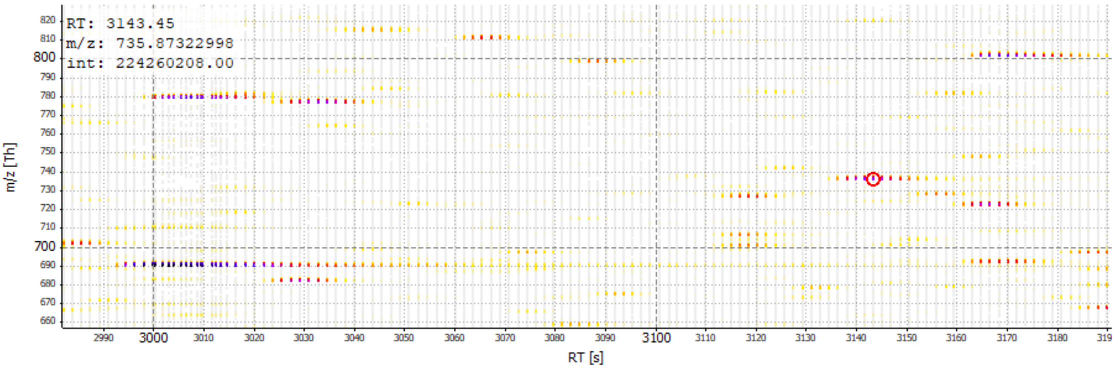


Figure 2.5: Part of an MS1 map with multiple features. The highest peak of one feature is selected and its attributes are shown on the top left. The intensity of peaks is encoded in a color gradient from grey for very low intensities over yellow and red for moderate and purple and blue for high intensities.

By stacking the MS1 spectra in the order of their acquisition, these spectra can be represented as a 3D map with the retention time (RT) in the LC column as a third dimension for each ion. The intensity peaks of a specific isotope composition of an ion

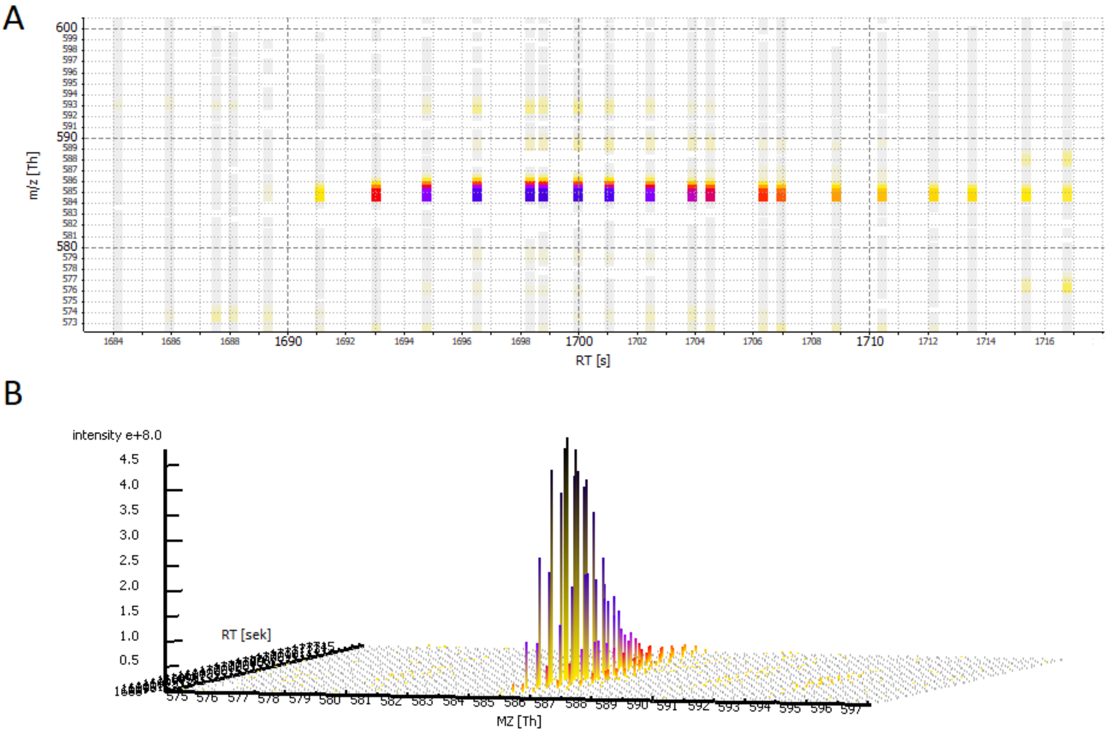


Figure 2.6: Part of an MS1 map showing an MS1 feature. The intensity of peaks is encoded in a color gradient from grey for very low intensities over yellow and red for moderate and purple and blue for high intensities. A) shows the feature in 2D and B) shows the same feature in a tilted 3D representation.

along the RT-axis form a roughly bell shaped curve. Along the m/z-axis the shape is determined by the isotope distribution of the molecule. This three-dimensional shape defined by all peaks that belong to the same ion is called a feature. The detection of features and the assignment of peaks that belong to a feature are important parts of many labeled and label-free quantification methods and are also required to make use of labeled cross-linkers. An example MS1 map with color coded intensities is shown in Fig. 2.5 and an alternative visualization of a single feature is shown in Fig. 2.6.

2.1.4 Basics of Peptide Identification

Among the several approaches for the identification of peptides using LC-MS/MS, the most efficient and most popular method is the database search approach. A protein database containing all protein sequences that are targeted in the search has to be provided to such an algorithm. This database is then digested *in silico* using the cleavage

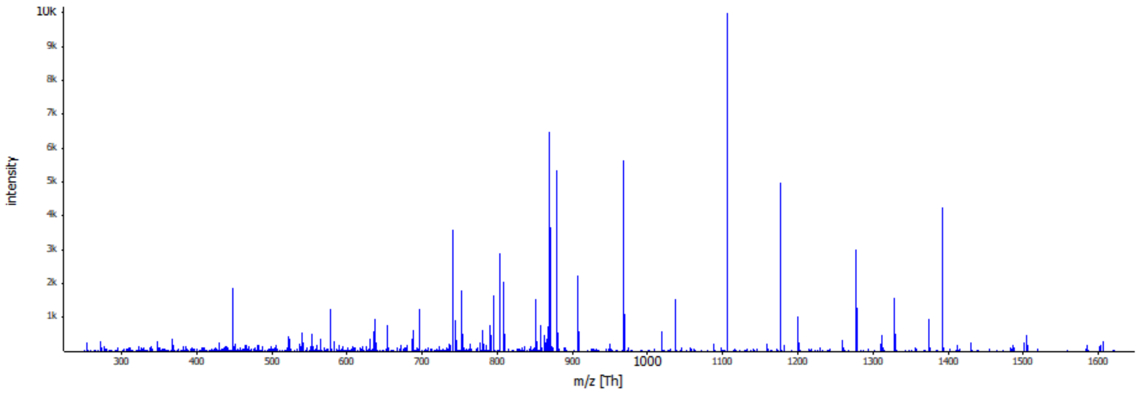


Figure 2.7: The MS2 spectrum of an unidentified peptide.

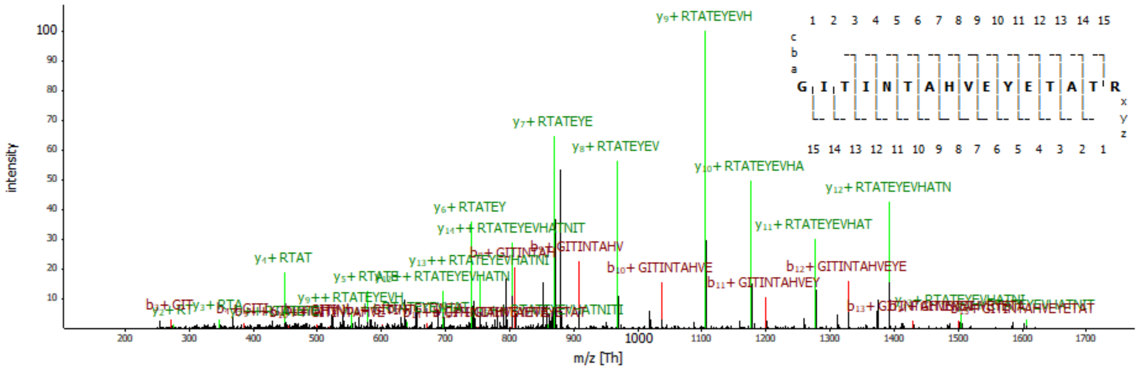


Figure 2.8: The MS2 spectrum of an identified peptide with sequence coverage and labeled matched peaks.

rules of the used enzyme and then fragmented *in silico* as well. The weakest bonds in a peptide are those along the backbone. Each fragmentation method causes the peptide to break at different bond types. For each residue there are three possible bonds along the backbone. Each fragmentation event results in two fragments, the N-terminal and the C-terminal side of the whole peptide. Depending on the broken bond the N-terminal fragments are called a-, b- or c-ions and the C-terminal fragments are called x-, y- and z-ions. The peptide amide bond that results in b- and y-ions is the weakest among these and the most common to break with collision induced dissociation. Using this knowledge together with the known masses of the canonical amino acids, the m/z-values for a theoretical spectrum can be computed for any peptide. The intensity is often set to a constant number, because it is not straightforward to predict. To identify the peptide that was fragmented in a given MS2 spectrum, the list of peptides in the searched database is first filtered by the precursor mass of the MS2 spectrum. Theoretical spectra from the peptides that have a mass within the tolerance of the precursor mass are then computed and compared to the experimental MS2 spectrum. This comparison can be done in many different ways, but it always results in a score: a numerical representation of either the similarity between the experimental and theoretical spectrum or the probability of the current candidate peptide being the correct one. A candidate peptide with its score is called a peptide-spectrum-match (PSM) or cross-link-spectrum-match (CSM) in the case of a cross-linked peptide pair. Using the score the best PSM is accepted as the identification of the current experimental spectrum. An example of a candidate peptide matched against a theoretical spectrum is shown in Fig. 2.8.

2.1.5 Common XL-MS Terminology

- **Dead-end- or mono-link**: If a cross-linker binds to an amino acid side chain with one side and the second reactive side is quenched before it can bind, it leads to a modification of a residue rather than a cross-link.
- **Loop-link**: A cross-link linking two residues on the same peptide after enzymatic digestion.
- **Cross-link**: The name usually used for a link between two peptides.
- α**-peptide**: The first of two cross-linked peptides. This is usually the peptide with the longer peptide sequence or, in case of equal lengths, higher mass.
- β**-peptide**: The second of two cross-linked peptides.

2.1.6 XL-MS Identification Algorithms

There are many tools already available for the identification of cross-linked peptides. The goal of the work described in this thesis was to develop a new tool that addresses some of the shortcomings of these already available tools.

xQuest and xProphet

The xQuest and xProphet pipeline[31–33] is a set of tools written in Perl. It consists of Perls scripts that are started from a command line and does not have a graphical user interface (GUI) for the purpose of setting up and running the tools. It is one of the oldest XL-MS identification tools that is still in popular use today. It was initially developed for lower-resolution MS instruments with ion trap mass analyzers. The developers made use of the search space reduction and spectrum filtering possibilities that labeled chemical cross-linkers make possible to make up for the lower accuracy. It is still the only tool with this capability aside from OpenPepXL.

xQuest also has the option of using pre-calculated fragment ion indices to retrieve peptides from the protein database based on observed fragment ions. This method is also used by several conventional linear peptide search tools like MSFragger[34] and has the potential to substantially speed up the database search by reducing the number of considered candidate peptides for each MS2 spectrum. Other than that the tool does not use any heuristics or shortcuts and scores each candidate pair of cross-linked peptides with its full score. Because of this it is among the slower tools available.

The xQuest algorithm first screens MS1 spectra for peptide masses with a characteristic isotopic shift. MS2 spectra of these precursors are compared to each other to find linear and cross-linked ions. Fragment ions present in both spectra do not contain the cross-linker and are therefore linear ions. Cross-linked ions are identified by their characteristic isotopic mass shift. Because these peak types only need to be matched against theoretical peaks of the same type, this improves the specificity of the later spectrum matching. The ion index is computed from the protein database. For each theoretical fragment ion m/z it contains all peptide sequences that can produce such a fragment. The linear ions are used to query this index and the peptides matched through this ion index form the search space for each spectrum pair. These peptides are then matched against the spectrum and only the best scoring peptides are combined into pairs that match the precursor mass and matched again as cross-link candidates. This pre-selection of peptides is optional and a full enumeration of peptide pairs fitting into the precursor tolerance window is also implemented. The xQuest score of a match between an experimental and a theoretical spectrum is a linear combination of multiple metrics including cross-correlation, the peak match probability based match-odds score and the percentage of the total ion current that was matched to theoretical peaks. The weights of these scores were determined by a linear discriminant analysis on a XL-MS dataset of monomeric standard proteins with manually verified cross-links.

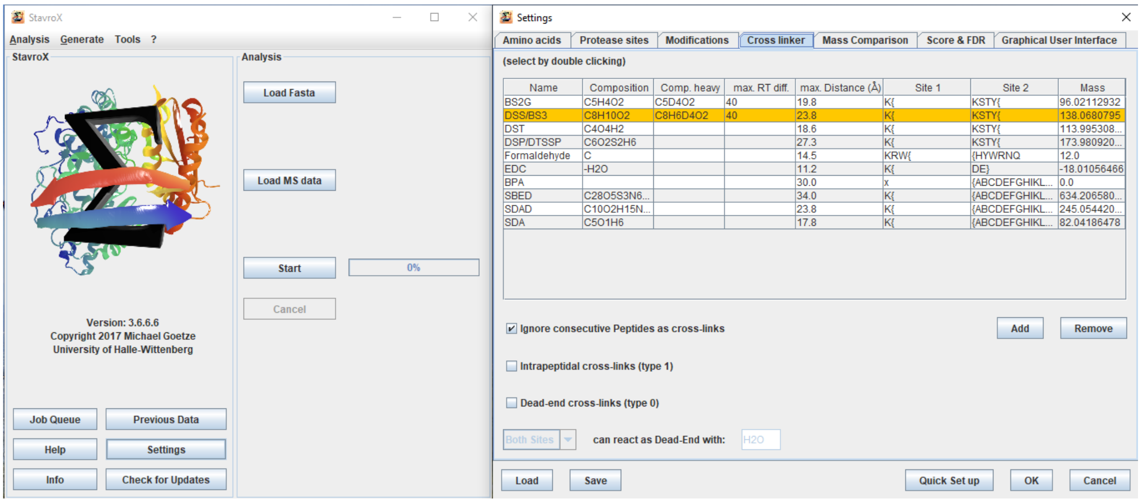
xProphet[33] was developed as a separate script in this pipeline to compute FDR. It separates the top ranked hits for each spectrum into intra- and inter-protein cross-links, as well as one group for loop- and mono-links. The FDR is calculated for each of these groups separately using this formula:

$$\frac{(TD + DT + DD) - 2 \cdot DD}{TT} \tag{2.1}$$

with TT being target-target, DD being decoy-decoy and TD and DT being target-decoy and decoy-target hits for the peptide pairs.

StavroX

StavroX[35] is a tool that scores pairs of cross-linked peptides directly. It is written in Java and can only be used through its GUI. It is also among the slower tools, but it allows to set up and run an analysis and explore the results in one integrated GUI. This makes it the most user friendly tool on this list for analyzing a small number of datasets. Unfortunately the requirement to use the GUI limits its options for automation. The user interface of StavroX is shown in Fig. 2.9. More recently the new tool MeroX[36] was released by the same group. Its major advancement is support for and focus on CID-cleavable cross-linkers. Today the functionality of StavroX is built into MeroX, which overall has the same user interface and functionality regarding non-cleavable cross-linkers as StavroX.



Version: 3.6.6.6
Copyright 2017 Michael Goetze
University of Halle-Wittenberg

(select by double clicking)

Name	Composition	Comp. heavy	max. RT diff.	max. Distance (Å)	Site 1	Site 2	Mass
BS2G	C5H4O2	C5D4O2	40	19.8	K{	KSTY{	96.02112932
DSS/BS3	C8H10O2	C8H6D4O2	40	23.8	K{	KSTY{	138.0680795
DST	C4O4H2			18.6	K{	KSTY{	113.995308...
DSP/DTSSP	C6O2S2H6			27.3	K{	KSTY{	173.980920...
Formaldehyde	C			14.5	KRW{	{HYWRNQ	12.0
EDC	-H2O			11.2	K{	DE}	-18.01056466
BPA				30.0	x	{ABCDEFGHIKL...	0.0
SBED	C28O5S3N6...			34.0	K{	{ABCDEFGHIKL...	634.206580...
SDAD	C10O2H15N...			23.8	K{	{ABCDEFGHIKL...	245.054420...
SDA	C5O1H6			17.8	K{	{ABCDEFGHIKL...	82.04186478

☑ Ignore consecutive Peptides as cross-links

☐ Intrapeptidal cross-links (type 1)

☐ Dead-end cross-links (type 0)

Both Sites ▾ can react as Dead-End with: H2O

Figure 2.9: The graphical user interface of StavroX showing the main menu and one tab of the search parameters.

The algorithm enumerates all cross-linked peptide pair candidates from the protein database based on the precursor mass tolerance windows of the input spectra. For each MS2 spectrum a signal-to-noise estimation is made per ion and ions exceeding a given noise level are excluded from further analysis. For each cross-link candidate to the spectrum the theoretical b- and y- ions are computed and this theoretical spectrum is matched against the experimental one. Neutral losses of water and ammonia are also considered for precursor ions and already matched b- and y-ions. The score is calculated from multiple spectrum quality and match quality metrics:

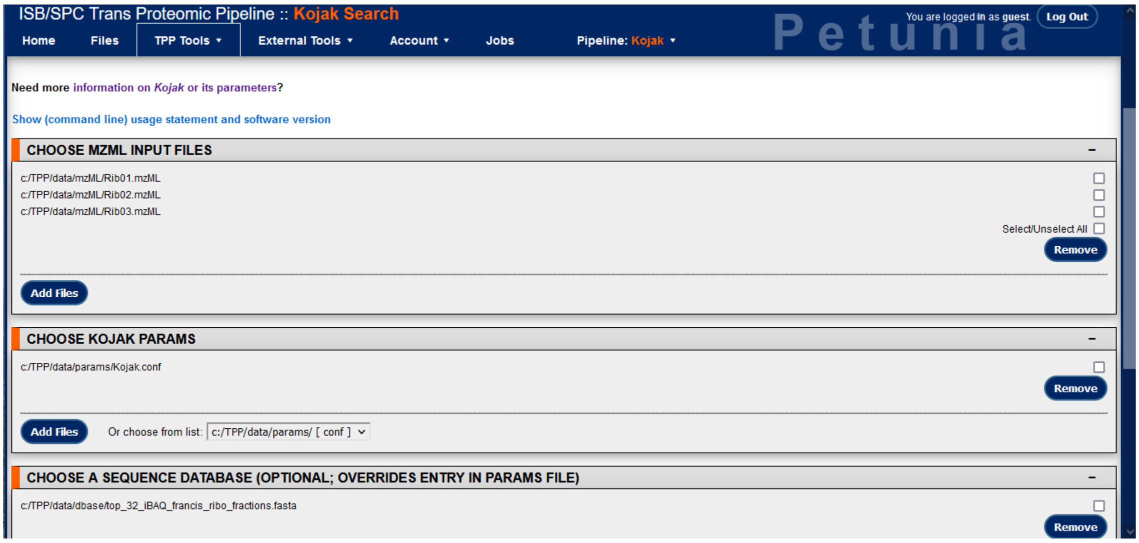
$$-50 \cdot log(\prod_{n} e^{-\frac{s_n}{p1+p2}} \cdot (0.2(1-e^{\frac{|d-300|^5}{10^{12}}}) + 0.2e^{-\frac{i}{6}} + 0.4e^{-\frac{7k}{i}} + 0.2e^{-\frac{20h}{d}})) \qquad (2.2)$$

with s_n: length of series n (b- or y-type ions); $p1$ and $p2$: lengths of cross-linked peptides; d: number of fragment ions in the experimental spectrum with acceptable signal-to-noise ratio; i: number of signals above 10% relative intensity; k: number of matched fragment ions; h: number of all matched ions.

The results are saved in files that can be loaded in later for further inspection and are also directly displayed in the interactive GUI. Such a result file is a collection of multiple non-standardized file types in a zipped folder. The GUI allows the exploration of the table of scored cross-links and matched spectra. StavroX uses an empirical estimation for the relationship between its score and FDR. The developers estimated the FDR for many score thresholds on a benchmark dataset. Setting an FDR is therefore simply done by applying a cut-off at the respective score. For example, the estimated FDR for all hits with a score above 100 is 2%.

Kojak

Kojak[37] is an open source C++ tool that can be used as a command line tool and is also integrated into the Trans-Proteomic Pipeline (TPP)[38] that provides a browser based GUI. The TPP interface is shown in Fig. 2.10.



You are logged in as guest.

Need more information on *Kojak* or its parameters?

Show (command line) usage statement and software version

CHOOSE MZML INPUT FILES

c:/TPP/data/mzML/Rib01.mzML □
c:/TPP/data/mzML/Rib02.mzML □
c:/TPP/data/mzML/Rib03.mzML □

CHOOSE KOJAK PARAMS

c:/TPP/data/params/Kojak.conf □

Or choose from list: c:/TPP/data/params/ [conf]

CHOOSE A SEQUENCE DATABASE (OPTIONAL; OVERRIDES ENTRY IN PARAMS FILE)

c:/TPP/data/dbase/top_32_iBAQ_francis_ribo_fractions.fasta □

Figure 2.10: The browser based graphical user interface of the Trans-Proteomic Pipeline[38] showing a part of the setup of a Kojak run.

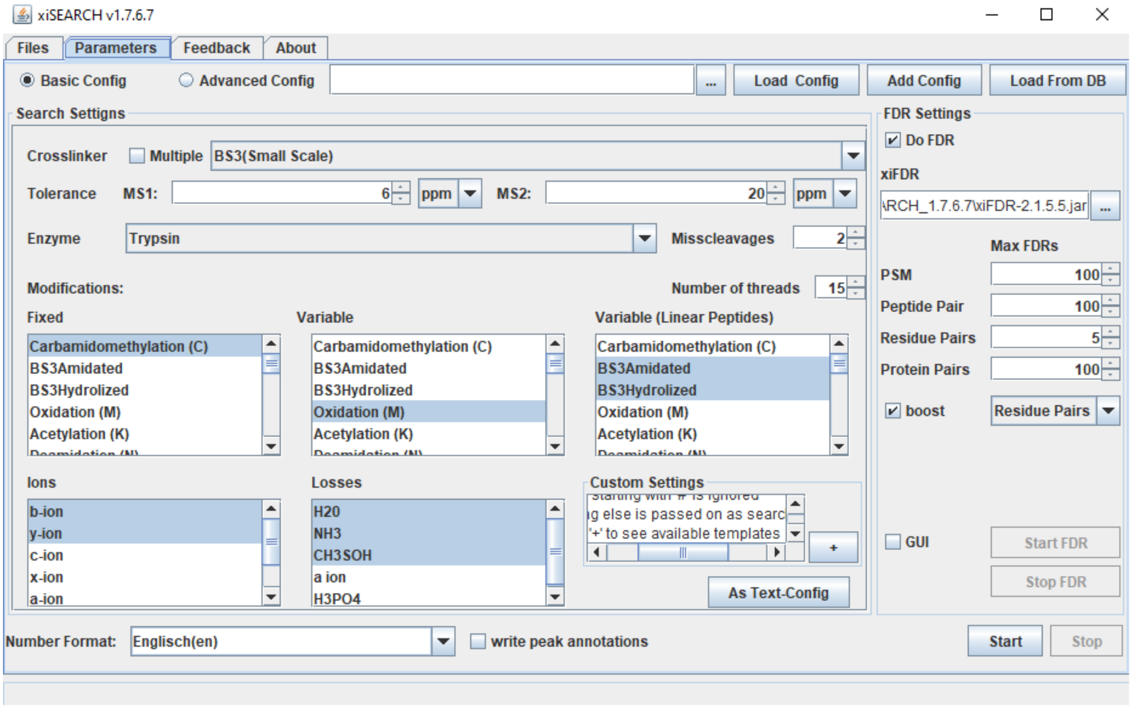
The algorithm starts by pre-processing the experimental spectra. A correction for the monoisotopic mass peak of the precursor ions is applied by predicting the ion isotope envelope from the entire MS1 feature using a model-based precursor fitting function. On high-resolution data an optional deisotoping procedure can be applied to summarize isotope envelopes of fragments into their monoisotopic peak. In an additional step the spectra are pre-processed to normalize the peak intensities in a special way to speed up the scoring later. This pre-processing involves binning the spectrum into a lower-precision sparse array that can be traversed very rapidly using a hash-like appraoch. Peptides are identified in two steps. First, single peptides are matched against the spectrum using an open-modification search strategy. The difference between the precursor mass and the peptide is assumed to be a modification on the linked residue. By default the top 250 scoring peptides are paired according to the precursor mass. Increasing this number improves the sensitivity of the tool in many cases, but due to the way Kojak is implemented, this increases the runtime and memory usage of the tool significantly. The final score for a cross-linked peptide pair is the sum of the individual peptide scores. Kojak scores matches between experimental and theoretical spectra with a fast cross-correlation algorithm derived from the Comet score[39] which itself is derived from the SEQUEST scoring algorithm,[40] but avoids SEQUESTs Fourier transform calculations. Kojaks theoretical spectra only contain b- and y-ions. The pre-processing of peak intensities in the experimental spectrum allows the score to be calculated simply by adding up the intensities of the matched peaks.

The confidence of Kojak identifications is estimated using PeptideProphet.[41] It uses an empirical non-parametric statistical model to compute posterior probabilities and derive a list of FDR controlled correct hits. Kojak is also compatible with the semi-supervised machine learning tool Percolator,[42][43] which can boost the number of correctly identified hits by recalibrating the match score and combining it with additional match quality metrics. Neither PeptideProphet nor Percolator use specific models for XL-MS data and treat target-decoy hits simply as decoy hits.

XiSearch

XiSearch[44–46] is a Java application that can only be used through its GUI and has therefore similar benefits and limitations to StavroX. The group of Xi tools also includes the FDR estimation tool XiFDR and the protein network visualization tool XiView.[47] All these tools support the HUPO-PSI identification format mzIdentML 1.2, which makes them also compatible with OpenPepXL. XiView in particular was used to visualize some networks resulting from OpenPepXL searches in this thesis. The user interface of XiSearch is shown in Fig. 2.11.

XiSearch starts by deisotoping and decharging the experimental spectra. They are then filtered by charge state and m/z to contain mostly linear fragments. All fragments with a charge of +1 and a mass smaller than half of the precursor mass are



● Basic Config ○ Advanced Config

Search Settings

Crosslinker ☐ Multiple BS3(Small Scale)

Tolerance MS1: 6 ppm MS2: 20 ppm

Enzyme Trypsin Misscleavages 2

Modifications: Number of threads 15

Fixed	Variable	Variable (Linear Peptides)
Carbamidomethylation (C)	Carbamidomethylation (C)	Carbamidomethylation (C)
BS3Amidated	BS3Amidated	BS3Amidated
BS3Hydrolized	BS3Hydrolized	BS3Hydrolized
Oxidation (M)	Oxidation (M)	Oxidation (M)
Acetylation (K)	Acetylation (K)	Acetylation (K)

Ions	Losses
b-ion	H20
y-ion	NH3
c-ion	CH3SOH
x-ion	a ion
a-ion	H3PO4

Custom Settings
starting with # is ignored
g else is passed on as searc
'+' to see available templates

FDR Settings
☑ Do FDR
xiFDR
\RCH_1.7.6.7\xiFDR-2.1.5.5.jar

Max FDRs

PSM	100
Peptide Pair	100
Residue Pairs	5
Protein Pairs	100

☑ boost Residue Pairs

☐ GUI

Number Format: Englisch(en) ☐ write peak annotations

Figure 2.11: The graphical user interface of XiSearch showing the tab with search parameters.

assumed to be linear fragments. The remaining cross-linked fragments are linearized by subtracting their mass from the precursor mass. This yields the mass of the linear fragment that was broken off the cross-linked fragment. This spectrum is then filtered to the top 10 most intense peaks to get a simple spectrum that most likely contains fragment peaks from the usually more intense α-peptide. An open-modification search is then performed to find a list of candidates for the α-peptide. The beta peptides are then extracted from the full peptide database by matching the masses to the difference between the α-peptides and the precursor mass. A second round of matching is now performed exhaustively between these cross-linked peptide pairs and the original experimental spectrum. The score XiSearch uses is an adapted ranked dot product (RDP), which is mostly used in spectral library searches.

$$RDP = \frac{S_r \times T_r}{\sqrt{S_r^2 \times T_r^2}} \tag{2.3}$$

where S_r and T_r are the vectors with the binned observed and theoretical peaks of a spectrum. The intensities of the peaks were replaced with their intensity ranks or set to zero if there is no matching peak in the bin.

XiFDR is a second tool in this pipeline and is responsible for FDR estimation.[44] It uses two different FDR formulas depending on the cross-linker chemistry. For some cross-linkers, for example heterobifucntional cross-linkers or other cross-linkers with asymmetrical structure, fragment spectra can look different even for the same peptide pair depending on the direction of the link. XiFDR uses two different formulas for such directional and symmetrical non-directional cross-linkers. The formula fo the directional FDR_d is

$$FDR_d = \frac{TD - DD}{TT} \tag{2.4}$$

and the formula for the non-directional FDR_{nd} is

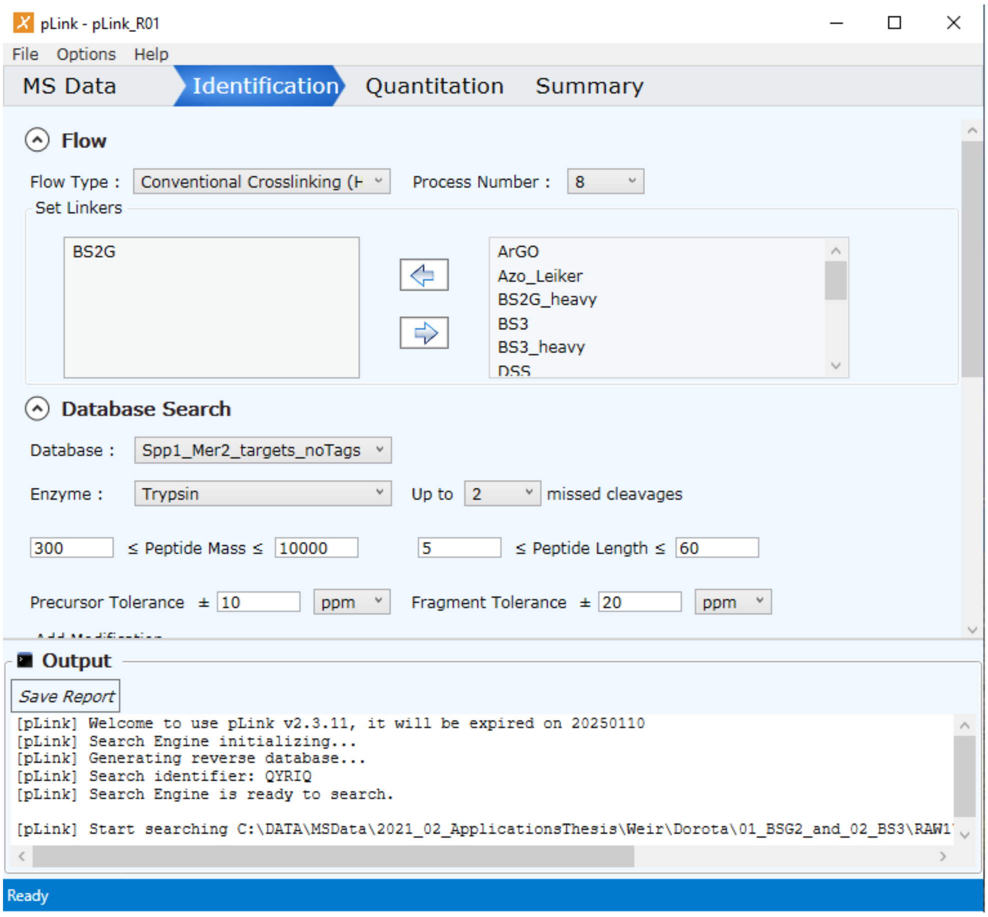
$$FDR_{nd} = \frac{TD + DD(1 - 2\frac{TD_{DB}}{TD_{DB} + \sqrt{TD_{DB}}})}{TT} \tag{2.5}$$

with TT and DD being target-target and decoy-decoy hits, TD being target-decoy and decoy-target hits and TD_{DB} being the number of all possible unique target-decoy or decoy-target peptide pairs in the database. XiFDR can compute FDRs at CSM, peptide-pair, residue-pair and protein-pair levels. XiSearch and XiFDR both support MzIdentML 1.2 additionally to a tool specific table format.

pLink2

pLink2[48] is among the more popular XL-MS identifications tools due to ease of use and performance. It uses several filters and heuristics that make it the fastest tool of this class. It is limited to Windows operating systems and also uses a GUI. Therefore it is user friendly for analyzing small numbers of datasets on Windows PCs, but it is very limited in the ways it can be deployed and automated. The user interface of pLink2 is shown in Fig. 2.12.

The algorithm of pLink2 starts by preprocessing the MS1 map using pParse[49] to correct the monoisotpic peaks of the precursors. Similarly to xQuest, pLink2 builds a fragment ion index from the protein database. It contains a mapping from fragment masses to lists of peptides that contain that fragment. Quality metrics are computed for all MS2 spectra and those that are deemed of low quality are not considered for the search at all. The remaining MS2 spectra are preprocessed and filtered to a smaller number of most intense peaks. These peaks are used to retrieve peptide candidates from the fragment index. These peptides are considered α-peptides and scored using an open modification search. The top 5 are kept and all possible β-peptides are matched to them using the difference to the precursor mass. These peptide pairs now



Flow Type : Conventional Crosslinking (H

Process Number : 8

Set Linkers

BS2G

ArGO
Azo_Leiker
BS2G_heavy
BS3
BS3_heavy
DSS

Database : Spp1_Mer2_targets_noTags

Enzyme : Trypsin Up to 2 missed cleavages

300 ≤ Peptide Mass ≤ 10000 5 ≤ Peptide Length ≤ 60

Precursor Tolerance ± 10 ppm Fragment Tolerance ± 20 ppm

Add Modification

```
[pLink] Welcome to use pLink v2.3.11, it will be expired on 20250110
[pLink] Search Engine initializing...
[pLink] Generating reverse database...
[pLink] Search identifier: QYRIQ
[pLink] Search Engine is ready to search.

[pLink] Start searching C:\DATA\MSData\2021_02_ApplicationsThesis\Weir\Dorota\01_BSG2_and_02_BS3\RAW1
```

Figure 2.12: The graphical user interface of pLink2 showing the tab with search parameters.

go through a second final scoring. Loop-linked, mono-linked and regular peptides are scored with the pFind algorithm,[50] which is the peptide identification tool by the same group.

The match scores used in pLink2 are the same as in its predecessor pLink[51]

$$pre\_score = Hyper(X \geq x) \cdot Norm(R \geq r) \cdot Taglen(T \geq t) \qquad (2.6)$$

with x being the number of matched fragments, r being the mean intensity ranking of matched fragments, t being the length of the longest sequence tag divided by the full sequence length. The three components of the equation are p-values for a random match having the same or better values in these three metrics based on a hypergeometric distribution of random fragment matches, a normal distribution for intensity ranks and a Monte-Carlo simulated precomuputed table for random sequence tag lengths.

The final refined score is the Kernel Spectral Dot Product (KSDP),[52] which was extended for XL-MS data. The SDP is equivalent to a cross-correlation, while the KSDP includes correlations between different fragment types to add additional weight to consecutive fragments and for example b- and y-ion pairs from the same fragmentation event.

For FDR control the top hits for each spectrum are separated into the groups of intra- and inter-protein cross-links, as well as loop-linked, mono-linked and linear peptides. They are re-ranked by a semi-supervised machine learning algorithm similar to Percolator[42 43] and filtered by a set FDR threshold.

Summary of available algorithms

Additionally to the details described above, every one of these tools uses binned spectra for their analysis. This speeds up the matching and scoring of spectra, but can cause inaccuracies in how the set fragment mass tolerance is applied in some cases. Of these tools xQuest and Kojak are the only ones usable from a command line interface, which allows for containerization using e.g. Docker, building pipelines with scripting languages like bash or python and embedding of the tools in third party tool and workflow managers like Galaxy or NextFlow. Additionally only XiSearch fully supports standard HUPO-PSI file formats for input and output. All the other tools output their results in simple non-standardized table formats.

2.2 Software and Libraries

2.2.1 OpenMS

OpenMS[53] is a framework for mass spectrometry data analysis. It contains a library of classes and algorithms for the processing of mass spectrometry and protein sequence data. It also contains the OpenMS proteomics pipeline[54] (TOPP), a set of more than one hundred executable tools that are based on the OpenMS library. Some tools can be chained together into more complex workflows and that enable the analysis of very diverse types of experiments, while others were developed specifically to handle a certain experimental protocol. The library and tools are all written in C++ and the project currently supports C++17. OpenMS has been developed as an open source project since the very beginning in 2006 by the Freie Universität Berlin and Eberhard Karls Universität Tübingen. It is licensed under the 3-clause BSD license and available for free to users and software developers. OpenPepXL was developed as a TOPP tool to make use of some of the data structures and algorithms in the library that have been maintained and optimized for more than a decade. OpenMS itself makes use of other open source libraries to handle specific file formats or to reuse optimized algorithms.

Boost and Sorting

The Boost library[55] is used in OpenMS for their many useful functions in mathematics and statistics. It also contains general use algorithms like sorting and efficient containers. In OpenPepXL Boost algorithms were applied for the cumulative distribution function of the binomial distribution to optimize the behavior of the OpenPepXL match score. Sorting algorithms are commonly needed in software development and their efficiency can be vastly influenced by the type of data that has to be sorted. The Boost library contains multiple such sorting algorithms with different optimal and worst-case conditions. Most of them were compared to find the optimal algorithm to sort the peaks of spectra by their m/z-values.

Qt for GUI Development

Qt is an application development framework written in C++. It is used in OpenMS for all tools with a GUI. Qt extends C++ with signals, slots and other features that enable event-driven programming, which is required for a responsive GUI design. TOPPView, the TOPP tool that visualizes MS data and the results of many TOPP tools including OpenPepXL, uses Qt extensively for the visual elements and control of the tool.

OpenMP Loop Parallelization

The OpenMP[56] application programming interface (API) allows to write multi-threaded code with shared memory. This is usually used to iterate over objects in an iterable data structure in parallel, by writing parallelized loops. This is necessary to make use of modern multi-core CPU architectures to increase the efficiency of computation and is used heavily in OpenPepXL and many other TOPP tools.

FeatureFinderMultiplex

The FeatureFinderMultiplex is a TOPP tool that can identify feature pairs with a specific mass difference in an MS1 map. It is a general tool that can be applied whenever labels that change the masses of molecules in MS1 spectra are used in an experiment. In the context of XL-MS it is used to identify and link the two MS1 features and the resulting MS2 spectra from the same cross-linked peptide pair with the two different cross-linker masses resulting from isotopic labeling.

2.3 File Formats

There are two major file types necessary as input for most MS database identification tools. These tools match mass spectra to protein sequences, so they need spectra and sequences as input files. In proteomics the FASTA format is used for protein databases almost everywhere, but for the MS input and the output data of tools many different formats are used. In the field of proteomics some of the most important formats were developed by the Human Proteome Organization's Proteomics Standards Initiative (HUPO-PSI). These formats include mzML for MS spectra and mzIdentML for peptide and protein identifications. These formats are designed to contain a lot of additional meta-information to ensure provenance and reproducibility.

2.3.1 MS File Formats

The raw data acquired by MS instruments is usually stored in vendor specific, proprietary formats that require specific libraries to parse them. Besides these there are open formats of varying complexity and standardization available. The conversion from a vendor specific format to an open common format is the first step in many MS data analysis workflows. There are simple formats like the Mascot Generic Format in which spectra are stored as a simple tab separated tables with two columns for m/z and intensity. However, most vendor specific formats contain a lot of additional information about the instrument settings and behavior during data acquisition and simple text based formats are inefficient in terms of storage space and reading efficiency. The currently most popular and best supported format is mzML.[57] It is an XML based format with a well defined but flexible structure and its own controlled vocabulary. It stores most of the auxiliary information as plain text but the spectra themselves are stored in a more efficient binary format and can also be compressed. This file format can also point to the original raw instrument files and keep track of processing steps applied to the spectra. MzML is supported by most current MS software and is also the primary input format for mass spectra in OpenMS and is therefore also supported by OpenPepXL. All other XL-MS tools described and used in this thesis also support it, except for xQuest. Another popular format is mzXML,[58] which is the alternative used by xQuest. It has a similar structure to mzML, but supports fewer types of data and metadata.

2.3.2 Identification Result File Formats

To be able to effectively make use of the results of a protein or cross-link identification tool, they have to be stored in a file format that captures all the relevant data in a way that is preferably readable by people and by other tools that are supposed to make use of the information. Many protein and XL-MS identification tools simply write out unstandardized tables that only capture data about the matches. MzIdentML[59] and mzTab[60] are two of the most widespread standardized result formats for proteomics identification tools. They are often used in tandem for different purposes. MzIdentML is a complex XML based format that supports a lot of metadata about the spectra and protein database input files, as well as the applied search tool and its settings. It is mainly used as a long term storage format to document results and the procedure used to arrive at them. However, because of its complexity it is not easy for people to read it directly. MzTab is a simpler tabular format that is meant to be easier to read, but it captures less information. The PRIDE[61] database and other databases in the ProteomeXchange[62] consortium collect data from many different types of proteomics and metabolomics experiments including the results of many different data processing pipelines. To be able to make use of that data later on and to potentially combine multiple datasets, they have to be stored in well defined ways. Therefore, standardized file formats like mzIdentML are required to make such data repositories useful. OpenMS supports these formats, but also uses the internal idXML format to forward results between different TOPP tools. IdXML is similar in structure to mzIdentML, but simpler and less redundant, because it only has to support the types of results that are already implemented in OpenMS so far.

2.4 Tools for Structural Validation and Visualization

To validate cross-links using known protein structures for benchmarking purposes, the cross-links have to be mapped onto a 3D model. The distances between the linked residues have to be calculated and the cross-links have to be visually represented as well. For the visual representation we used Chimera[63] with the Xlink Analyzer plugin.[64] Chimera can be used to render protein structures and the Xlink Analyzer plugin calculates euclidean distances between linked residues and visualizes the links as straight lines between C_α atoms. Euclidean distances can be very inaccurate, because the cross-linker has to span the distance between the linked residues along the surface of the protein. The Xlink Analyzer plugin was only used for visualization, because the straight lines look clean and are easy to see and understand. To get more accurate numbers for the actual validation of the cross-links, we used TopoLink.[65] It calculates topological distances between Cα or Cβ atoms along the surface of a protein structure.

Chapter 3

Algorithms and Implementation

3.1 Algorithms

3.1.1 Introduction

Given the state of existing tools in this domain as described in Chapter 2, new tools are needed that support standardized file formats and are deployable on various computing infrastructures. Additionally, many existing tools in this domain use heuristics, filters, or other shortcuts simply because the large search space requires a lot of CPU time and computer memory to search through. One goal of this study was to avoid that as much as possible and see how effective XL-MS identification could be if we can brute force our way through the entire square search space. This requires an efficient implementation that can handle this task for datasets of a size that reflects real-world research.

3.1.2 OpenPepXL Overview and Design Intentions

OpenPepXL is one of the algorithms that score an entire candidate molecule of two covalently cross-linked peptides against an experimental spectrum without pre-scoring linear peptides by doing an open-modification search. It applies an exhaustive strategy that tries to consider as many potential candidate matches as possible. In this sense it has more in common with xQuest[31] and StavroX[35] than with pLink2,[48] Kojak[37] or XiSearch.[44] OpenPepXL also has very lenient filters for the spectra. The main goal of this exhaustive approach is to not throw away potential matches before actually scoring them. This is done with the goal of keeping the sensitivity as high as possible. Most of the algorithms discussed in this chapter have the purpose of doing this exhaustive search as efficiently as possible and ensuring specificity.

An overview of the workflow is shown in Fig. 3.1. The analysis starts with reading the raw spectra from an mzML file and the protein database from a FASTA file. The spectra are deisotoped and filtered to reduce the noise. The protein sequences are split into peptides with modifications and stored in a permanent list. After that, the algorithm loops over the spectra and enumerates all possible peptide combinations that would fit each spectrum's precursor mass. The peptide pair and modified peptide candidates are fragmented *in silico*. These theoretical fragment spectra are compared and scored against the experimental spectra. The matches for each spectrum are ranked according to this score and the top-ranked match is considered as the final resulting identification for this spectrum. These 1st ranked cross-link-spectrum-matches (CSMs) are reported in the written-out result files of OpenPepXL. An additional tool called XFDR then takes these results as its input and sorts all those CSMs by their score. Using the target-decoy approach the False Discovery Rate (FDR) is calculated for each match. Now the final results can be filtered by a chosen FDR threshold. In the end, the decoy CSMs are removed and the remaining target CSMs below the FDR threshold are reported as the final search result.

3.1.3 Preprocessing of Spectra

Deisotoping and Filtering

As a first step, all spectra are normalized to a standard intensity scale to reduce the effect of absolute fragment intensities on the scoring. Fragment intensities are divided by the maximum peak intensity of each spectrum resulting in a normalized intensity between zero to one.

The deisotoping algorithm has the purpose of summarizing the information of an isotopic pattern into one peak. This reduces the total number of peaks and therefore reduces noise and makes the other algorithms more efficient. It also gives us another piece of information about the peaks, namely their charge state. It starts with a pointer to the first peak and searches for the next one in the pattern that is supposed to be 1.0033548378 u higher than the first. That number represents the difference between ^{12}C and ^{13}C. If it finds a second peak, it looks for a third one with the same distance to the second. If there is no next peak, and the current pattern has at least three peaks, only the first peak is kept and the rest is removed as noise. Because we are working with experimental data here, the matching of peaks allows for a mass tolerance and does not have to match exactly. This is generally true for any comparison of masses that involves at least one experimental value. After the search for a pattern is finished, the first pointer is moved to the next peak that was not part of any pattern so far and the algorithm is repeated starting from this new peak. This is then also repeated multiple times for different charge states, because the mass difference is divided by the number of positive charges. When a pattern is found with a specific assumed charge state, we now also know the charge state of the monoisotopic peak. This information is used in later steps as a criterion when matching to other masses to improve the match specificity.

For normal peptides it is usually assumed that the monoisotopic peak with only ^{12}C has the highest intensity, the second peak has the second highest intensity and so on. Therefore the monoisotopic peak can be enough to represent the whole pattern by itself. The decreasing intensity is also used as a criterion for including the next peak in the pattern. For cross-linked peptide fragments that is not always true. Because of the higher average mass of these fragments, the second peak can be higher. So the criterion of a decreasing intensity does not always apply and the monoisotopic peak does not always represent the intensity of the whole pattern. Therefore, a few additions were made to the algorithm specifically for XL-MS spectra. One was to start using the decreasing intensity criterion from the third peak in a pattern, in case the second peak has a higher intensity than the first. And the second function was to sum up the intensities of the peaks in the pattern into the monoisotopic peak before removing the other peaks. This way the intensity of the monoisotopic peak represents the intensity of the whole pattern better. Peaks that do not fit into any pattern are left as they are, as sometimes the pattern is not detected by the instrument and we only have one or two peaks for a fragment in the raw spectrum.

In the next step, the spectra are filtered by keeping only the 20 highest intensity peaks in a jumping window of 100 along the m/z axis. This will keep at most 400 peaks below 2000 m/z and there are barely any above that. This is again done to remove low-intensity noise and reduce the number of peaks, because that is a major factor in the efficiency of the matching and scoring algorithms. The summing up of intensities during deisotoping is also meant to give those deisotoped peaks more weight through having a smaller chance of being filtered out at this step. This filtering method is common in mass spectrometry, because the middle of the spectrum generally has more high-intensity peaks and many in the lower and higher m/z ranges would be filtered out, if we just kept the 400 most intense peaks in the entire spectrum. Filtering evenly across the spectrum ensures the full range of m/z values is kept and the peptide sequences are covered as well as possible.

Spectra that contain less than two times the minimal peptide size of peaks after these steps are not analyzed further. The default minimal peptide size is five, so usually entire spectra with less than ten peaks are removed. Even if all the peaks matched theoretical fragments, with so few peaks it would be realistically impossible to cover so there is no point in spending more time analyzing these.

Spectrum Matching

The spectrum matching algorithm is used whenever two spectra need to be compared to each other. It returns a list of pairs of peaks. Each pair represents two peaks that were matched between the two spectra, because they have the same m/z value and fit some additional criteria. In the context of OpenPepXL that is needed at three different steps that will be touched upon later.

The peak matching algorithm to match two spectra in OpenPepXL is a linear sliding window algorithm that moves along the peaks in the first input spectrum. For every peak in the first spectrum, the algorithm searches for a second peak with the same m/z in the second spectrum. At least one of the two spectra is always an experimental one, so the second peak is looked for within a small tolerance window. At first the peak with the smallest mass in that tolerance window is checked for the correct charge. If the charge of both peaks is known, we can enforce that it matches at this step. If the charge for at least one of them is not known, then we ignore the charge state. If the charge does not match, the next peak in the second spectrum is selected, until it reaches the end of the tolerance window or a peak with a matching charge is found. When a peak matches, then it is kept as a potential match. Optionally the algorithm also allows for the additional criterion of having a similar intensity. This is defined as having a ratio of more than 0.3 between the lower and the higher of the two intensities. This is a lenient filter, but if one of the peaks has one of the highest intensities in its spectrum, it will not be matched to a peak that is among the low-intensity noise peaks in the other spectrum. Even after finding a match within the m/z tolerance window with a matching charge and intensity, the search does not necessarily stop. The rest of the peaks within the tolerance window in the second spectrum are still checked to find the matching peak that has the smallest m/z difference to the peak from the first spectrum.

Preprocessing Stable Isotope Labeled XL-MS Data

Additional preprocessing of the spectra differs between the workflows for labeled and label-free cross-linkers. For labeled cross-linkers an additional consensusXML file produced by the TOPP tool FeatureFinderMultiplex is read in that contains a pairing of MS1 features. These are used to link together fragment spectra with the light and heavy versions of stable isotope labeled cross-links. In the case of multiple fragment spectra being assigned to one feature, all possible combinations between one light and one heavy fragment spectrum from the two features are considered as separate pairs. In the end the top-ranking matches to all these spectrum pairs are summarized into one ranked list per light spectrum. The purpose of the algorithm described next is to use the spectrum with the heavy isotope label to filter peaks from the spectrum with the light label and to gain additional information about each remaining peak. The heavy isotope labeled spectrum itself is not used afterward.

The fragment spectrum pairs are combined by first applying the spectrum matching algorithm with the matching intensity ratio rule turned on. Both of these spectra are experimental and should contain fragments from the same peptides, just with different cross-linker masses. Therefore the fragmentation pattern and peak intensities of fragments without a cross-link should be comparable, while fragments with the cross-linker should not match. The peak matching at this step can also consider the charges discovered through deisotoping to increase the specificity for those peaks where charge information is available. The matched peaks from the light spectrum are stored in a new spectrum representing the linear peptide fragments without cross-linkers. All matched peaks are also removed from the original spectra before the next step. The entire remaining heavy spectrum is now shifted by the mass difference of the light and heavy labeled cross-linker and the spectrum matching is done again. This mass shift also depends on the charge and is done multiple times to represent the different possible charge states. Every new matching peak is a fragment that most likely contains a cross-link. These matching peaks are summarized over the multiple charge states into one new spectrum representing the cross-linked fragments. This process is summarized in Fig. 3.2.

For those peaks where the charge could not be determined by deisotoping, the assumed charge for the mass shift of the heavy spectrum is set as the known charge. This way charge determination from multiple sources is combined and used to filter out false positive peak matches later on, since we also know the expected charge for every theoretical peak. For the label-free algorithm only the charge information from deisotoping can be considered.

3.1.4 Candidate Enumeration

OpenPepXL keeps a list of all linear peptides and their modified variants with their masses after *in silico* digestion of the protein database. This list is sorted by their masses. The digestion is done according to the known protease cleavage rules of the set protease, usually trypsin. For each spectrum a list of peptide pairs has to be enumerated. The correctly matching pair of cross-linked peptides has to have the same total mass as the precursor mass of the fragment spectrum. This is also one of the two most critical steps for the entire tool's efficiency and sensitivity. The enumeration has to be done as quickly as possible, but also has to be complete. For the α-peptide we already have some restrictions we can place upon its mass. Denoting the mass of any molecule with $M(molecule)$, it cannot have a mass higher than $M(precursor) - M(cross\text{-}linker) - M(\beta\text{-}peptide)$ and we can plug in the peptide with the smallest mass as the β-peptide to set a ceiling for the mass of the first peptide. The enumeration of peptides for the α-peptide in a pair will now iterate from the smallest peptide in the list up to this ceiling. The settings for the digestion step allow for missed cleavages that can result in very large peptides. Using this mass ceiling many of those will never be considered, if there is no large enough precursor mass in the data. This makes the tool quite robust, as very lenient and ambitious digestion settings for the tool will still be constrained by the actual data and not cause the efficiency to suffer needlessly. After the first peptide of a pair is fixed, the second peptide must have a mass equal to $M(precursor) - M(\alpha\text{-}peptide) - M(cross\text{-}linker)$ with a small error tolerance. Now a quick binary search is used to find the positions of the lower and upper bounds of this tolerance in the peptide list. The candidates for the β-peptide have to be within these bounds and iterate from the lower to the upper bound. When a suitable β-peptide is found, an object representing the cross-linked peptide pair is created and saved in a list of candidate cross-links for the current spectrum. The process is summarized in Fig 3.3.

Although we can set some hard limits on the possible combinations of peptides during the enumeration step, the resulting list of candidate peptide pairs for one spectrum can be longer than the list of peptides from the entire protein database. This is due to the large mass range possible for the α-peptide and the nature of the squared complexity of this enumeration. Therefore this list is only kept temporarily until the analysis of the current spectrum is finished and then discarded. The way the list is constructed also ensures that no memory is wasted. Instead of making new structures with the entire peptides, a linked peptide pair is represented by two indices pointing towards the peptides in the list of digested. That way the complex structures representing the modified peptides only have to be stored once. The structure of the peptide pairs additionally contains the assumed linked positions on the two peptides and some additional necessary information, but it is designed to use as little memory as possible while allowing quick access to all information necessary to generate the theoretical spectrum of the cross-linked peptide pair. At this step mono-linked and loop-linked peptides are also enumerated, but they only have one peptide each. Their enumeration is more straightforward and is done similarly to the iteration over β-peptide candidates for a fixed α-peptide within a small mass tolerance window.

Sequence Tags

During the enumeration, there is also an additional optional filtering step using sequence tags. They are short segments of peptide sequences that are generated in linear time from the experimental fragment spectrum. The algorithm works similarly to the deisotoping algorithm. It starts at the first peak of the fragment spectrum and looks for another peak with a specific distance to the first one, but this time there are many possible distances to look for. The masses of all 20 canonical amino acids, also considering multiple charge states from +1 to the maximal set precursor charge (usually +7), and also taking into account variable modifications of residues. After a matching peak has been found, the one-letter-code of the corresponding residue name is saved and the next peak is looked for. This algorithm has a high complexity, but we do not need to look for sequence tags of indefinite length, but rather only a fixed size. Looking only for tags with a length of two keeps the time for the search low, but even such short tags can often be used to filter out more than half of the peptide pair candidates for a spectrum. The filtering is done by applying an additional criterion for accepting a peptide pair: at least one of the peptides has to have one of the sequence tags as part of its sequence.

Sequence tagging often gets rid of decoys and false hits that otherwise would have negatively impacted the FDR calculation. For this reason sequence tagging is known to boost the sensitivity of linear peptide searches most of the time, when done correctly.[67] Before testing this algorithm in OpenPepXL, we expected the sensitivity would drop most of the time. The reason is the fact that the segments of peptides that could potentially be sequenced this way are much shorter for cross-linked peptides compared to linear ones. The mass difference between a normal residue and a neighboring cross-linked residue suddenly jumps by the mass of the cross-link and the entire second peptide. Without knowing the mass of one of the peptides beforehand it is impossible to calculate that mass difference. Any part of the sequence that overlaps with a cross-linked residue can therefore not be sequenced by sequence tagging. The actual results of this algorithm varied between different datasets. In our experiments it always reduced the total runtime significantly. In more cases than expected it actually boosted the sensitivity of OpenPepXL as well, but that is not guaranteed and one should generally expect a decrease in sensitivity when using this option. Due to both its potential benefits and unpredictability we made this filtering step optional. It can be very useful to speed up repeated runs while optimizing other search parameters or as a quick check whether there is anything useful to be found in a new dataset before doing a full exhaustive search.

3.1.5 Theoretical Spectrum Generation

The generation of theoretical spectra is the second of the two most critical algorithms for the efficiency of the tool. As mentioned before the list of peptide pair candidates for an experimental fragment spectrum can be extremely long. Now at this step the tool has to calculate a theoretical fragment spectrum for each of those candidates. This involves calculating the masses of all possible fragments from the two peptides. This algorithm has two main parts. First, all the masses have to be added as peaks into a new spectrum. And then this list of peaks has to be sorted. The first part is straightforward and the only way to make it as efficient as possible is to precalculate the masses of as many building blocks as possible to simplify the math. The masses of the 20 residues and the mass differences between a-, b-, c-, x-, y-, and z-ions, as well as the masses of the two most common neutral losses H_2O and NH_3 were precalculated. Other losses were not considered, because after some testing no other losses were seen consistently enough to make a difference in the scoring and to keep this algorithm simple.

For linear peptide fragments the algorithm starts with the mass of the first N-terminal or C-terminal residue mass, creates a peak with that mass, and adds it to the spectrum. For the next peak the mass of the next residue is added to the previous mass and another peak is created and added to the spectrum. During this the type of the new residue is checked for possible neutral losses and a simple number flag represents the state. A 0 for no losses, 1 for a possible H_2O loss, 2 for NH_3 and 3 for both being possible. According to this number additional peaks with the respective masses subtracted from the fragment mass are created.

This continues until the cross-linked residue of the peptide is reached. For cross-linked fragments the algorithm works backward by starting with the precursor mass and subtracting residue masses one by one. This makes it simpler as both algorithms only need to know the sequence of one peptide and the cross-linked position. Neutral loss peaks are also added in the same way as already described. All of those peaks have their monoisotopic mass, but as mentioned before cross-linked fragments sometimes have a higher intensity on their second isotopic peak and the deisotoping algorithm does not deisotope patterns with less than 3 peaks. In some cases the less prominent monoisotopic peak is not detected or filtered out as noise. For those cases a second isotopic peak is added for every peak type. Although this doubles the size of the spectrum and the complexity of most of the following algorithms, it significantly improves the sequence coverage for cross-linked peptide pairs.

The last major step is to sort the spectrum. Algorithms for sorting numbers have been a staple of computer science since the very beginning and new ones are still being developed. That is because these sorting algorithms can have widely different runtime behaviors depending on the initial state of the list to be sorted. In our case we have multiple patterns in the data. There are partially presorted linear fragment spectra, where peaks were added with increasing masses. Then we have the cross-linked fragment spectra with decreasing masses. We have tried out several different algorithms and approaches. One of them was to reverse cross-linked spectra, because partly presorted lists tend to be easier for most sorting algorithms. It turned out that the time spent on reversing the spectra made this approach one of the slowest we tried. In the end we settled on the Pattern-Defeating-Quicksort (pdqsort) algorithm from the Boost library. It was the most efficient for both the linear and cross-linked fragments. On our test data it was on average about 20% faster than the standard std::sort algorithm. It works mostly like the quicksort algorithm, but tries to detect patterns that would be a bad fit for quicksort and applies heapsort to these patterns. Therefore it has the benefits of quicksort while avoiding its worst-case scenarios.

3.1.6 Match Score: Developing the Formula

OpenPepXL started out with our own implementations of the scores from xQuest, since it still is a successful tool for its purpose. The match-odds score will be described in detail later. The ion intensity ratio score represents the percentage of the experimental spectrum intensity that was matched to peaks from the theoretical spectrum. The cross-correlation score is a measure of the similarity of two spectra and is computed as a sliding dot product. It is a reimplementation of the fast SEQUEST cross-correlation[68] used ubiquitously in peptide identification. Additionally to the scores used in xQuest we implemented and tried multiple additional scores. X!Tandem's[69] HyperScore combines the dot product of the cross-correlation score with a count of matched b- and y-ions. The AScore[70] was developed for the localization of phosphorylation sites and we tested whether it could be applied to cross-linking sites as well.

Additional types of data were also collected in attempts to incorporate them into the scoring. The sizes of the isotope patterns from deisotoping were kept for each monoisotopic peak. The idea was that larger patterns could be a sign of better spectrum quality and validation for the deisotoping working correctly. Another type of data was the mass errors for each matched fragment peak. The idea here was that even though a certain error is allowed for a match (usually 20 ppm), most correct matches would have a mass difference much closer to zero than the maximally allowed tolerance. We tried using the average of these matched peak errors as an additional match quality measure. The same idea also applies to the tolerance associated with the precursor mass (usually $5-10$ ppm). The precursor is also another interesting target because of the isolation window for fragmentation. That window with a usual size of 1.0 Da is much wider than the tolerance window. That means everything within this window is fragmented at once and measured in the same fragment spectrum. This is done to capture multiple isotopes of the fragmented molecule, but can also lead to co-fragmentation of additional entities. We implemented an algorithm to look into the isolation window of the precursor peak in the MS1 spectrum. It identifies the actually selected precursor peak and its isotope pattern. The intensity of these is summed up and divided by the total summed up intensity within this window to calculate a ratio for how much of the spectrum's intensity is expected to come from the actual precursor molecule.

We tried to integrate this data into the scoring and to combine multiple scores together. We discovered that xQuest's match-odds score was consistently outperforming all of the others and most combinations only made the final results worse. xQuest itself combined a few scores, but it was developed and optimized for older ion trap instruments and was mostly used with absolute fragments mass tolerances of about $0.1-0.3$ Da most of the time. On higher resolution data where we mostly use relative fragment tolerances around $10-20$ ppm the other scores xQuest used turned out to be less useful. Scores like the HyperScore and AScore work well with high-resolution data, but were slightly worse than the match-odds score, at least specifically for XL-MS. Combining these well performing scores also did not result in any benefit, since they correlate very strongly and do not provide additional information. Incorporating more data like the isotope pattern sizes and fragment match errors was also unsuccessful, aside from one exception. The precursor mass error turned out to be the only measure that improved upon the results of the match-odds score. Coincidentally this worked mostly in our favor, because from all the scores we tried the match-odds score and the precursor mass error are among the fastest measures to compute. The most computationally intensive part of the score is applying the spectrum matching algorithm discussed before.

The original xQuest match-odds score was like the rest of the tool developed with lower-resolution ion-trap data in mind. And although it worked quite well from the beginning, some adjustments were made before it became the most successful score in our comparisons. Because we also felt that the name did not actually fit the way it is calculated, we called our own version the log-occupancy score.

Log-Occupancy Score

The log-occupancy score represents the probability of a match between a peak from the experimental spectrum and a peak in the theoretical spectrum by random chance. For this these parameters are used: the mass tolerance window tol, the number of peaks in the theoretical fragment spectrum s, the mass range of the theoretical spectrum r, the number of considered charges for all theoretical peaks c and most importantly the number of matched peaks between the two spectra k. The probability of one such match is calculated as:

$$p = 1 - (1 - \frac{2 \cdot tol}{\frac{1}{2}r})^{\frac{s}{c}} \tag{3.1}$$

To calculate the probability of matching more than k peak pairs by random chance, the cumulative distribution function of a binomial distribution with sample size s and probability p is used:

$$P(X > k) = \Sigma_{i=k+1}^{s} \binom{s}{i} p^i (1-p)^{s-i} \tag{3.2}$$

For higher numbers of k, this probability tends toward zero, which is associated with a good match, since it is unlikely to have happened by chance. Using a negative logarithm, the probability is turned into a score with higher and simpler to process numbers for the better matches:

$$lo = -log(P(X > k)) \tag{3.3}$$

This log-occupancy score lo is combined with the precursor error pe, which is the difference between the theoretical and experimental precursor masses, in the following formula to get the final OpenPepXL score:

$$score = 0.2 \cdot log(10^{-7} + lo) - 0.03 \cdot |pe| \tag{3.4}$$

A linear regression and a linear discriminant analysis were applied on several XL-MS datasets to calibrate the weights of the two components.

Mathematically the log-occupancy score is identical to the original match-odds score from xQuest, but the original implementation in Perl has a few numerical issues. With the smaller tolerances of high-resolution data p also became much smaller in most cases. A straightforward computation of the cumulative distribution function hit on numerical limits that caused the score to have a low dynamic range. $P(X > k)$ had the same four low values for many different spectrum matches and a maximal value was reached for many good matches. It seemed like the sum of many small ps would always land on the same few values. To work around these issues we used the Boost library's implementation of the cumulative distribution function and used its complement version. This function, therefore, computed the complement of the cumulative distribution:

$$P(X <= k) = \Sigma_{i=0}^{k} \binom{s}{i} p^i (1-p)^{s-i} \tag{3.5}$$

In our case this formula has to deal with less extreme values and the result can be easily transformed into what we need with the formula:

$$P(X > k) = 1 - P(X <= k) \tag{3.6}$$

This gave the score a much smoother distribution and a better match always results in a higher log-occupancy score.

3.1.7 Scoring and Post-Processing

The generation of theoretical spectra and the scoring is done in one simple loop over the list of cross-link candidates for one experimental spectrum. Four separate theoretical spectra are computed as needed: the spectrum containing linear fragments from α-peptide fragmentation, the spectrum containing cross-linked fragments from α-peptide fragmentation, and the same two spectra for the β-peptide. For labeled cross-linkers, the linear theoretical spectra are matched to the linear fragment spectrum and the cross-linked theoretical spectra are matched to the cross-linked fragment spectrum. For label-free cross-linkers all four theoretical spectra are matched to the whole experimental fragment spectrum. The log-occupancy score is computed based on the number of matched peaks for each of these matches and the average of all four is the total log-occupancy score. The precursor mass error and following that the final score is computed for the current candidate as described above. The scored candidates are sorted by the score and the best match is kept. Afterward, in an additional loop over all spectra more information is computed for each of those best matches. This includes fragment peak annotations for visualization and some of the additional match quality metrics that were mentioned before, but did not make it into the final score. They are still useful for manual inspection of the results and visualization.

3.1.8 False Discovery Rate Estimation

For FDR calculation we implemented the unchanged formula from xProphet[33] into the OpenMS tool XFDR. The algorithm separates the hits into intra- and inter-protein cross-links, as well as an additional class for mono-links (or dead-end links) and loop-links. Hits to the target and decoy versions of the same protein are considered intra-protein cross-links. The FDR is then computed separately for these three groups, because their score distributions are different from each other.

The formula uses the counts of target-target (TT), decoy-decoy (DD), target-decoy (TD) and decoy-target (DT) hits:

$$\frac{(TD + DT + DD) - 2 \cdot DD}{TT} \tag{3.7}$$

The formula assumes that all four categories have equal score distributions for random incorrect matches, but there are three decoy categories and only one target category. The standard target-decoy approach in peptide identification only has one decoy and one target category. It can be illustrated with the simple formula $\frac{D}{T}$. This approach assumes that, among the random incorrect matches, the score distributions of targets and decoys are equal and therefore at any score cut-off there are roughly the same number of targets and decoys. Using this formula directly for the four XL-MS categories would result in a three-fold overestimation of the number of false matches. The subtraction of $2 \cdot DD$ is used to normalize the number of decoys to match the number of targets under the assumption of only random scores.

3.1.9 Parallelization

The list of enumerated candidates for each spectrum is a major contributor to OpenPepXL's memory requirements. It is also the major reason behind an important decision about the parallelization of OpenPepXL. The tool has two main loops. One loop over all spectra and another nestled loop over all the peptide pair candidates for the current spectrum. The outer loop over all spectra in OpenPepXL is not parallelized, otherwise the tools would need to keep an additional list of candidates for each additional thread. Having to keep multiple candidate lists would practically multiply the memory requirements by the number of threads and make the memory requirements for multi-threaded operation prohibitively large. Our experiments have shown that parallelizing the entire outer loop would be slightly more efficient in terms of runtime, but we felt that it was not worth it. Most parts within the outer loop itself were parallelized. The enumeration of candidates and the loop to score the candidates are all done in parallel.

The preprocessing of spectra and the post-processing of the results are also done on their own parallel loops. This means there is more overhead for starting and finishing multiple parallel loops in sequence, but the memory requirements for multiple threads stay almost as low as when using only one thread.

3.2 Runtime and Memory Complexity

The runtime complexity of OpenPepXL and OpenPepXLLF is mostly determined by the two types of input data. Assuming the MS data only contains one MS2 spectrum and its precursor mass, the precursor correction setting is turned off and we have a fixed precursor mass tolerance, we can approximate the runtime complexity based on the database size. The runtime to score one candidate peptide pair against one MS2 spectrum is approximately constant and negligible on its own. The spectrum alignment and scoring have a fixed upper ceiling through pre-filtering spectra to a maximal number of most intense peaks. The complexity of theoretical spectra is limited by the highest reasonable precursor masses. We will assume that the precursor mass is within the normal range of peptide pair masses in the database. It follows that the fraction of the peptide pair search space that falls into the precursor mass tolerance window should be roughly similar for any size of the input database. This means the search space for this MS2 spectrum increases at the same rate as the entire search space of the full database. Let p be the number of peptides in the full list after in silico digestion, including modified peptides. The full search space of modified peptide pairs is then p^2. The specific search space of one MS2 spectrum is a small but constant fraction of that search space. The fraction of the search space that applies to each spectrum can be controlled by the precursor mass tolerance, but is always a small fraction between 0 and 1. If we now increase the number of spectra, each spectrum adds its own precursor mass tolerance window of the same size and has on average the same search space size as any other spectrum. Therefore the runtime complexity increases linearly with the number of MS2 spectra n. The method of precursor mass correction implemented in OpenPepXL allows the user to set additional fixed tolerance windows shifted by the mass difference between ^{12}C and ^{13}C. Each of these windows adds a search space equivalent to an additional spectrum to each spectrum in the input data and therefore multiplies the runtime by the number of set windows. However, it is not reasonable to set this to a number above the single digits and is therefore negligible in the overall theoretical complexity calculation. The ability to reduce the search space by using isotopic labeling simply decreases the number of MS2 spectra that needs to be searched to a small fraction of the entire MS data, but this is again always a fraction between 0 and 1.

The runtime complexity of OpenPepXL and OpenPepXLLF is therefore

$$O(n \cdot p^2) \tag{3.8}$$

with the number of spectra n and the number of peptides in the search database p.

Memory complexity scales similarly, since the current algorithm reads in whole mzML files and whole FASTA files at once. Although the enumeration of peptide pairs is done very efficiently, for very large databases most of the memory is still being used for the list of candidate peptide pairs for the currently analyzed MS2 spectrum, while the memory needed for the MS data and the list of single modified peptides becomes negligible. This could be improved with a batch-wise enumeration and processing of candidates by always keeping only a small subset of best hits per batch, but memory usage was never a bottleneck with using OpenPepXL so far.

To estimate how well this tracks empirically, the first mzML file of the ribosomal fraction dataset with a size of 771 MB and 53799 MS2 spectra was searched with OpenPepXLLF against multiple databases of various sizes, doubling in the number of proteins with each step, starting with 32 different proteins. The largest database has 512. However, the proteins have different lengths and the actual database size is not always doubled in this process. Table 3.1 contains the resulting runtime and memory usage with the FASTA file sizes in KB as a metric of actual database size differences. The analysis was run on a laptop with an Intel i7-11850H CPU using 8 threads. From the smallest database to the largest the number of proteins increased by a factor of 16 and the file size of the database increased by a factor of about 31. This means according to our formula the runtime should increase by a factor of about $31^2 = 961$. Between the smallest and the largest database the runtime increased by a factor of 406 and the memory usage by a factor of about 2. The jump in runtime by doubling the database was a factor of 4.75 on average, while the file size increased by a factor of 2.4 on average. This follows the expected trend of the squared search space very well.

3.3 Visualization in TOPPView

Visualizing the results is an important factor in making them useful. In XL-MS experiments it is not uncommon for the number of identifications to be in the dozens. Many experiments target a single purified protein complex. Therefore manual review and validation of CSMs is feasible and common. This needs to be enabled and facilitated by

Table 3.1: Runtime and memory usage of OpenPepXLLF on different database sizes. The database sizes are given as the number of proteins and the file size in KB. The last three rows show the factor by which the size, runtime and memory usage increases from the previous database.

Database	32 (10 KB)	64 (21 KB)	128 (53 KB)	256 (126 KB)	512 (311 KB)
Runtime (s)	101	604	2854	6888	41056
Memory (MB)	1767	2381	2816	3144	3500
DB size mult.		x2.1	x2.5	x2.4	x2.5
Runtime mult.		x5.9	x4.7	x2.4	x6.0
Memory mult.		x1.3	x1.2	x1.1	x1.1

the software. For this we implemented a visualization of CSMs for manual validation in TOPPView. By loading in an mzML file with spectra and an idXML or mzIdentML file with OpenPepXL identifications from that mzML file, a table with CSM data becomes accessible. The table contains all important information about the CSMs and one additional column called 'selected' with clickable checkboxes. If an idXML or mzIdentML file is written out again from TOPPView, this column is saved as boolean values and can be used to filter the data later. This way the checkboxes can be conveniently used to either accept or reject specific CSMs from the results. The table also shows the score and other match quality metrics and can be sorted by any of them. Clicking on any line in the table outside of the checkbox will open up the corresponding MS2 spectrum with the matched peaks highlighted and labeled with the description of the specific ion it was matched to. On the top right the two linked peptides are shown with symbols indicating where they are supposed to be linked and which parts of the sequences were covered by matching fragments. The peak labels can be moved around, edited, or deleted and custom labels can be added to peaks. The spectrum viewer also supports zooming into and scrolling along the spectrum. On the top left is a text field showing the precise intensity and m/z-value of the peak closest to the mouse cursor. The current view of the spectrum can be exported as a PNG image file. Examples are shown in Fig. 3.4 and Fig. 3.5. These features facilitate a detailed examination of a match and the preparation of figures for publications.

3.4 MzIdentML 1.2

MzIdentML is a standardized format for results of proteomics experiments. Version 1.1 supported many types of protocols with a main focus on peptide and protein identification. The specification defines multiple sections of the XML-based format that store detailed information in a way that is as little redundant as possible. The first major section covers peptide sequences with detailed information about residue modifications with their masses and positions. Each peptide has a unique ID number that can be used from other parts of the document to refer to the peptide. A second major section links these peptide IDs to the proteins the peptides are from. Each peptide can be linked to multiple proteins for ambiguous cases. The third major section describes PSMs. Multiple PSMs can be described for each spectrum. This section has detailed information about the spectra and links them to the identified peptides through peptide IDs. This way the detailed peptide entries can be reused if exactly the same peptide is identified multiple times. Standard file formats developed by the Human Proteome Organization's Proteomics Standards Initiative (HUPO-PSI) are designed to contain enough meta-information to ensure provenance and reproducibility.

We collaborated with the HUPO-PSI in the development and publication of the updated mzIdentML 1.2 specifications[59] that among other minor changes added support for the scoring of modification positions, some proteogenomics approaches and cross-link identification results. The cross-links are represented as two new types of peptide modifications. A cross-link donor modification with a position and the mass of the cross-link spacer and a cross-link acceptor with a position and a mass of zero. These modifications can be linked by a cross-link ID to represent the linkage of the two peptides. In the PSM section the two peptides being identified in the same spectrum are represented as two PSMs that are also linked by a cross-link ID. The reading and writing of cross-link identification results according to these specifications has been implemented into OpenMS and can be used for the output of OpenPepXL and also the input and output of XFDR. MzIdentML 1.2 is the only standardized identification result format for XL-MS data so far and there are already a few other tools available that support it, like the visualization tool XiView[47] and the FDR tool XiFDR.[44]

3.4.1 Support for CID-Cleavable Cross-Linkers

Support for CID-cleavable cross-linkers was implemented in OpenPepXL in the context of a Master thesis[71] by Ruben Grünberg under my supervision. This mainly required two additional steps in the workflow: The detection of the characteristic peak pairs from the fragmentation of cleavable linkers in MS2 spectra and the filtering of cross-link candidates by the peptide masses derived from those peak pairs. Additionally the theoretical spectrum was extended by full peptides and peptide fragments with the remaining fragments of the cross-linker after its cleavage as modifications.

To detect the peak pairs, the algorithm iterates over all peaks in an experimental spectrum and for each peak looks for a peak with a specific higher mass, . The higher mass is determined by the mass difference created by the asymmetric fragmentation of the cross-linker. This search is done in m/z space, so multiple charges are considered. Every found pair can represent either a full peptide or just a peptide fragment, because a fragmentation of both the cross-linker and the peptide frequently happens together. Multiple levels of specificity were implemented for this matching. The least specific mode assumed each pair to be a potential full α-peptide and uses the precursor mass to calculate a mass for the second peptide. The enumeration of peptide pair candidates is then constrained by the multiple masses of potential α-peptides additionally to the precursor mass. The stricter mode requires two pairs of peaks that correspond to two peptides with a combined mass that fits the precursor mass. This reduces the search space drastically, but can also impair sensitivity, since one of these four specific peaks can be missing.

3.5 Summary: OpenPepXL Features

OpenPepXL supports the current versions of Windows, macOS and Linux. It can be installed on most local and remote computing environments. It can be set up to search all labeled and label-free CID-cleavable and non-cleavable cross-linkers in tandem MS spectra. It makes use of preprocessing of spectra with labeled linkers to reduce the search space and denoise MS2 spectra, in a similar way to xQuest. OpenPepXL is to our knowledge the only tool that applies this processing alongside support for high-resolution orbitrap spectra, combining the benefits of this approach with the improved specificity of modern instruments. To allow the field of XL-MS to mature, more interoperability between tools and standardization of file formats are necessary. OpenMS supports many of the open file formats developed and standardized by the HUPO-PSI like mzML for raw MS data and the MS identification data format MzIdentML. OpenPepXL supports the XL-MS data extension of the MzIdentML 1.2 specification. The OpenMS Pipeline (TOPP) contains many tools for MS data processing and analysis, including several quantification methods and correction of monoisotopic peak assignment. OpenPepXL is fully integrated into the OpenMS ecosystem and can be easily combined with these tools to build complex workflows that are deployable in versatile ways. The MS data visualization tool TOPPView was extended for XL-MS data and can visualize the MS1 features on an MS1 map, peak spectra, the precursor isolation windows of MS2 spectra, annotations of matched peaks and the sequence coverage for both cross-linked peptides. The spectrum visualization is interactive with the capability to zoom and scroll along the spectrum. Peak annotations are generated by the search, but can be freely added, removed, edited and moved to aid in manual validation and preparation of images for publication.

Table 3.2 shows a list of desirable features for XL-MS identification tools and which tool which supports these features. Table 3.3 shows which parameters can be changed in each tool's settings.

Table 3.2: Feature comparison of the different tools. A √shows that the feature is either built in into the tool or part of the same integrated pipeline. A (√) shows a partially supported feature (xQuest has a toggle for a −1 *Da* precursor correction, but no option for −2 *Da* or more). M.p.c. = monoisotopic peak correction; a.m. = additional precursor masses; MS1 f.a. = MS1 feature analysis; CLI = Command Line Interface; CSM GUI = a GUI for interactive manual validation of CSMs; PSI formats = Open standard HUPO PSI formats (like mzML, MzIdentML); Non-PSI formats = Open standard non-PSI formats (mzXML, pepXML, MGF); MOS = Multiple OS support (Win, Lin, macOS). The tools and tool versions were OpenPepXL 1.1, Kojak 1.6.0, pLink 2.3.5, XiSearch 1.6.731, StavroX 3.6.6.5, and xQuest 2.1.3

	xQuest	**StavroX**	**Kojak**	**XiSearch**	**pLink2**	**OpenPepXL**
M.p.c. (a.m.)	(√)	√		√		√
M.p.c. (MS1 f.a.)					√	√
CLI	√		√		√	√
CLI documentation	√		√			√
Tool GUI		√	√	√	√	√
CSM GUI	√	√			√	√
PSI format in		√		√		√
PSI format out				√		√
Non-PSI format in	√		√		√	
Non-PSI format out			√			
MOS	√	√	√	√		√
Open Source Code	√		√	√		√
Labeled linker filtering	√					√
Labeled linker quant	√				√	√

Table 3.3: Search parameters accessible in each tool. The tools and tool versions were OpenPepXL 1.1, Kojak 1.6.0, pLink 2.3.5, XiSearch 1.6.731, StavroX 3.6.6.5, and xQuest 2.1.3. For xQuest precursor monoisotopic peak correction for an up to 2 Da difference can be turned on or off, but the maximal correction difference cannot be changed. The mass and RT tolerance for isotopic pair detection for OpenPepXL is not directly built into the tool, but can be adjusted in the tool FeatureFinderMultiplex.

	xQuest	**StavroX**	**Kojak**	**XiSearch**	**pLink2**	**OpenPepXL**
precursor mass tolerance	✓	✓	✓	✓	✓	✓
precursor charges	✓					✓
precursor mono-isotopic peak corrections	(✓)		✓	✓		✓
fragment mass tolerance	✓	✓	✓	✓	✓	✓
fixed modifications	✓	✓	✓	✓	✓	✓
variable modifications	✓	✓	✓	✓	✓	✓
max variable mods per peptide	✓	✓	✓	✓	✓	✓
enzyme	✓	✓	✓	✓	✓	✓
min peptide length	✓	✓		✓	✓	✓
min peptide mass		✓	✓		✓	
max peptide mass		✓	✓		✓	
max missed cleavages	✓	✓	✓	✓	✓	✓
cross-linker name	✓	✓	✓	✓	✓	✓
cross-linker mass	✓	✓	✓	✓	✓	✓
mono-link masses	✓	✓	✓	✓	✓	✓
linked residues	✓	✓	✓	✓	✓	✓
isotopeshift	✓					✓
isopair Mass tolerance	✓					(✓)
isopair TR tolerance	✓					(✓)

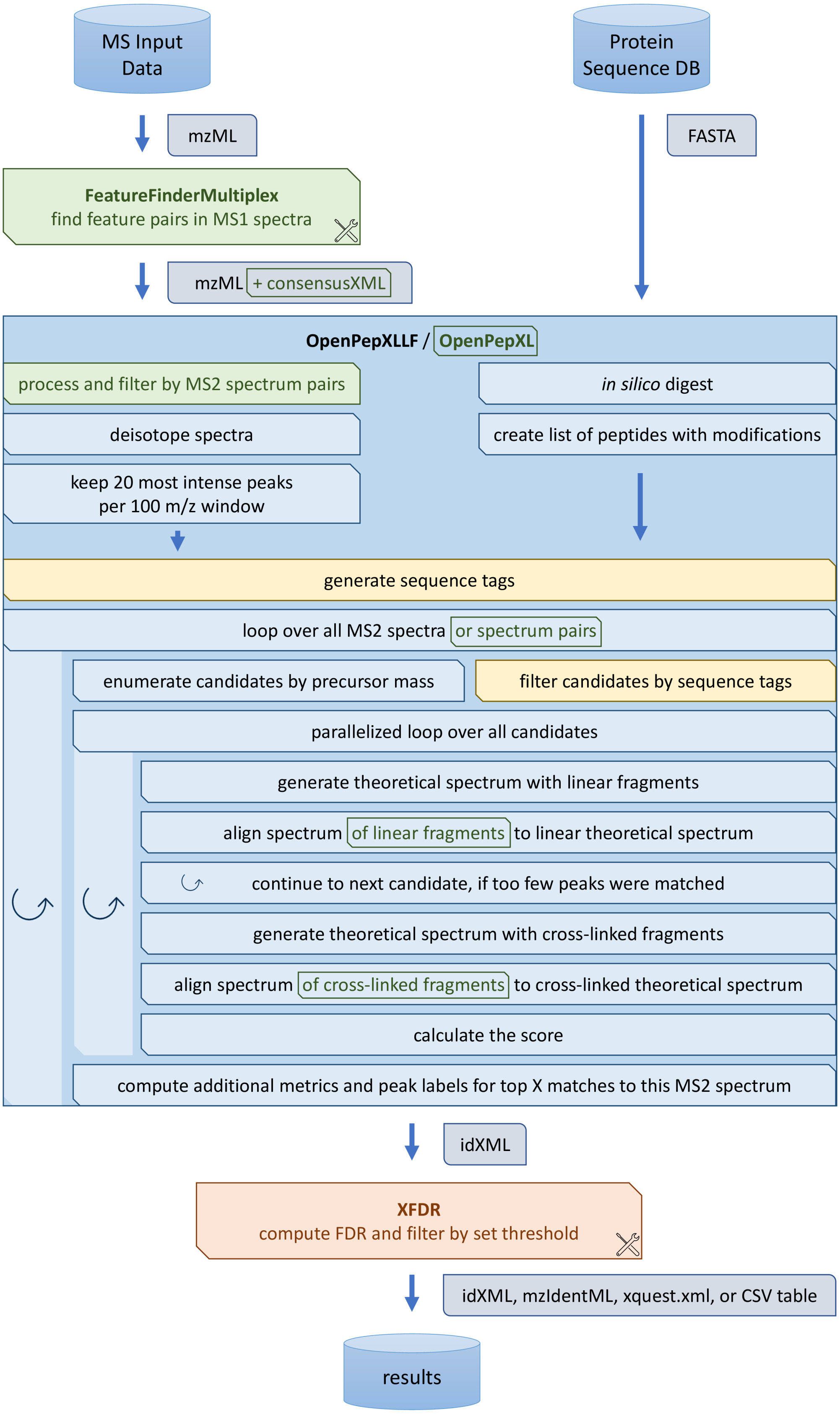

MS Input Data
Protein Sequence DB
mzML
FASTA
FeatureFinderMultiplex
find feature pairs in MS1 spectra
mzML + consensusXML
OpenPepXLLF / OpenPepXL
process and filter by MS2 spectrum pairs
in silico digest
deisotope spectra
create list of peptides with modifications
keep 20 most intense peaks per 100 m/z window
generate sequence tags
loop over all MS2 spectra or spectrum pairs
enumerate candidates by precursor mass
filter candidates by sequence tags
parallelized loop over all candidates
generate theoretical spectrum with linear fragments
align spectrum of linear fragments to linear theoretical spectrum
continue to next candidate, if too few peaks were matched
generate theoretical spectrum with cross-linked fragments
align spectrum of cross-linked fragments to cross-linked theoretical spectrum
calculate the score
compute additional metrics and peak labels for top X matches to this MS2 spectrum
idXML
XFDR
compute FDR and filter by set threshold
idXML, mzIdentML, xquest.xml, or CSV table
results

Figure 3.1: The workflow of OpenPepXL / OpenPepXLLF. The green boxes and text are specific to the OpenPepXL workflow using labeled cross-linkers. FeatureFinderMultiplex and XFDR are their own executables and can be optional depending on the workflow. The yellow boxes refer to the optional sequence tagging feature.

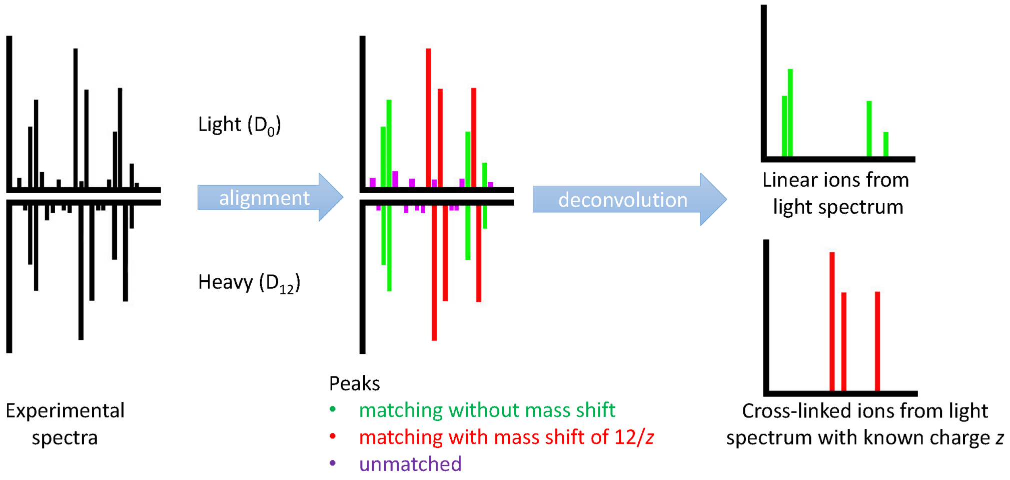
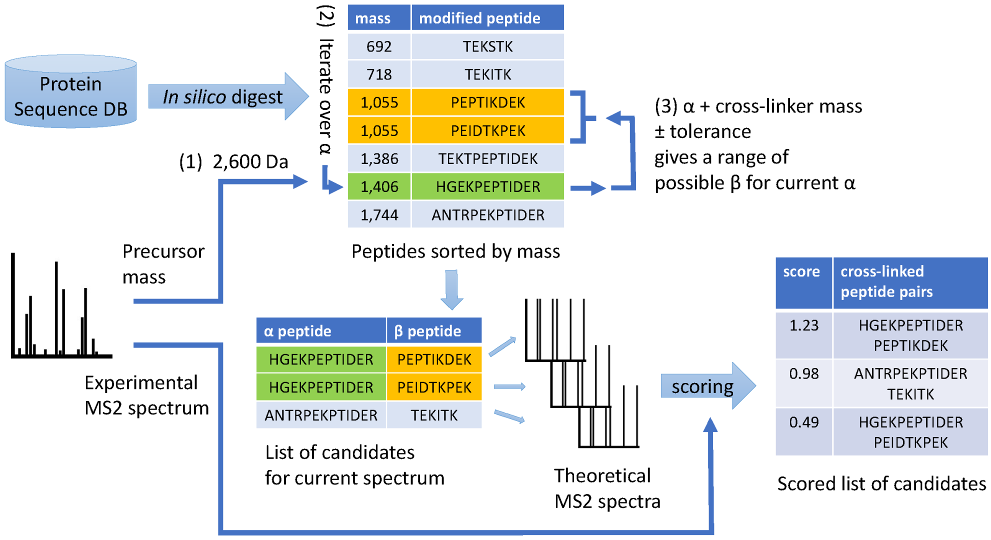

Light (D_0)
alignment
Heavy (D_{12})
deconvolution
Linear ions from light spectrum
Cross-linked ions from light spectrum with known charge *z*
Experimental spectra
Peaks
• matching without mass shift
• matching with mass shift of 12/*z*
• unmatched

Figure 3.2: Workflow for preprocessing of pairs of experimental spectra for experiments with labeled linkers. DSS-d_0/d_{12} is used as an example. Two experimental spectra of the same cross-link with different linker masses are matched against each other without a mass shift and with a shift corresponding to the mass difference of the isotopic labeling. Multiple charges are considered. Peaks without matches are discarded and two spectra of matched peaks are kept, a linear fragment spectrum without known charges and a cross-linked fragment spectrum with now determined ion charges. This allows for these groups of peaks to be matched against theoretical spectra separately.
Figure originally published in Netz *et al.*[66].


Protein Sequence DB
In silico digest
(2) Iterate over α
(1) 2,600 Da
(3) α + cross-linker mass ± tolerance gives a range of possible β for current α
Peptides sorted by mass
Precursor mass
Experimental MS2 spectrum
List of candidates for current spectrum
Theoretical MS2 spectra
scoring
Scored list of candidates

mass	modified peptide
692	TEKSTK
718	TEKITK
1,055	PEPTIKDEK
1,055	PEIDTKPEK
1,386	TEKTPEPTIDEK
1,406	HGEKPEPTIDER
1,744	ANTRPEKPTIDER

α peptide	β peptide
HGEKPEPTIDER	PEPTIKDEK
HGEKPEPTIDER	PEIDTKPEK
ANTRPEKPTIDER	TEKITK

score	cross-linked peptide pairs
1.23	HGEKPEPTIDER PEPTIKDEK
0.98	ANTRPEKPTIDER TEKITK
0.49	HGEKPEPTIDER PEIDTKPEK

Figure 3.3: Overview of peptide pair candidate enumeration. Proteins are digested *in silico* and a sorted list of modified peptides with precomputed masses is made. For each spectrum the precursor mass (1) is used to calculate the possible mass range for the α-peptide. The slice of the peptide list that corresponds to that range is iterated over (2) to find potential β-peptide candidates (3). A list of peptide pairs is generated and theoretical spectra of these pairs are created and scored against the experimental spectrum.
Figure originally published in Netz *et al.*[66].

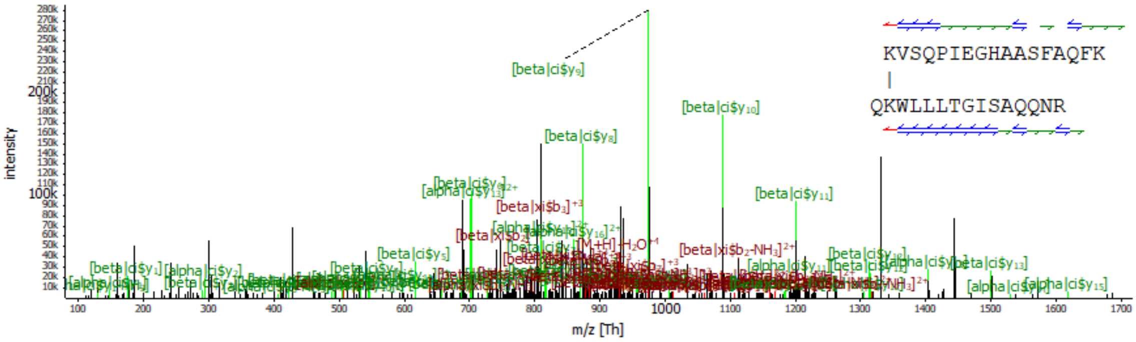


Figure 3.4: The matched and labeled MS2 spectrum of a cross-link, visualized with the new capabilities in TOPPView.

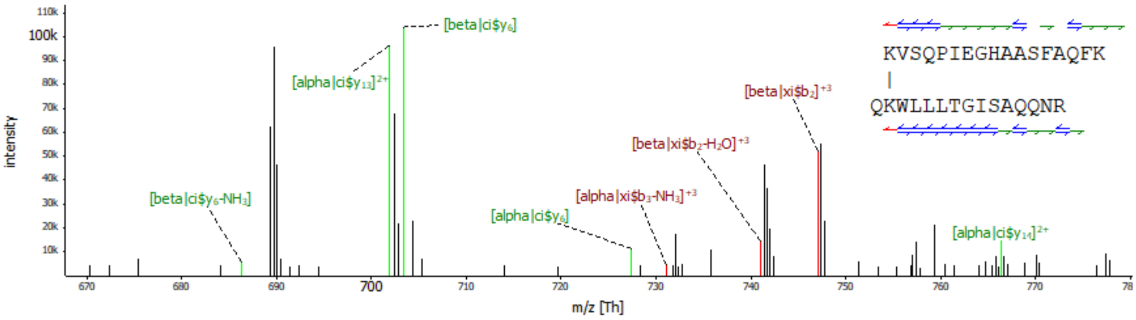


Figure 3.5: The matched and labeled MS2 spectrum of a cross-link, visualized with the new capabilities in TOPPView. The view is zoomed in to a small region of the spectrum and labels are rearranged for better visibility.

Chapter 4

Benchmarking OpenPepXL

4.1 Introduction

To show that OpenPepXL works as intended and is competitive in the field of XL-MS identification algorithms, we compared it to other tools in the field. We chose pLink2, Kojak, StavroX and XiSearch as they are some of the most popular tools for the identification of unlabeled non-cleavable cross-linkers. We also included xQuest to compare the performance with labeled linkers, for which xQuest is optimized. Parts of this chapter were adapted from the OpenPepXL publication and several of the figures were reused.[66]

4.2 Experiments and Datasets

4.2.1 Mass spectrometry of CRM Complex

The CRM complex experiment was designed and performed by Ralf Ficner and Thomas Monecke (protein expression and purification), as well as Henning Urlaub and Olexandr Dybkov (cross-linking and MS data acquisition) and published as part of the OpenPeXL publication.[66] This section describes that work. The computational analysis regarding this data as described in the rest of this chapter was performed by me.

Human full-length wild type CRM1 and SNP1 as well as a 1-180 fragment of Ran carrying a Q69L mutation, which blocks GTPase activity, were recombinantly expressed in E. coli, purified and a trimeric complex thereof was assembled as described before.[72] Prior to cross-linking, the trimeric complex was dialyzed against a buffer containing 20 mM HEPES-KOH, pH 7.9, 50 mM NaCl, 2 mM Mg(CH_3COO)$_2$, 1 mM DTT. One hundred pmol of the trimeric complex was cross-linked with
150 μM bis(sulfosuccinimidyl)suberate (BS3, Thermo Fisher Scientific) at 25 °C for

30 min in a volume of 50 μl and subsequently quenched for 5 min with 0.1 M Tris-HCl, pH 8. The sample was resolved on a NuPAGE 4-12% Bis-Tris protein gel (Thermo Fisher Scientific), followed by Coomassie staining. A slow migrating band corresponding to the BS3-cross-linked products was excised and in-gel digested with trypsin. Extracted peptides were dissolved in a sample solvent (5% acetonitrile and 0.1% formic acid). Approximately 5 pmol peptides were injected into an EASY-nLC 1000 HPLC system coupled to a Q Exactive mass spectrometer (Thermo Fisher Scientific) in duplicates under each of the three tested normalized collision energy (NCE) conditions using a 50-min method. A 20 cm long C18 analytical column with inner diameter of 75 μm self-packed with 5 μm beads (pore 120 Å, Dr. Maisch) was used for on-line HPLC. MS1 and MS2 resolution were set to 70000 and 17500, respectively. Fifteen most abundant precursors with charge of 3-7 (350-1600 m/z, isolation window 2 m/z) were selected for MS2 fragmentation at NCE 20, 24 or 28%. MS2 AGC target and injection time were limited to 200000 and 60 ms, respectively. Dynamic exclusion of 15 s was applied. The protein database only contains modified versions of the three UniProt sequences O14980, O95149 and P62826. The modifications reflect changes made to optimize the protein expression and purification. The mass spectrometry proteomics data including the modified protein sequences have been deposited to the ProteomeXchange Consortium[62] via the PRIDE[61] partner repository with the dataset identifier PXD014359.

4.2.2 Used Public Datasets

Three additional datasets were kindly provided to us by other laboratories or downloaded from public repositories.

We chose a more complex publicly available XL-MS dataset derived from a HEK293 cell lysate (ProteomeXchange ID: PXD006131[73]). A crude ribosomal fraction obtained by size exclusion chromatography from the cell lysate was cross-linked with BS3. The sample contained thousands of proteins, which were quantified by label-free quantification of peptides. The most abundant proteins were put into multiple databases for the XL-MS search. The first database has the 32 most abundant proteins, the second database has the 64 most abundant, and so on. The databases double in size up to 512 proteins for the largest one. The HCD fragmented subset of the XL-MS data contains about 170,000 MS2 spectra. We searched them against the target database with 128 proteins.

Additionally, we analyzed a dataset with labeled DSS-d_0/d_{12} and PDH-d_0/d_{10} (pimelic acid dihydrazide) cross-linkers. In this case d_0/d_{12} denotes the two versions of the cross-linker: d_0 with no deuterium and d_{12} with 12 hydrogens replaced by deuterium. Commercial Bovine Serum Albumin (BSA; Sigma-Aldrich) was cross-linked with the mentioned cross-linkers in two experiments. The samples were analyzed using HCD fragmentation and high-resolution MS/MS detection (Orbitrap Fusion Lumos), as well as ion trap CID fragmentation with low-resolution MS/MS detection (Orbitrap Elite). An analysis of this dataset had been previously published[74] and the raw data was kindly provided to us by Alexander Leitner upon request. The protein database only contained BSA.

Furthermore, we used a dataset of synthetic peptides (ProteomeXchange ID: PXD014337[75]). Tryptic peptides with one internal lysine from the *S. pyogenes* Cas9 protein were synthesized for the study. The termini of the peptides were modified to make it impossible for lysine reactive cross-linkers to link to the N-termini or C-terminal lysines. The peptides were kept in 12 separate samples without overlapping peptide sequences between them. Each sample was cross-linked with DSS and after quenching the reaction, the samples were mixed to simulate one more complex sample. MS data acquisition of three technical replicates was done using HCD fragmentation on an Orbitrap Q-Exactive HF-X. With this method identified cross-links with both cross-linked peptides coming from the same group are almost certain to be valid identifications, while links between peptides from different groups can be considered false positives. The protein database we searched against was the *S. pyogenes* Cas9 sequence with nine common contaminant proteins.

4.3 Data Processing and Tool Comparison

4.3.1 MS Data Processing and XL-MS Identification

The .RAW files of all datasets were converted into mzML, mzXML, and MGF files using MSConvertGUI from the ProteoWizard toolkit version 3.0.10577.[76] 64-bit encoding precision was used and the options to write indices and TPP compatibility were turned on, while compression for mzML files was turned off. For each dataset, reversed sequence decoy proteins were added to the protein databases using the TOPP tool DecoyDatabase. OpenPepXLLF 1.1 (OpenPepXL Label-Free) with the TOPP tool XFDR for False Discovery Rate (FDR) estimation, TPP 5.1.0 with Kojak 1.6.0 and Peptide-Prophet for FDR estimation, XiSearch 1.6.731 with xiFDR 1.1.27 for FDR estimation, xQuest 2.1.3 with xProphet for FDR estimation as well as StavroX 3.6.6.5 and pLink 2.3.5 with their built-in FDR estimators were used to analyze the label-free datasets. The parameters of the multiple tools were set to the same values where possible and to reasonable or similar values otherwise (Appendix Table E.1, Table E.2, Table E.3). Additional post-processing was partly done with the TOPP tools IDFilter, IDMerger and TextExporter for OpenPepXL and xQuest output. Most of the additional processing, comparisons and plots were done with R scripts. An FDR cut-off of 5% on the cross-link spectrum match (CSM) level was applied to every tool and every dataset, except for a few specific experiments with 1% FDR on the synthetic peptides dataset. Additionally unique residue pairs (URPs) were only kept if they were supported by at least two CSMs. A filter for link distance was applied to loop-links. Linked residue pairs were only kept, if they were at least 4 residues apart in the peptide sequence. This was done to make the tool results more comparable, since this cut-off was not a changeable parameter in any of the tools, but there were differences in the implementations. All tools were compared on the same Windows 10 PC with an Intel(R) Core(TM) i5-6500 CPU and 8 GB of RAM using one CPU core. The labeled cross-linker datasets were processed with only OpenPepXL and xQuest. For OpenPepXL, the TOPP tool FeatureFinderMultiplex was applied to connect pairs of MS1 features.

The synthetic peptides dataset was analyzed with OpenPepXL using search settings and result filter criteria equivalent to those used in the published benchmark analysis by Beveridge *et al.*[75] Results from the publication were used for the other tools.

The mass spectrometry proteomics data from the CRM dataset, including search results from all tools compared in this study have been deposited to the ProteomeXchange Consortium[62] via the PRIDE[61] partner repository with the dataset identifier PXD014359. The reanalyzed ribosomal fraction dataset was deposited with identifier PXD014520 and the BSA dataset with identifier PXD014523. The OpenPepXL results for the synthetic peptide dataset were deposited with identifier PXD021417.

4.3.2 Sensitivity and Specificity

In this study the number of reported cross-links under a fixed FDR cut-off was used as a substitute for the real sensitivity of a search. We compared the sensitivity of all tools at the same FDR setting of 5% at the CSM level, because of the trade-off between sensitivity and specificity. Additionally only URPs that were supported by at least two CSMs were considered for all tools. For OpenPepXL we recalculated the FDR at the URP level by not filtering out decoy hits early and recalculating the FDR for the summarized list of URPs. When it was feasible, we validated the URPs against structural data from the PDB.[77] The synthetic peptide dataset enables the validation of the identified cross-links even more objectively than using protein structures.

4.3.3 Structural Validation

TopoLink[65] was used for the structural validation by computing solvent accessible surface distances (SASD) between cross-linked residues. A cut-off of 35 Å was used for all datasets, as the differences between the spacers of all used cross-linkers was within 1 Å. SASD was measured between Cβ atoms while ignoring the rest of all side chains. UCSF Chimera[63] with the Xlink Analyzer[64] plugin was used for a euclidean distance visualization of the identified cross-links. A cut-off of 35 Å was chosen for the link coloring. Cross-links consistent with the structures are blue, inconsistent cross-links are red. The CRM and BSA datasets were validated on XRC structures (CRM: PDB ID 3GJX;[72] BSA: PDB ID 4F5S, chain A).

The ribosomal fraction dataset was validated on a set of XRC and cryo-EM structures. BLAST[78] was used to search the PDB database for suitable structures. Generally the structure with the best E-value was chosen, but in some cases alternative structures were selected manually, if the first choice did not cover enough identified cross-links. Because of the untargeted nature of the dataset, most of the links that could be verified with this method were intra-protein links.

4.4 Benchmark Results

In order to assess the performance of OpenPepXL, we compared it to five currently popular XL-MS search engines (pLink2,[48] XiSearch,[44] xQuest,[31 32 33] StavroX,[35] and Kojak[37]) on a number of data sets. The tools were set up with equivalent settings to the extent that was possible (see Appendix Tables E.1, E.2 and E.3 for all settings). The chosen datasets differ in their size and complexity: The more dense ribosomal fraction sample with more than a hundred searched proteins gives insight into sensitivity on complex data and computational performance, but only about a third of the cross-links could be structurally verified. On the highly purified sample of the CRM complex with a known three-dimensional structure we could fully verify the output of all tools. Then OpenPepXL was evaluated on samples with labeled cross-linkers and linkers with alternative linking chemistries to demonstrate its versatility. Additionally the synthetic peptide dataset enables an alternative method of cross-link validation and avoids the biases that a single rigid protein structure might introduce.

4.4.1 Ribosomal Fraction of a Cell Lysate

In this experiment about 170,000 MS2 spectra were searched against a protein database of 128 target and 128 decoy proteins on one CPU core of a desktop PC. This

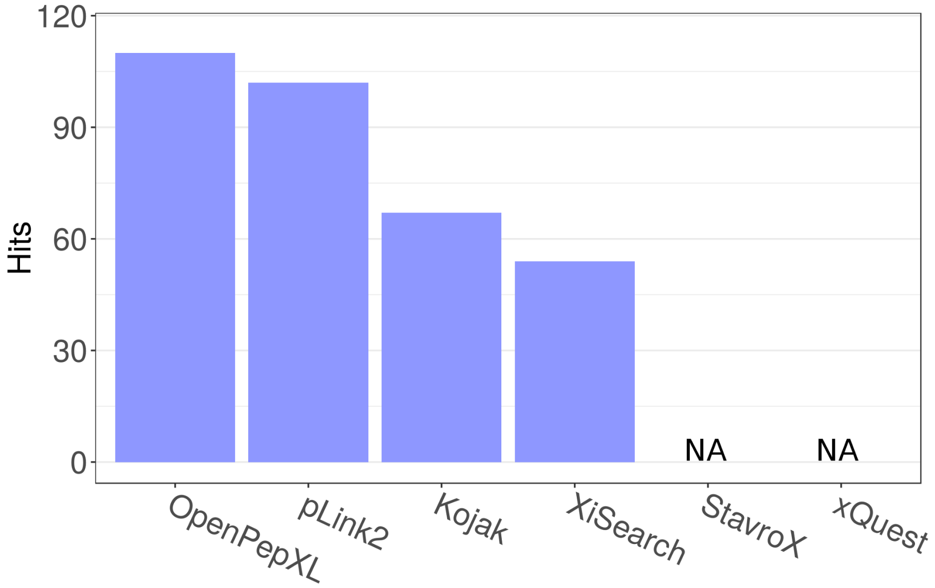


Figure 4.1: Results from the analysis of the ribosomal fraction data set. Identified URPs in the ribosomal fraction dataset after searching against 128 target proteins. OpenPepXL identified 110 URPs, pLink2 identified 102 URPs, Kojak 67 URPs and XiSearch 54 URPs. StavroX crashed after exceeding the available memory of 8 GB. The xQuest search was canceled after a week, because the estimated runtime under these conditions was unreasonable. Figure originally published in Netz *et al.* [66].

dataset size was the largest database that OpenPepXL and XiSearch could handle in a reasonable runtime of less than three days. OpenPepXL identified 110 unique residue pairs (URPs), pLink2[48] 102 (Fig. 4.1). The calculated FDR at URP level (URP-FDR) was 8.8% for OpenPepXL. The network of cross-links identified by OpenPepXL is shown in Fig. 4.3. An UpSet plot showing the overlap of identifications between the tools is shown in Fig. 4.4. Only four URPs were identified by all four tools. The highest overlap was between the two most sensitive tools OpenPepXL and pLink2, but even here the overlap is less than half of the URPs identified by each tool.

StavroX[35] crashed due to high memory requirements. The xQuest[31] search was canceled after a week, because the estimated remaining runtime under these conditions was unreasonable. As a side note: xQuest can use multiple threads and can run on most HPC environments, so finishing this search even with the limited amount of computer memory is probably within its capabilities. It is also meant to be used with labeled cross-linkers and can analyze most such datasets within feasible runtimes. The sensitivity of OpenPepXL comes at the cost of an exhaustive search of the squared search space, which necessitates longer runtimes than would be otherwise possible. pLink2 took only 15 minutes to analyze this dense dataset, Kojak about 3 hours, OpenPepXL 28

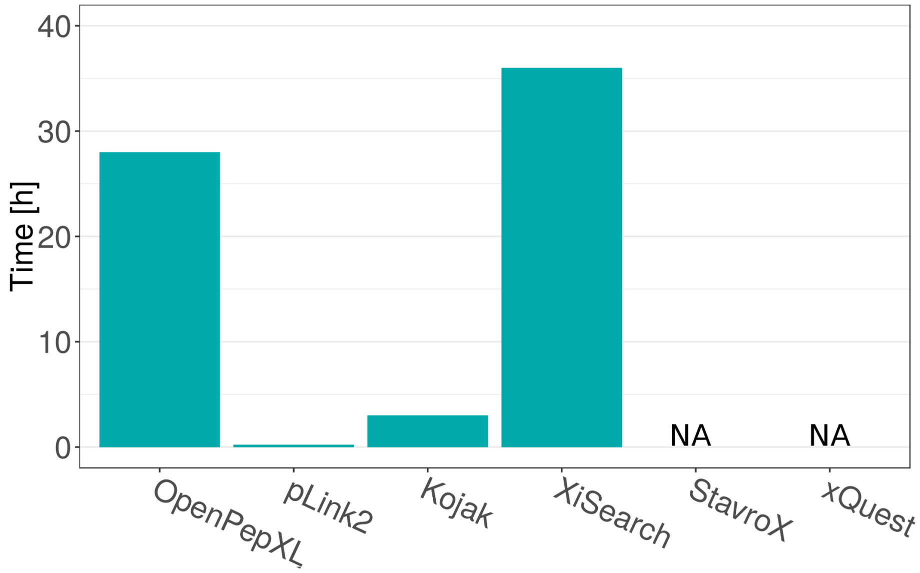


Figure 4.2: Results from the analysis of the ribosomal fraction data set. Runtime of each tool using one CPU core in hours. pLink2 needed 15 min. Kojak took 3 h, OpenPepXL 28 h and XiSearch 36 h. Figure originally published in Netz *et al.* [66].

hours and XiSearch 36 hours (Fig. 4.2). OpenPepXL can also be installed on Linux computing clusters and a speedup by a factor of 15 can be achieved by running the tool on 25 cores (Fig. 4.5). How the runtime and memory usage of OpenPepXL scales with the size of the dataset has already been discussed in Chapter 3.2. The structural validation of this dataset was largely fruitless, because most of the identified cross-links linked residues that were not covered by PDB structures, very often being found in unresolved and likely intrinsically disordered regions of existing structures. Results from the links that could be validated are shown in the appendix (Fig. D.1 and Fig. D.2). Surprisingly, among the links we were able to test not a single one from any tool was inconsistent with a structure. This, combined with the low overlap of identified URPs between the tools suggests that there are many more cross-links in this complex dataset than any one of the compared tools was able to identify.

4.4.2 CRM Complex

The dataset from the highly purified sample of the trimeric CRM complex with a known three-dimensional structure was searched by all compared tools to make sure that the higher sensitivity of OpenPepXL does not stem from just reporting more false positives and that the cross-links are just as useful for structural modeling as the ones found by the other tools. Both OpenPepXL and pLink2 found 78 URPs. Kojak reported 61

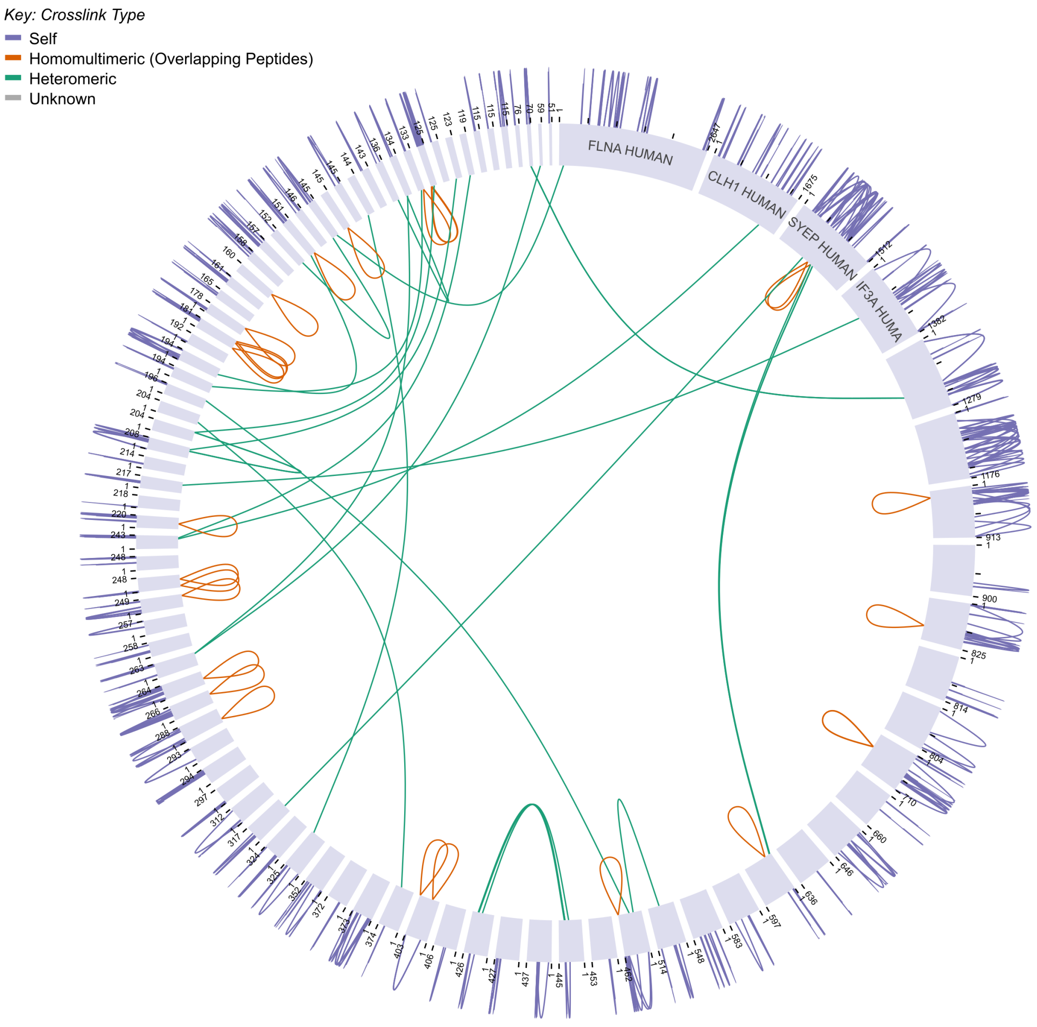


Figure 4.3: Network of cross-links identified by OpenPepXL in the ribosomal fraction dataset. Proteins are shown as grey bars in a circle. The length of the bars is proportional to the protein lengths. Intra-protein links are represented as purple loops on the outside, inter-protein links as green lines on the inside. Inter-protein links between copies of the same protein are shown as orange loops inside of the circle. Figure created with XiView.[47]

URPs. The cross-links that were covered by the structure were 45 for OpenPepXL, 41 for Kojak and 40 for pLink2 (Fig. 4.6). OpenPepXL's URP-FDR was calculated to be 12%. TopoLink was used to evaluate these cross-links on the PDB structure 3GJX using the solvent accessible surface distance (SASD). Among OpenPepXL's URPs was one that exceeded the cut-off of 35 Å, Kojak's links were all validated and pLink2 had 2 URPs that were inconsistent with the structure (IWS) (Fig. 4.6, Fig. 4.8). OpenPepXL's error rate on this dataset is similar to the other tools with comparable sensitivity. An

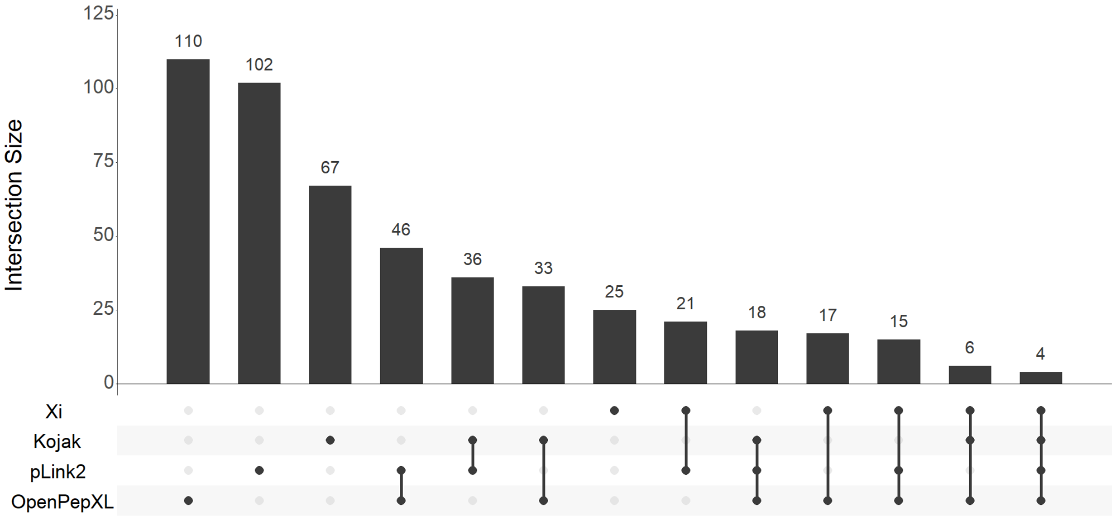


Figure 4.4: UpSet plot showing the intersections of identified URPs among the tools for the ribosomal fraction dataset.

UpSet plot showing the intersections of identified URPs between the tools is shown in Fig. 4.9. OpenPepXL identified 14 URPs, that were not identified by any of the other tools and 6 of them could be structurally validated. Only 13 URPs, 17% of the links identified by OpenPepXL or pLink2, were identified by all tools, but this overlap is substantially higher than for the more complex ribosomal fraction dataset.

4.4.3 BSA with Labeled Cross-Linkers

The BSA dataset used the labeled cross-linkers DSS-d_0/d_{12} and PDH-d_0/d_{10}. PDH cross-links the carboxylic acids of aspartate and glutamate. OpenPepXL and xQuest were applied on this data to assess the performance of OpenPepXL on data with labeled cross-linkers and different linking chemistries. xQuest was originally developed with stable isotope labeled cross-linkers in mind. Its algorithm and score were calibrated for lower-resolution ion trap spectra. OpenPepXL was calibrated using HCD fragmentated spectra from orbitrap instruments. OpenPepXL's spectrum alignment and deisotoping algorithms rely on high-resolution spectra. xQuest does not apply deisotoping, but can make use of the stable isotope labels to denoise spectra. In this dataset equal BSA protein samples were cross-linked with two very different labeled linkers and measured using both types of instruments. Since the target protein is very simple, we also searched for DSS cross-links to serine, threonine and tyrosine. For the PDH data we set aspartate, glutamate and the C-terminus as potentially linkable sites. In the orbitrap data OpenPepXL found 65 DSS and 22 PDH URPs. In the ion trap data xQuest

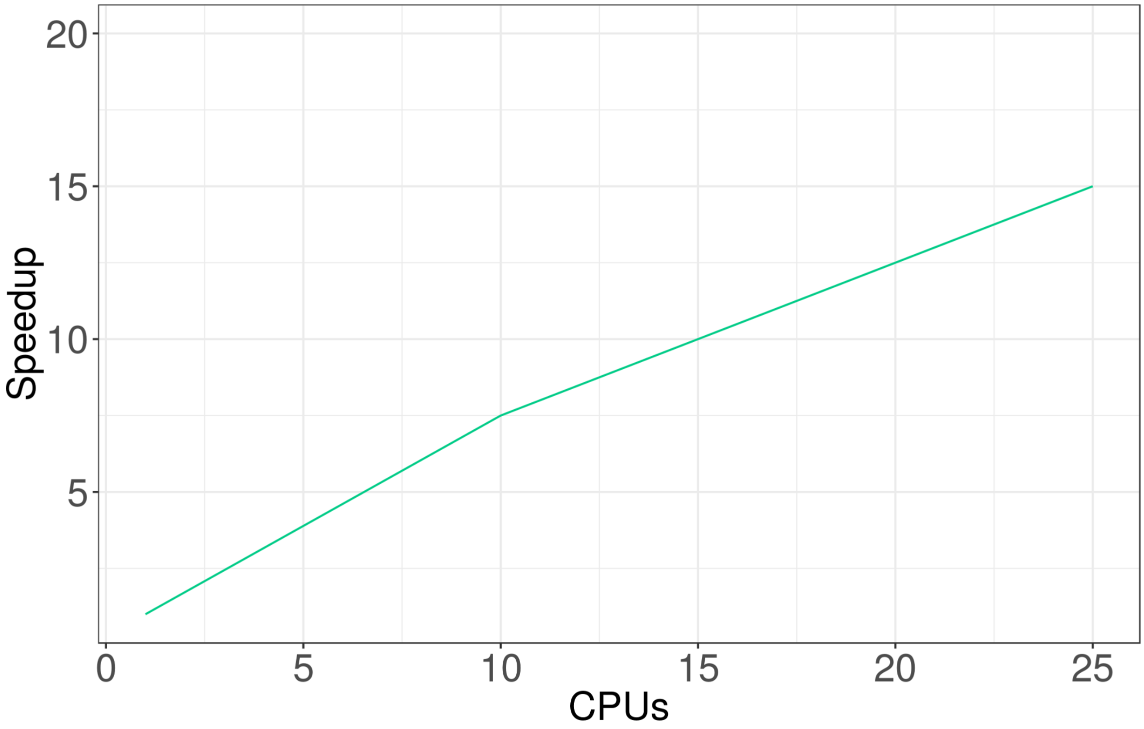


Figure 4.5: Speedup of the OpenPepXL algorithm when using up to 25 CPU cores relative to the speed of using one core. When using 25 CPU cores to run one analysis in parallel, OpenPepXL runs 15 times faster than on one core and uses almost the same amount of memory. The runtime was measured with 1, 10 and 25 CPU cores on the same computer analyzing the ribosomal fraction dataset.

found 57 DSS and 9 PDH URPs (Fig. 4.7). These results are the most comparable, since they represent the strengths of both tools. The URP-FDR for OpenPepXL was 1.5% for the DSS and 7.1% for the PDH results. TopoLink was used to evaluate these cross-links on chain A of the PDB structure 4F5S. OpenPepXL had three DSS URPs and one PDH URP exceeding the SASD cut-off. xQuest had one DSS URP that was inconsistent with the structure (Fig. 4.7, Fig. 4.10).

4.4.4 Synthetic Peptides

Structural validation of cross-links can be biased due the nature of XRC structures. They are rigid snapshots of a low energy state and it is difficult to consider the protein's flexibility and dynamics when measuring distances between residues. As we have seen in the ribosomal fraction dataset, cross-links also tend to link residues in very flexible or even intrinsically disordered regions, because this flexibility favors their chemical reactions. It is therefore desirable to find alternative methods of validating the correctness of cross-link identifications, independently from available protein struc-

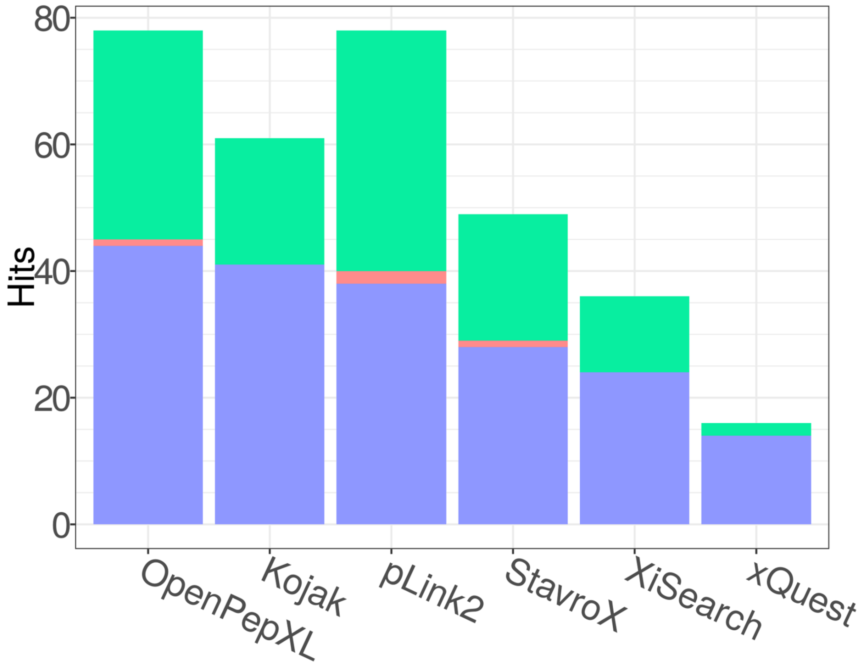


Figure 4.6: Identified URPs in the CRM data set. TopoLink was used to evaluate the cross-links on the PDB structure 3GJX. The red bars show cross-links inconsistent with the structure (IWS). That means they are either not solvent accessible, or their distance exceeds 35 Å according to the SASD. The green bars show the proportion of URPs that were not covered by the structure. OpenPepXL identified 78 URPs. 44 URPs could be validated and one link exceeds the distance cut-off with 37.4 Å between the linked residues. Kojak identified 61 URPs, 41 could validated. pLink2 identified 78 URPs of which 38 were validated and two are IWS. One of them is shared with OpenPepXL and the second one has a distance of 40.4 Å. StavroX found 48 URPs including 28 validated ones and one IWS. XiSearch identified 36 URPs including 24 validated ones and xQuest found 16 with 14 validated links. Figure originally published in Netz *et al.*[66].

Table 4.1: Reported CSMs from the synthetic peptide dataset at a 5% FDR cut-off. All data except for OpenPepXL was taken from Beveridge *et al.*[75]

Search Engine	Number of CSMs								
	Correct			Incorrect			Calculated FDR (%)		
	R1	R2	R3	R1	R2	R3	R1	R2	R3
OpenPepXL	822	938	895	28	24	34	3.3	2.5	3.7
pLink2	639	712	683	27	27	39	4.1	3.7	5.4
StavroX	378	434	419	9	12	10	2.3	2.7	2.3
XiSearch	491	498	547	20	13	10	3.9	2.6	1.8

tures. The synthetic peptides dataset provides such an opportunity. It enables a more objective validation of the cross-links found in it.

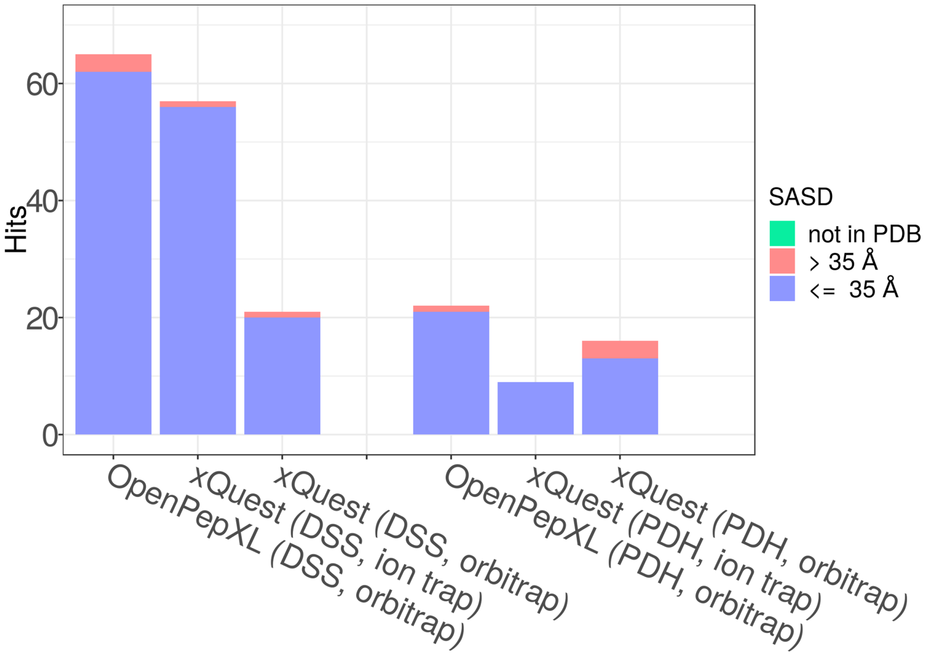


Figure 4.7: Identified URPs in the BSA dataset. OpenPepXL and xQuest were compared on lower-resolution ion trap and high-resolution orbitrap data with the two isotopically labeled linkers DSS-d_0/d_{12} and PDH-d_0/d_{10}. TopoLink was used to evaluate the cross-links on the PDB structure 4F5S. The red bars show cross-links inconsistent with the structure (IWS). That means they are either not solvent accessible, or their distance exceeds 35 Å according to the SASD. The green bars show the proportion of URPs that were not covered by the structure.
OpenPepXL found 65 URPs in the DSS orbitrap dataset, with three links that exceed the cut-off, but are all below a distance of 40 Å. It found 22 URPs in the PDH orbitrap dataset, including one link that exceeds the cut-off with 59.3 Å. xQuest reported 16 URPs in the PDH orbitrap dataset, with 3 IWS links. It found 21 URPs in the DSS orbitrap dataset, with one IWS link. It reported 9 URPs in the PDH ion trap dataset and 57 URPs in the DSS ion trap dataset, with one link exceeding the cut-off with a distance of 70.2 Å. Figure originally published in Netz *et al.* [66].

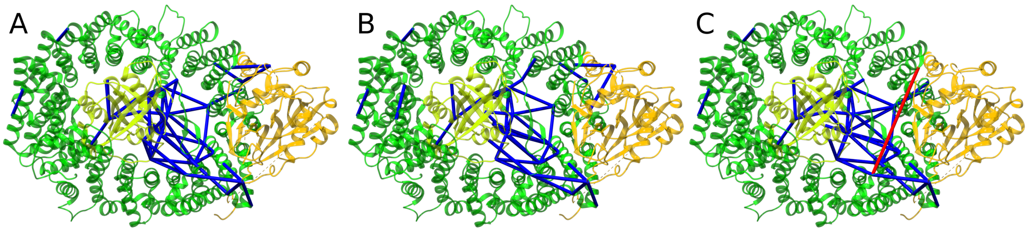


Figure 4.8: Cross-links mapped onto the CRM complex (PDB ID 3GJX). Cross-links from **(A)** OpenPepXL, **(B)** Kojak and **(C)** pLink2. Cross-links exceeding a Euclidean distance of 35 Å are red and those consistent with the structure are blue. Figure originally published in Netz *et al.* [66].

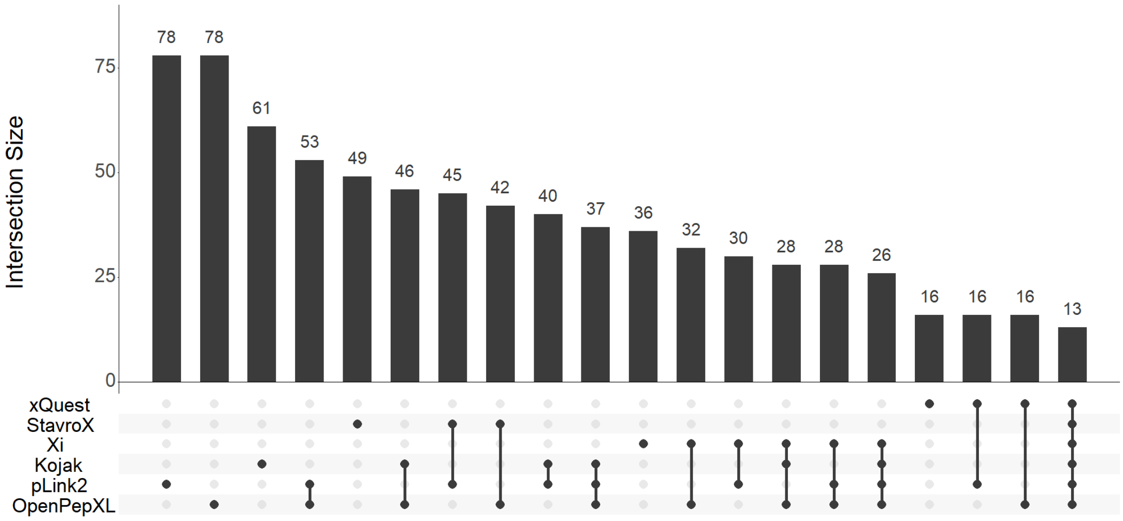


Figure 4.9: UpSet plot of identified URPs for the CRM complex dataset.

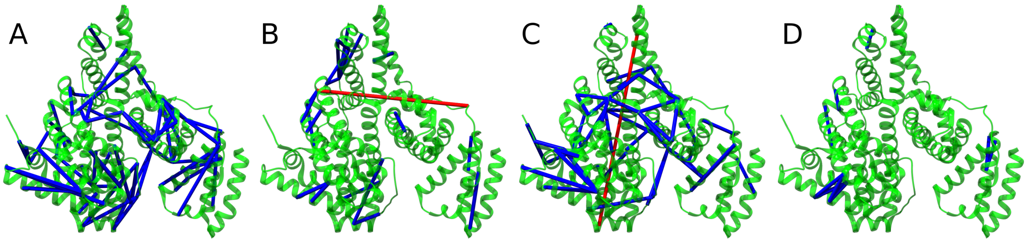


Figure 4.10: Cross-links mapped onto BSA (PDB ID 4F5S).**(A)** DSS URPs found by OpenPepXL in the orbitrap data. **(B)** PDH URPs found by OpenPepXL in the orbitrap data. **(C)** DSS URPs reported by xQuest in the ion trap data. **(D)** PDH URPs reported by xQuest in the ion trap data. Figure originally published in Netz *et al.*[66].

Table 4.2: Reported URPs from the synthetic peptide dataset at a 5% FDR cut-off. All data except for OpenPepXL was taken from Beveridge *et al.*[75]

Search Engine	Number of URPs								
	Correct			Incorrect			Calculated FDR (%)		
	R1	R2	R3	R1	R2	R3	R1	R2	R3
OpenPepXL	242	250	237	21	21	21	8.0	7.7	8.1
pLink2	217	230	203	26	24	33	10.7	9.4	14.0
Kojak	220	225	217	68	60	67	23.6	21.0	23.6
StavroX	159	175	154	8	10	9	4.8	5.4	5.5
XiSearch	179	183	179	18	11	7	9.1	5.7	3.8

For this comparison we searched the data with OpenPepXL and compared the results to the benchmark results reported in Beveridge *et al.*[75] To make this comparison valid, we also used search settings that are as close as possible to the settings used in

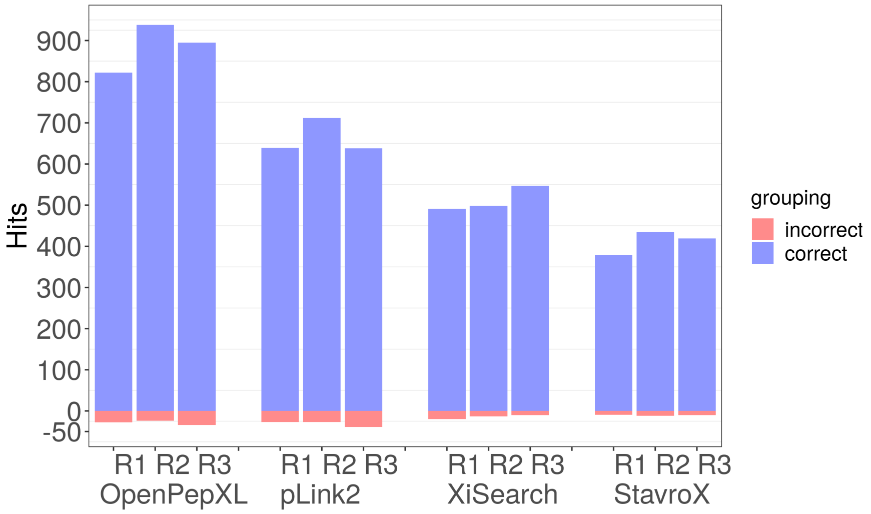


Figure 4.11: Reported CSMs from synthetic peptides data set at a 5% FDR cutoff. Shows the three replicates R1, R2 and R3. Validated CSMs are blue and false positives are red on the negative y-axis. All data except for OpenPepXL was taken from Beveridge *et al.*[75] The exact numbers are in Table 4.1. Figure originally published in Netz *et al.*[66].

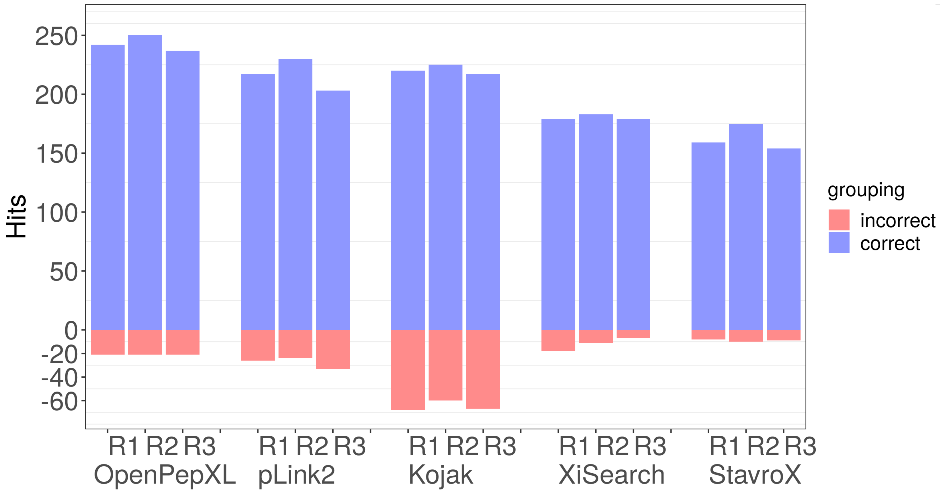


Figure 4.12: Reported URPs from the synthetic peptides data set at a 5% FDR cutoff. Shows the three replicates R1, R2 and R3. Validated URPs are blue and false positives are red on the negative y-axis. All data except for OpenPepXL was taken from Beveridge *et al.*[75] The exact numbers are in Table 4.2. Figure originally published in Netz *et al.*[66].

that study. Some comparisons are not complete, since that data was omitted in that study. For example, they did not include xQuest and did not report 1% FDR results for Kojak. From the multiple results available for Kojak we chose the 5% FDR results with Percolator using only unique cross-links, since these results are the most favorable for Kojak. The results are shown in Fig. 4.11, Fig. 4.12, Fig. 4.13, Fig. 4.14 and Fig. 4.15 and Tables 4.1, 4.2, 4.3, and 4.4. At 5% FDR OpenPepXL found on average 242

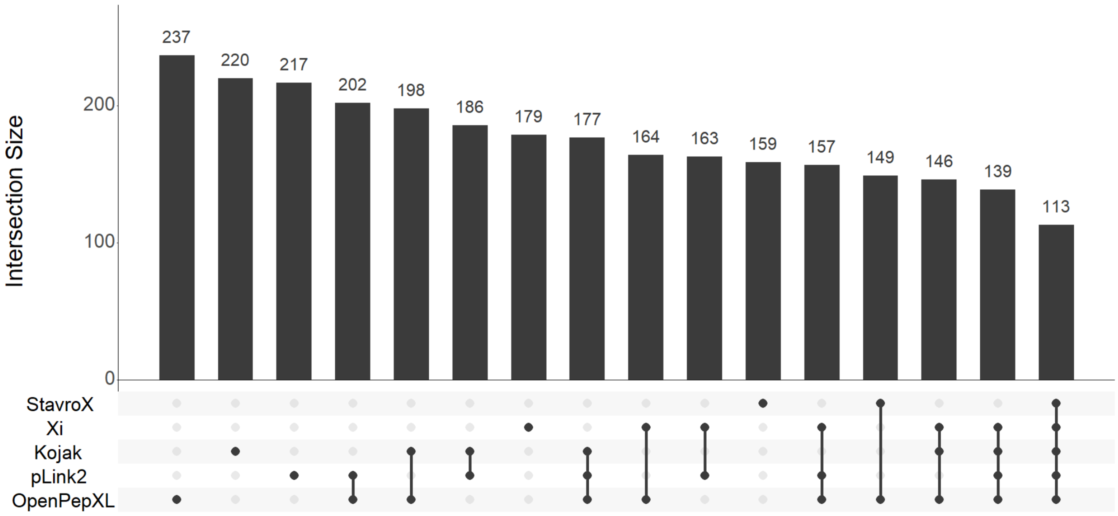


Figure 4.13: UpSet plot of identified URPs at 5% FDR for the synthetic peptides dataset.

Table 4.3: Reported CSMs from the synthetic peptide dataset at a 1% FDR cut-off. All data except for OpenPepXL was taken from Beveridge *et al.*[75]

Search Engine	Number of CSMs								
	Correct			Incorrect			Calculated FDR (%)		
	R1	R2	R3	R1	R2	R3	R1	R2	R3
OpenPepXL	368	506	365	4	5	4	1.1	1.0	1.1
pLink2	594	644	585	10	13	25	1.7	2.0	4.1
StavroX	265	157	160	4	0	1	1.5	0	0.6
XiSearch	312	352	438	2	4	5	0.6	1.1	1.1

Table 4.4: Reported URPs from the synthetic peptide dataset at a 1% FDR cut-off. All data except for OpenPepXL was taken from Beveridge *et al.*[75]

Search Engine	Number of URPs								
	Correct			Incorrect			Calculated FDR (%)		
	R1	R2	R3	R1	R2	R3	R1	R2	R3
OpenPepXL	161	196	148	2	4	3	1.2	2.0	2.0
pLink2	215	218	189	9	12	22	4.0	5.2	11.6
StavroX	124	91	90	4	0	1	3.1	0	1.1
XiSearch	141	152	163	2	3	5	1.4	1.9	3.0

validated URPs with an average URP-FDR of 7.9%. pLink2 reported on average 217 validated URPs with an average URP-FDR of 11.4% (Fig. 4.12, Table 4.2).

At 1% FDR OpenPepXL found on average 168 validated URPs with an average URP-FDR of 1.7%. pLink2 reported on average 207 validated URPs with an average

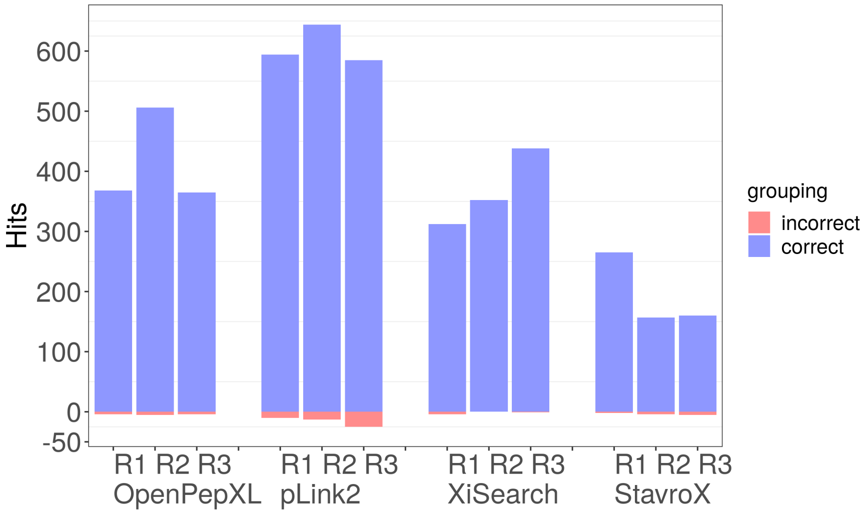


Figure 4.14: Reported CSMs from the the synthetic peptides dataset at a 1% FDR cut-off. Shows the three replicates R1, R2 and R3. Validated CSMs are blue and false positives are red on the negative y-axis. All data except for OpenPepXL was taken from Beveridge *et al.*[75] The exact numbers are in Table 4.3. Figure originally published in Netz *et al.*[66].

Table 4.5: Targets and decoys assigned to spectra in total, as well as validated CSMs above and URPs assigned below the 5% FDR cut-off for the first replicate (R1) of the synthetic peptide dataset. All data except for OpenPepXL was taken from Beveridge *et al.*[75] The number of assigned decoys for StavroX is missing, because they are not reported by StavroX.

Search Engine	Total target CSMs	Total decoy CSMs	Validated CSMs above 5%	Additional URPs below 5%
OpenPepXL	2029	2156	822	80
pLink2	1006	383	639	4
StavroX	1322	NA	378	58
XiSearch	1686	2677	491	11

URP-FDR of 6.9% (Table 4.4). This dataset has a much stronger overlap in reported URPs between the tools compared to the other datasets in this study (Fig. 4.13). At 5% FDR 113 URPs were identified by all tools, almost half of all URPs found by OpenPepXL. OpenPepXL identified 22 URPs that are not found by either pLink2, StavroX or XiSearch. 17 of those were also identified by the Kojak search with a very high average calculated FDR of 22.7%. The UpSet plot for this dataset in Fig. 4.13 shows the highest overlaps of identified cross-links yet. The highest overlap is 85% between OpenPepXL and pLink2, and almost half of the links identified by OpenPepXL were also found by all other tools in the comparison.

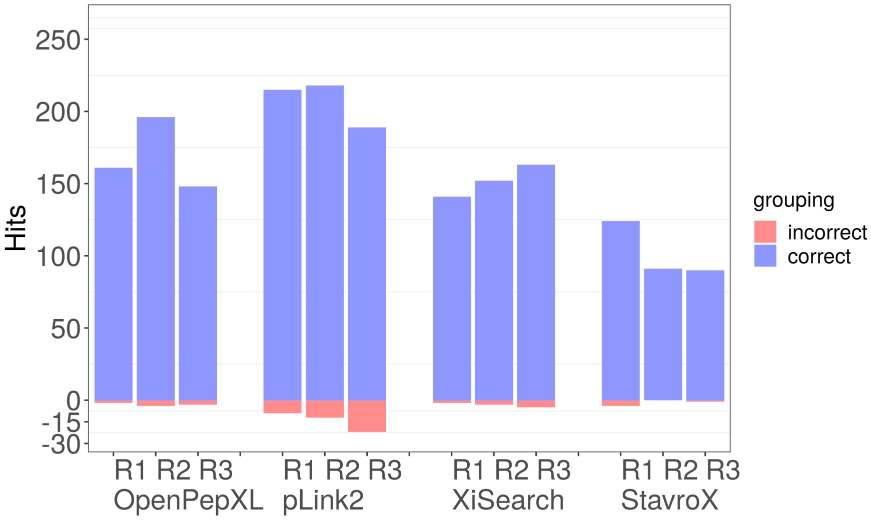


Figure 4.15: Reported URPs from the synthetic peptides dataset at a 1% FDR cut-off. Shows the three replicates R1, R2 and R3. Validated URPs are blue and false positives are red on the negative y-axis. All data except for OpenPepXL was taken from Beveridge *et al.*[75] The exact numbers are in Table 4.4. Figure originally published in Netz *et al.*[66].

We took a look into how the spectra of this dataset were utilized by the different tools using the first replicate as an example (Table 4.5). The file contains 5022 MS2 spectra. Before FDR filtering, OpenPepXL returned at least one hit for 4185 spectra. Of the 1st ranked hits 2029 were targets and 2156 decoys. After filtering by 5% FDR, validated cross-links were matched to 822 spectra above that cut-off. 80 valid cross-links were found below that cut-off. pLink2 returned hits to only 1389 spectra with 1006 targets and 384 decoys. Valid cross-links were matched to 639 spectra above the FDR cut-off and 4 valid links were below. XiSearch returned hits to 4363 spectra with 1686 targets and 2677 decoys. Valid cross-links were matched to 491 spectra above the FDR cut-off and 11 valid links were below.

4.4.5 Performance with CID-Cleavable Cross-Linkers

As a part of Ruben Grünbergs Master thesis,[71] OpenPepXL's performance with CID-cleavable cross-linkers was compared to one of the most effective and commonly used tools in this domain: MeroX.[36] The comparison was done on multiple experiments using the cross-linkers DSBU and DSSO with stepped HCD fragmentation in MS2. OpenPepXL performs roughly equally well with MeroX in terms of sensitivity and specificity, but has an advantage in runtime and memory requirements. Example results for identified cross-links, runtime and memory usage are shown in Fig. 4.16, Fig. 4.17 and Fig. 4.18. Using OpenPepXL with cleavable cross-linkers improves its runtime and memory usage significantly compared to using non-cleavable linkers, allowing it to search through a database of 10,000 proteins against more than 31,000 MS2 spectra in about 40 minutes on a single CPU core.

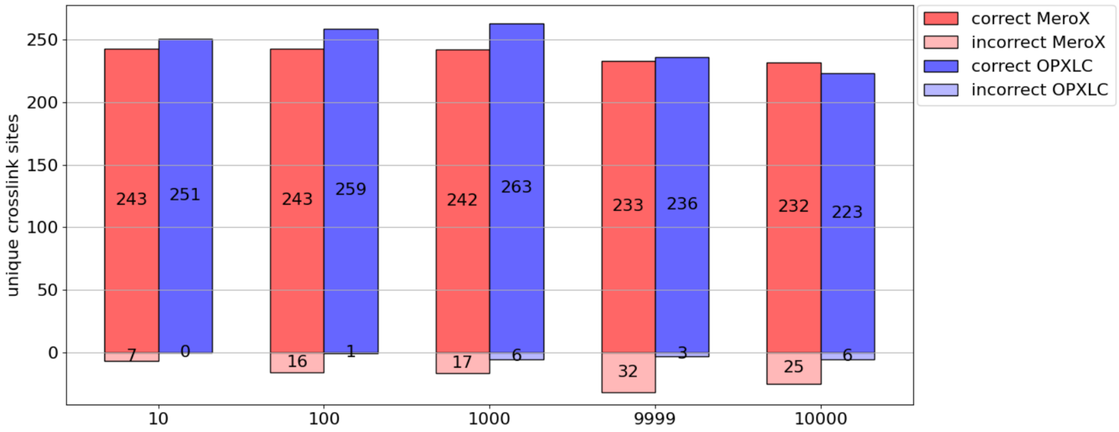


Figure 4.16: Example result of a benchmark for cleavable cross-links with OpenPepXL and MeroX from Ruben Grünbergs Master thesis.[71] The sample contained only BSA cross-linked with DSBU, but an entrapment database with up to 10,000 human proteins was added to the search space. The size of the entrapment database is shown on the X-Axis. The one sequence missing in the 9,999 database is HSA, which has a 77% sequence similarity to BSA and seems to affect OpenPepXLs performance significantly.

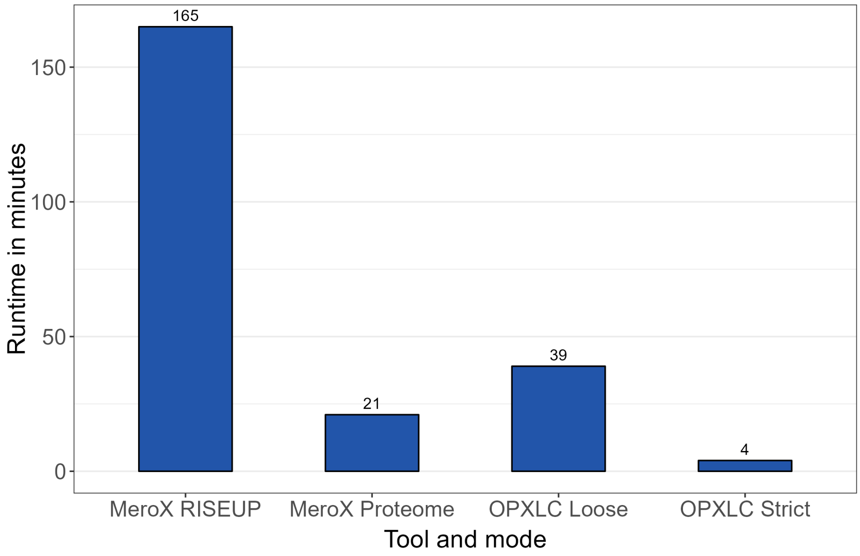


Figure 4.17: Runtime of OpenPepXL and MeroX with cleavable cross-links. The sample contained only BSA cross-linked with DSBU, but an entrapment database with 10,000 human proteins was added to the search space. The dataset had 31,000 MS2 spectra and was searched on a single CPU core of a higher end laptop CPU (Intel i7-11850H).

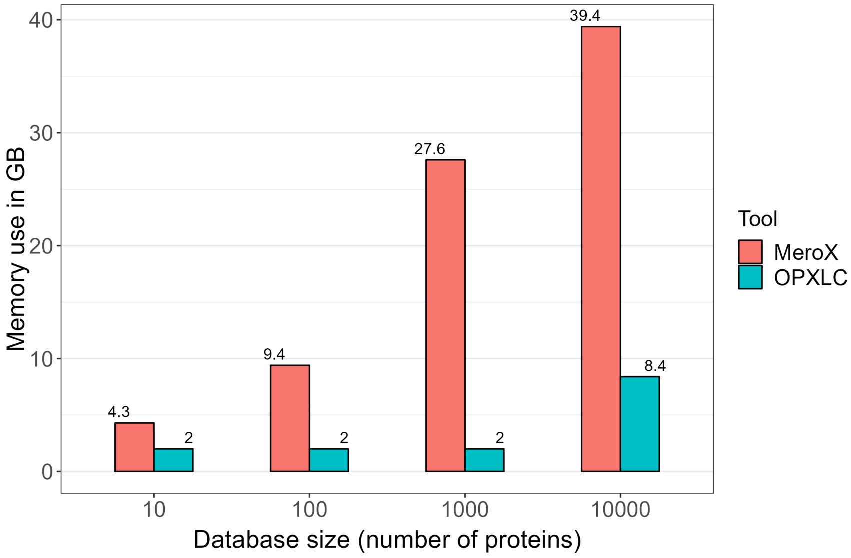


Figure 4.18: Memory usage of OpenPepXL and MeroX with cleavable cross-links. The sample contained only BSA cross-linked with DSBU, but an entrapment database with 10,000 human proteins was added to the search space. The dataset had 31,000 MS2 spectra.

4.5 Discussion

Using structural validation was in some cases insufficient to satisfactorily compare the performance of the tools. For the CRM dataset the raw count of identified URPs put OpenPepXL and pLink2 on equal terms. For the ribosomal fraction dataset OpenPepXL had a few more links in total. In both cases only the URPs covered by protein structures in the PDB could be verified. The proportions of identified URPs that could be verified was not equal across tools. The best we can say is that in both cases the number of URPs that were structurally verifiable was highest for OpenPepXL. The synthetic peptides dataset provided us with the opportunity of a more direct comparison, because every URP for this dataset can be verified.

Comparing the results after different FDR cut-offs, the 5% FDR results have a visible pattern in the validated cross-links across the different replicates (Fig. 4.12). The second replicate has the most valid URPs and the third the fewest. This is reproducible with all compared tools. In the 1% FDR results this is only recognizable in OpenPepXL's and pLink2's URPs. The results look noisier overall (Fig. 4.15, Tables 4.3 and 4.4). This time pLink2 found more valid cross-links than OpenPepXL, but it also had an unusually high actual FDR of 4.1% at CSM level and an URP-FDR of 11.6%.

Looking at all the hits assigned to spectra by the tools (Table 4.5), XiSearch matched hits to the highest number of spectra and also matched the highest ratio of decoys.

This might make it more stringent and its calculated FDR values at both the CRM and URP level are indeed lower on this dataset than those of pLink2 and Kojak. However, they are not significantly different from OpenPepXLs FDR values (Tables 4.1 and 4.2). pLink2 matched the fewest hits even without any FDR filtering, probably due to its several heuristics to filter out low quality spectra and candidates before computing the full score. Most candidates do not make it far enough in the pipeline to reach the FDR filtering step. This can be felt very prominently in its blazing runtime, but it seems the small number of hits can cause issues for its FDR estimation for low cut-offs. And it is also likely that some valid cross-links were filtered out already before the FDR filter as well. OpenPepXL had a very even ratio of target and decoy hits. It matched valid cross-links to the most spectra and in the end reported the most valid cross-links above the FDR cut-off in most cases. The 80 valid cross-links below the 5% FDR cut-off indicate that there might be room for improvement in separating correct from incorrect CSMs. However, these might also have been matches made correctly just using the precursor mass and without sufficient fragmentation to get a high match score.

The UpSet plots for all these datasets show that the intersections of identified cross-links between the tools decrease with a higher complexity of the datasets. The more complex the sample is and the more proteins are in the searched protein database, the more unique cross-links are identified by each tool. This probably means that the truth is closer to the union of the correct identifications, rather than the intersection. In the more complex datasets there are likely many correct identifications under the FDR cut-off for each tool, and due to the different scores sorting the CSMs in different orders, distinct sets of cross-links come up on top above the cut-off. This means there should still be room for improvement in the scores of all these tools to lift more correct identifications in the ranking.

Chapter 5

Applications

5.1 Introduction

Large transmembrane protein complexes and proteins with disordered functional regions have always been difficult to solve for most of the traditional methods in protein structure research.[79–82] Recently, Cryo-EM and XL-MS have become very successful in tackling these problematic cases. We applied XL-MS in two research problems involving such difficult complexes.

One project was centered around protein complexes involved in the control of DNA structure during meiosis. Many of these proteins are very flexible or have disordered regions, some of which are directly involved in binding to other proteins. During this project we analyzed several different but related protein complexes with XL-MS to gain insights into their topologies. This work was done in collaboration with John Weir, Dorota Rousova, Magdalena Firlej and Veronika Altmannova from the Weir lab of the Friedrich Miescher Laboratory in Tübingen.

The second project was related to research regarding the structure and assembly of the type III secretion system of *Salmonella*. It is an overwhelmingly complex and important piece of protein machinery that is embedded in both the inner and outer membrane. It is directly involved in the pathogenicity of some bacteria and therefore of high interest for medical research. The fact that it is embedded in membranes and its size make it difficult to study the details of the interactions between its many components. In this project we tried to apply XL-MS in a more targeted way. This work was done in collaboration with Samuel Wagner and his group at the Interfaculty Institute of Microbiology and Infection Medicine in Tübingen.

5.2 DNA Structure During Meiosis

5.2.1 Introduction and Common Methods

These experiments followed a common protocol. Proteins were expressed in *S. cerevisiae* or insect cells and purified. Some of the proteins were expressed as maltose binding protein (MBP) fusion proteins to improve solubility and for affinity purification. The MBP tags were cleaved off and the targeted protein complexes were reconstituted in a solution. A cross-linking reagent was then used to cross-link the complexes and the reaction was quenched. Cross-linked molecules were separated from other proteins by gel electrophoresis. To reduce potential disulfide bonds between cysteines DTT was used and the reduced side chains were blocked by carbamidomethylation. The cross-linked proteins were digested in gel using either trypsin or Lys-C. The digested sample was then analyzed with LC-MS/MS using a 60 minute LC gradient and a Q Exactive HF mass spectrometer (Thermo Fisher Scientific).

5.2.2 Double-Strand Break Initiation, Mer2 and Spp1

Background

Meiotic recombination is a central mechanism of meiosis and an important contributor to genetic diversity. During meiosis, homologous chromosomes are paired and recombined into new molecules by breaking and rejoining the double-stranded DNA strings.[83]

One feature of meiotic recombination hotspots shared by many different organisms is the methylation of histone H3K4. The PHD finger protein Spp1 recognizes this modification and binds to the histone. The double-strand break (DSB) protein Mer2 then interacts with Spp1 to promote DSB formation.[23] The structure of Mer2 has not been solved yet and only some fragments of Spp1 were crystallized in other complexes before. How these two proteins interact is also not known in detail. However, it is known that *S. cerevisiae* Mer2 forms a homotetramer via its core domain and forms a parallel-antiparallel coiled-coil. Spp1 alone is a monomer, but two Spp1 proteins bind to the tetramerized coiled-coil core domains of a Mer2 tetramer via Spp1's C-terminus.[24] A single amino acid substitution of the Val residue at position 195 in Mer2 can break its interaction with Spp1.[84] A cartoon representation of the Mer2-Spp1 complex binding to a H3K4me3 nucleosome is shown in Fig. 5.1.

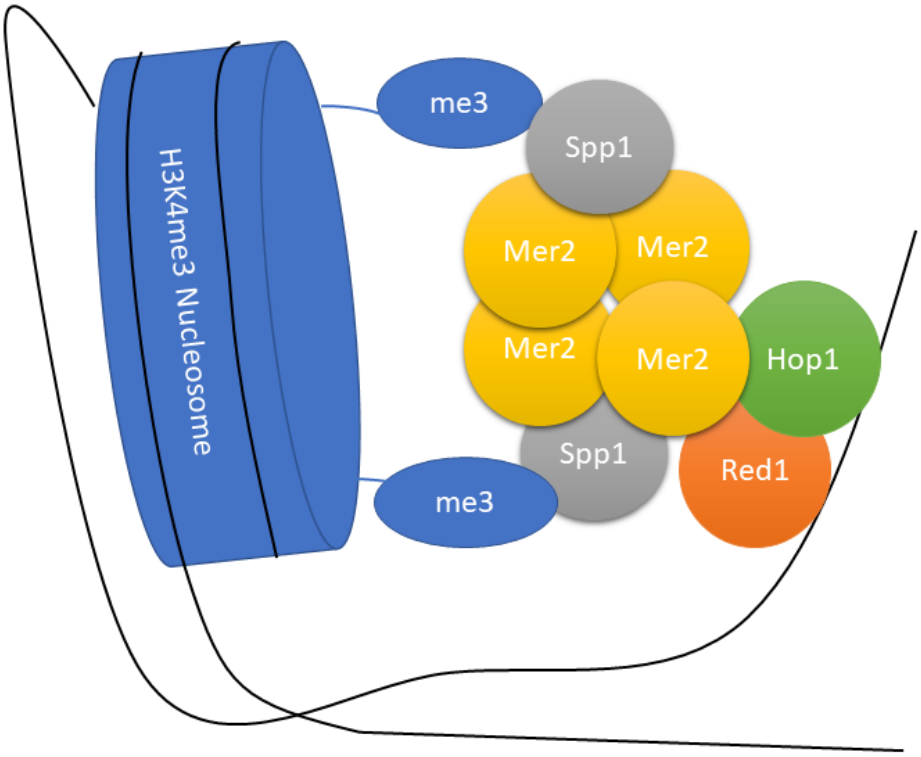


Figure 5.1: Cartoon of meiotic loop axis structure with the roles of Mer2 and Spp1. The proteins Red1 and Hop1 together with cohesin bind to a DNA double strand and form a loop of extruded chromatin. The Mer2-Spp1 complex directs the Spo11 complex (not shown), that makes the DNA breaks, to the proximity of a H3K4me3 nucleosome.

Experiments

In the first experiment Mer2, Spp1 from *S. cerevisiae* and trimethylated H3K4 nucleosomes from *Xenopus laevis* were cross-linked with BS2G. This cross-linker is among the shorter ones and, if the experiment is successful, provides slightly shorter and therefore more useful distance restraints. In one sample Mer2 and Spp1 were cross-linked alone, in three further samples they were cross-linked together with nucleosomes. In the second experiment the labeled cross-linker BS3-d_0/d_{12} in a ratio of 1:1 was used to cross-link a similar set of samples. Two samples only contained Mer2 and Spp1 and in three more samples they were cross-linked together with nucleosomes again. In the third experiment two samples with only Mer2, two samples with Mer2 and Spp1 and two more samples with Mer2, Spp1 and nucleosomes were cross-linked with BS3-d_0/d_{12} again, with a higher cross-linker concentration. The fourth experiment was done with the MS-cleavable cross-linker DSBU. In one sample Mer2 was cross-linked alone, in the second sample Mer2 and Spp1 were cross-linked and in the third sample

Mer2, Spp1 and nucleosomes were cross-linked together again. The BSS2G and BS3 datasets were mainly analyzed with OpenPepXL. To make sure we are not missing out on identifications due to software issues some data was additionally searched with pLink2. The DSBU data was analyzed with MeroX.[36]

Results

Many intra-protein cross-links were identified in these datasets. The structures of these proteins are not fully resolved, so this additional information is helpful to get insight into their structure. A total of 9 unique linked residue pairs were identified within the Spp1 sequence, while a total of 112 were identified in Mer2. Of the cross-linked Mer2 peptide pairs 10 had overlaps in their sequences, meaning the two peptides must have been from two different copies of the protein. These were found to link a Mer2 homodimer at multiple positions between 79 and 141, which is consistent with a parallel dimerization. All 9 links in Spp1 were identified in the first experiment and confirmed in some of the others. Of the Mer2 cross-links 110 were identified in the first two experiments and many of them were confirmed in the other two, with only two new ones identified in the third experiment. Comparing the identified cross-links of OpenPepXL and pLink2, the trend was similar to the results of the benchmarks. For example in the first BS3 sample of the second experiment, OpenPepXL identified 34 linked residue pairs and pLink2 32. No inter-protein links between Mer2 and Spp1 were identified by pLink2 either. This trend continued for most of the samples with pLink2 finding a few more cross-links in a few cases. This shows that OpenPepXL works on real world data comparably well as on the benchmark datasets and the search algorithms were most likely not the problem in this experiment.

Unfortunately, one of the major goals of these experiments was to study the interactions between these proteins, but no cross-links between Spp1 and Mer2 or the nucleosome could be reliably identified in any of the experiments. Hadeer Elhabashy used AlphaFold[85] to model a Mer2 monomer. The inter-protein links between Mer2 subunits were used to create a dimer model with Haddock 2.2[86] (shown in Fig. 5.2A). I suggested a possible tetramer model shown in Fig. 5.2B, C and D, based on what is known about its interaction sites. The two Spp1 subunits are expected to bind to the Mer2 tetramer at the top and bottom of Fig. 5.2C and D, roughly midway along the bundle of coiled coils, since that is where V195 is located.

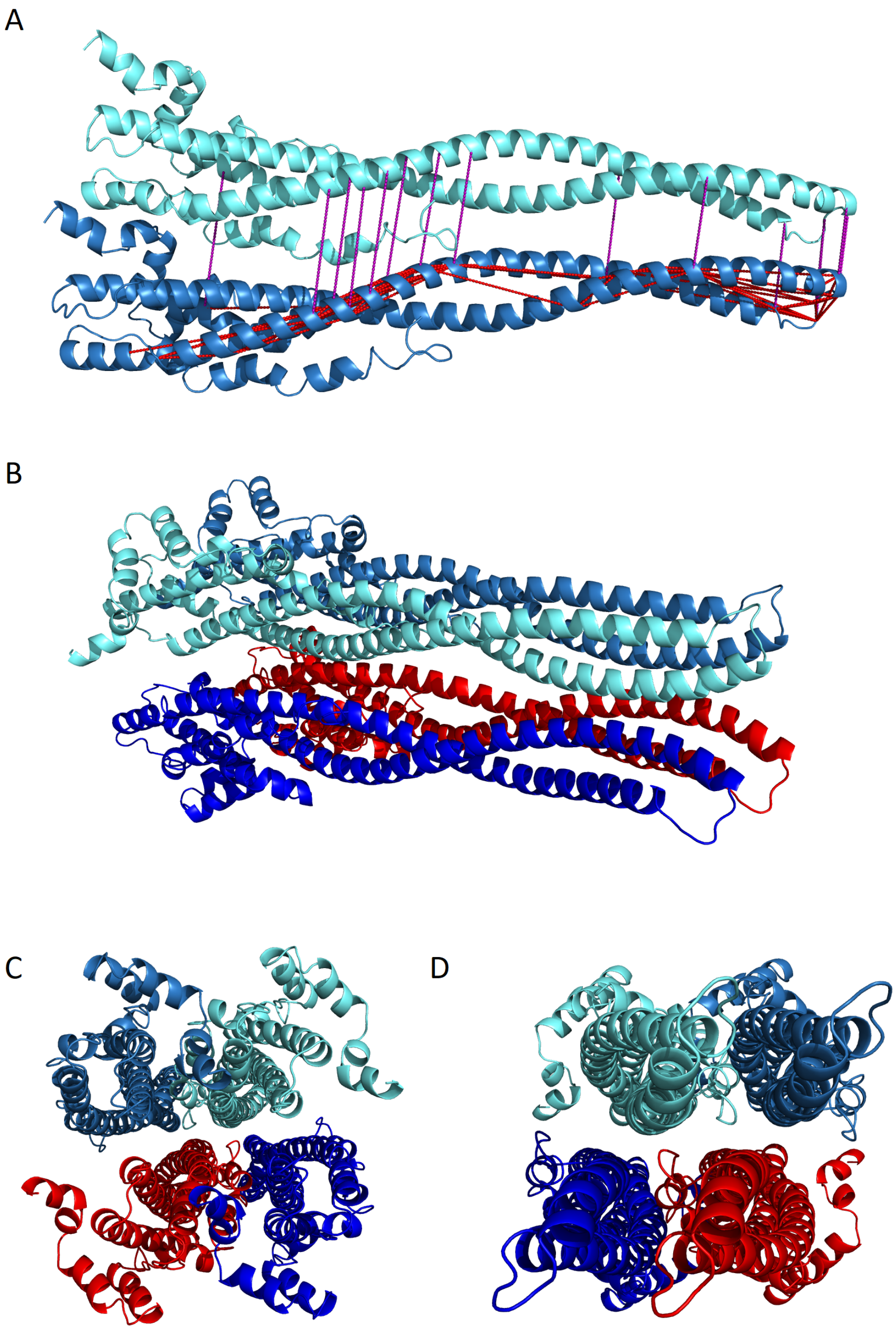


Figure 5.2: Mer2 homomultimer structures. A) Mer2 dimer with intra-protein cross-links in red and inter-protein cross-links in purple. B) C) and D) Suggested model for the Mer2 tetramer structure. C) and D) show the structure in B) from the left and right ends.

5.2.3 Miotic Replisome, Mer3 with Mlh1-Mlh2

Background

In *S. cerevisiae* the meiotic Mer3 helicase interacts with the mismatch repair related MutLβ complex Mlh1-Mlh2 and recruits it to recombination hotspots. Mer3 and MutLβ preferentially recognize displacement loops and contribute to the limitation of gene conversion due to meiotic recombination.[87] How Mer3 and Mlh1-Mlh2 interact is not known and was the main focus of this experiment.

Experiments

In this experiment the complex of Mer3 with Mlh1-Mlh2 was analyzed. The first sample only contained cross-linked Mer3. In the second sample the Mlh1-Mlh2 complex was cross-linked and in the third sample Mer3 was cross-linked together with Mlh1-Mlh2. Unlabeled BS3 was used in this experiment.

Results

The results in these experiments were very similar to the Mer2-Spp1 case. In the first experiment 103 cross-links were identified within Mer3. 254 cross-links were identified within the Mlh1-Mlh2 complex. Of those, 35 were inter-protein links between Mlh1 and Mlh2. No cross-links between Mer3 and Mlh1 or Mlh2 were found.

5.2.4 Miotic Replisome, Top3-Rmi1-Dmc1

Background

After the initial formation of a displacement loop, it is often dissolved by the proteins Sgs1, Top3 and Rmi1.[88] It is likely that Top3 and Rmi1 interact with Mer3 and/or Mlh1-Mlh2, but how they interact is unknown. Dmc1 is the central recombinase in the homologous recombination of double-stranded DNA breaks during meiosis and is also known to interact with Mer3. The target of this experiment was to find and model interactions between the Mer3-Mhl1-Mhl2 complex and Top3, Rmi1, or Dmc1.

Experiments

In the first experiment the complex of Mer3 and Mlh1-Mlh2 was analyzed together with Top3-Rmi1 using the MS-cleavable cross-linker DSBU. In the first sample Top3, Rmi1, Mlh1 and Mlh2 were cross-linked together without Mer3. In the second sample

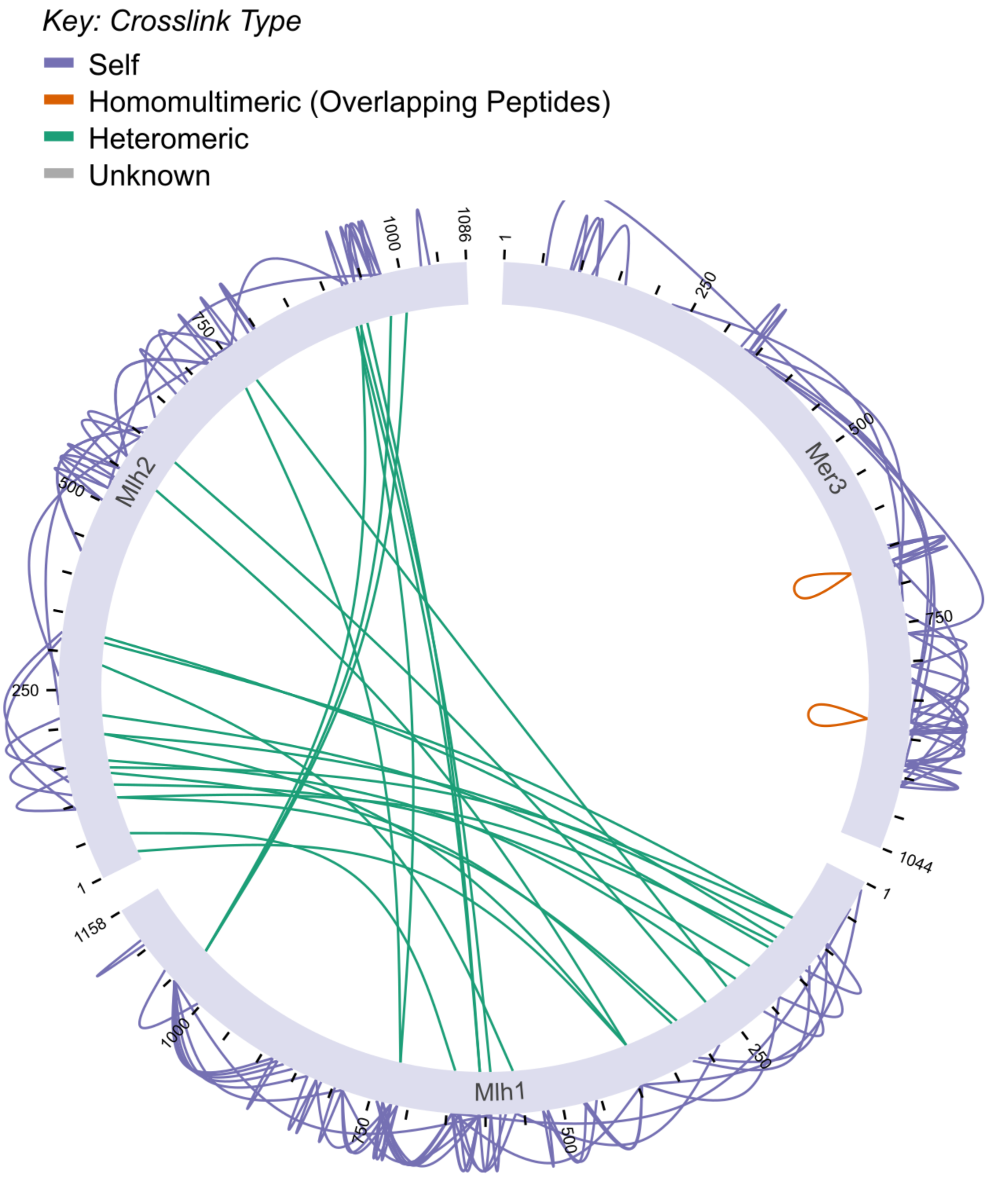


Figure 5.3: Network of cross-links identified in the Mer3-Mlh1-Mlh complex. No cross-links were identified between Mer3 and the other two proteins. Figure created with XiView.[47]

Mer3 was added as well. The third, fourth, fifth and sixth samples were cross-linked in the presence of DNA. In the third sample Mer3 and Mlh1-Mlh2 were cross-linked, in the fourth Mer3, Top3 and Rmi1, in the fifth Top3, Rmi1, Mlh1 and Mlh2 and finally in the last sample Mer3, Top3, Rmi1, Mlh1 and Mlh2. In the second experiment the complex of Mer3, Top3 and Rmi1 was analyzed together with Dmc1 using unlabeled

BS3. In the first sample Mer3 was cross-linked alone, in the second Dmc1 was cross-linked alone and in the third Top3 was cross-linked with Rmi1. In the fourth and fifth samples Mer3 was cross-linked together with Dmc1 and in the last two samples Mer3 was cross-linked together with Dmc1, Top3 and Rmi1.

The data analysis was done with OpenPepXL with mostly default settings. The cross-linker definition and the enzyme settings were adjusted to each experiment. MeroX[36] was additionally used to confirm the results of experiments using the MS-cleavable cross-linker DSBU, because OpenPepXL did not consider fragment ions produced by cross-link cleavage at the time.

Results

This time some of the Mer3 and Mlh1-Mlh2 cross-links from the previous study were confirmed, but still no inter-protein links between Mer3 and Mlh1 or Mlh2 could be identified. Additionally, 6 cross-links within Rmi1 and 11 within Top3 were identified. One inter-protein link between Rmi1 and Top3 was also found. In the second experiment 18 cross-links within Dmc1 were identified, 5 of them were peptide pairs with overlapping sequences and therefore show potential contacts between multiple Dmc1 proteins. Other than that, some cross-links in Rmi1, Top3 and Mer3 were confirmed again.

Like in the Mer2-Spp1 case, we could not find any cross-links between Mer3 and any of the other proteins. Interactions between Dmc1 and other proteins could also not be identified. Although we gained some structural insights into these protein complexes, the interactions between Mlh1 and Mlh2, or Rmi1 and Top3 are already known and were not the focus of these experiments. The cross-links we found were consistent with prior knowledge about these complexes so it seems that the experiments went well, but it looks like no cross-links could be formed between Mer3 and Dmc1 and the other two complexes.

5.3 Photo-Cross-Linking of the Type III Secretion System

5.3.1 Background

The type III secretion system (T3SS) of *Salmonella* is a large protein complex that likely shares a common ancestor with the bacterial flagellum. Most of the base of the system is built as multiple circular homomultimeric protein complexes that are stacked upon each other.[25] The homomultimeric complexes built from PrgH and PrgK are two such ring systems that are in direct contact with each other. This is a well

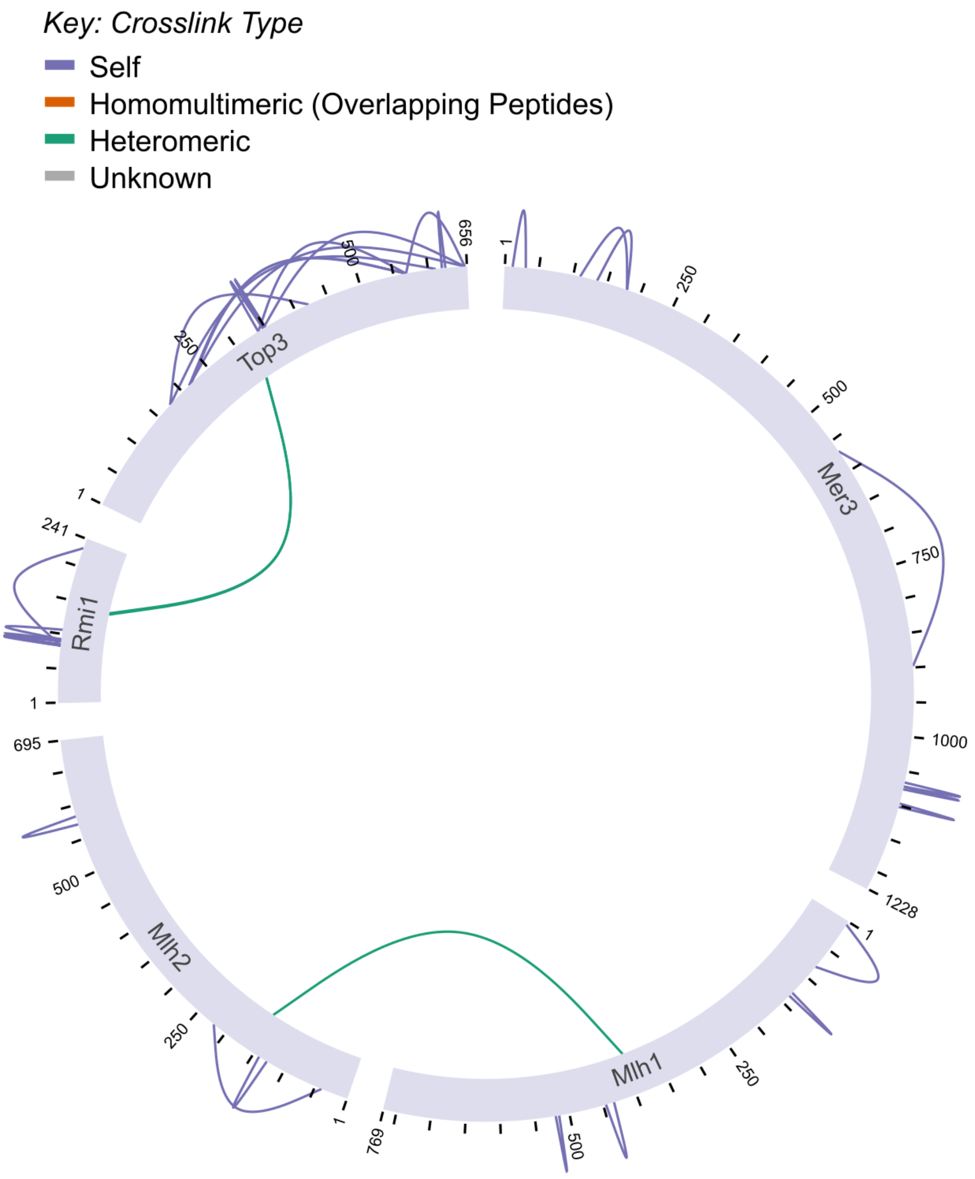


Figure 5.4: Network of cross-links identified in the Mer3-Mlh1-Mlh complex with Top3 and Rmi1. Figure created with XiView.[47]

known interaction that can be considered to be resolved.[89] The genes coding for these and other proteins involved in this system are clustered together on the Salmonella pathogenicity island. OrgA is also one of these genes and is known to be essential for the invasion and secretion functions of the T3SS. SptP is also part of the cluster and is involved in mediating the recovery of the host cytoskeleton after infection. It requires the chaperone SicP to be stable enough to function.[90] SicP forms two homodimers that

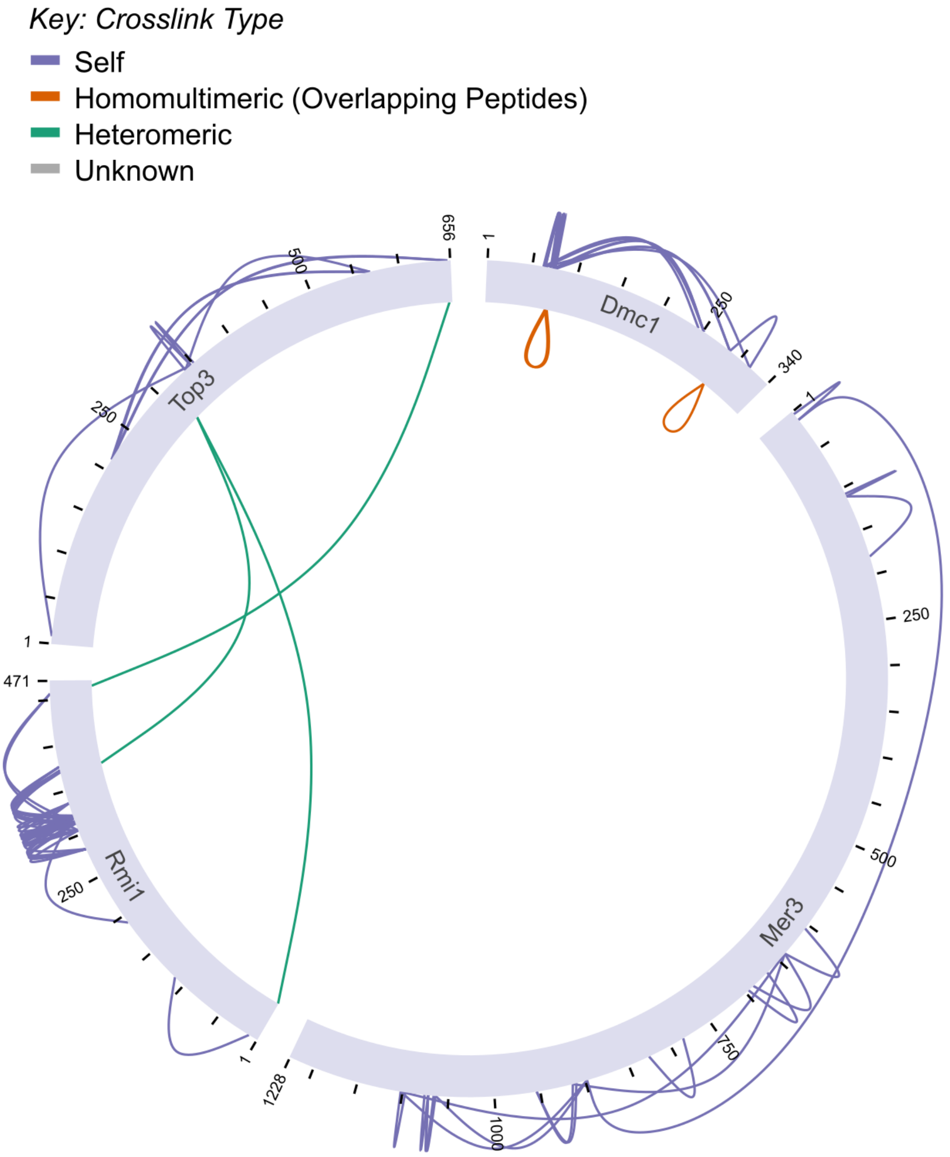
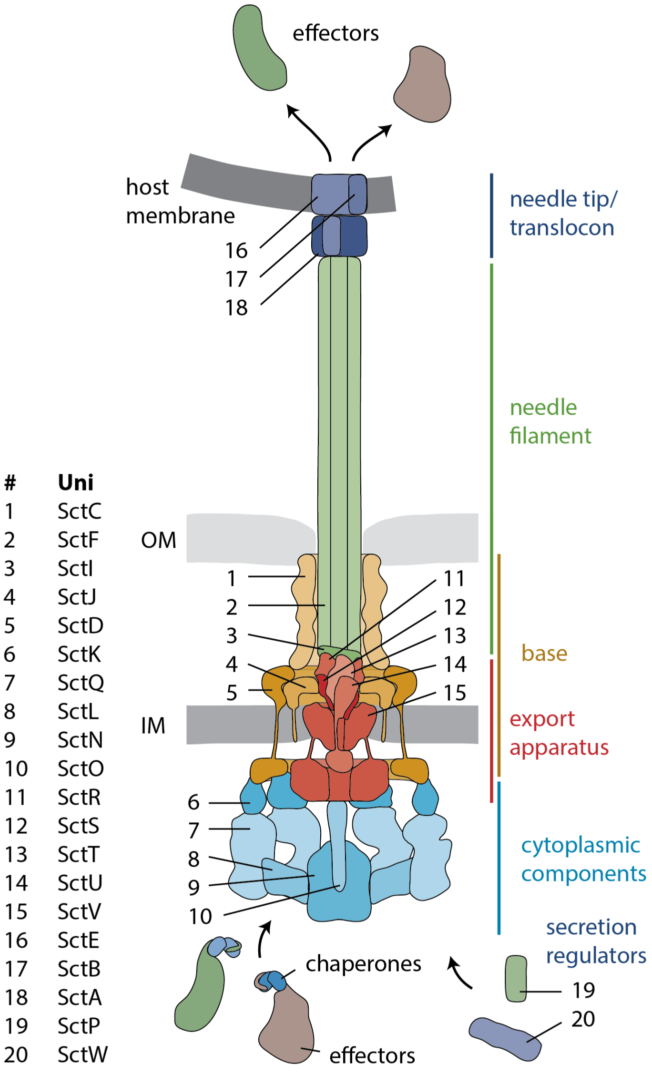


Figure 5.5: Network of cross-links identified in the experiment with Mer3, Top3 and Rmi1. Figure created with XiView.[47]

bind two SptP molecules. These two subunits interact through an interface between the two SptP molecules.

5.3.2 Experiments and Datasets

In these experiments the labeled cross-linker pBpa was used. This unnatural photo-cross-linking amino acid can be incorporated in a protein sequence *in vivo*.[93] Irradiation



#	Uni
1	SctC
2	SctF
3	SctI
4	SctJ
5	SctD
6	SctK
7	SctQ
8	SctL
9	SctN
10	SctO
11	SctR
12	SctS
13	SctT
14	SctU
15	SctV
16	SctE
17	SctB
18	SctA
19	SctP
20	SctW

Figure 5.6: A schematic overview of the Type III Secretion System.[91] The figure uses the suggested unified nomenclature of T3SS proteins.[92] SctJ is PrgK, SctD is PrgH and SctK is OrgA.

with UV light activates its side chain, which preferentially reacts with carbon-hydrogen bonds. It can therefore cross-link to the backbone of a protein. This results in a unique identification task, where one cross-linked residue in a pair is predefined by the incorporation of pBpa into the protein sequence and the other side can cross-link to any residue. This method facilitates the testing of very specific interaction hypotheses. However, it also has the risk of not yielding any useful information if the hypothesis can not be confirmed. The chemical composition of unlabeled pBpa is $C_{16}O_2NH_{13}$ and its residue mass is 251.0946212 u. In the stable isotope labeled version, nine of the C-atoms are replaced with ^{13}C isotopes and one N with its ^{14}N isotope, making the mass difference between the two molecules 10.0272342 *u*. Although pBpa has been used in various types of cross-linking experiments,[94] at the time of writing this thesis there are to our knowledge no published studies identifying the linked positions on a residue level resolution by directly matching the cross-link to a fragment spectrum. Sometimes elaborate approaches are used to arrive at specific link positions.[95] To make sure this approach is viable, in a first experiment we incorporated pBpa into PrgH, replacing the glutamine residue at position 179. We expected to find a cross-link to PrgK to confirm these cross-links can form and we can identify them. The second experiment was an attempt to find an interaction between PrgH and OrgA. For this the proline residue at position 68 of PrgH was replaced with pBpa. The third experiment was an attempt to find an interaction between SptP and its chaperone SicP. For this the phenylalanine at position 36 of SicP was replaced with pBpa.

The data analysis was done with OpenPepXL with default settings, except for the cross-linker definition. OpenPepXL does not have a setting to restrict the search for cross-links to only one specific position in a protein sequence. Therefore the residue type that was replaced by pBpa in each experiment was set cross-linkable by one side of the cross-link and the other side allowed all residue types to cross-link. The mass of the cross-link was also adjusted by the difference between the replaced residue type and pBpa. In the second and third experiments, pLink2 was also used to confirm the results.

5.3.3 Results and Discussion

The first experiment to confirm the viability of this approach was very successful. A cross-link between the pBpa residue at position 179 of PrgH and the proline at position 201 of PrgK was found. The same cross-link was identified in six different spectra below an FDR of 5% with several different peptide pairs containing the same positions. However, with such small numbers of cross-links FDR estimation is not reliable and we know one of the linked positions in advance. Therefore we filtered out cross-links to other glutamine residues and manually validated the CSMs. Many slightly less confident identifications also pointed towards the second linked position being the glycine at position 199 of PrgK. In total more than 30 CSMs identified a link between the pBpa on PrgH and the proline or glycine residues on PrgK. The two positions are very close to each other and since pBpa is able to link to the protein backbone, both cross-links may be correct.

In the second experiment with PrgH and OrgA, not a single cross-link could be identified with either OpenPepXL or pLink2.

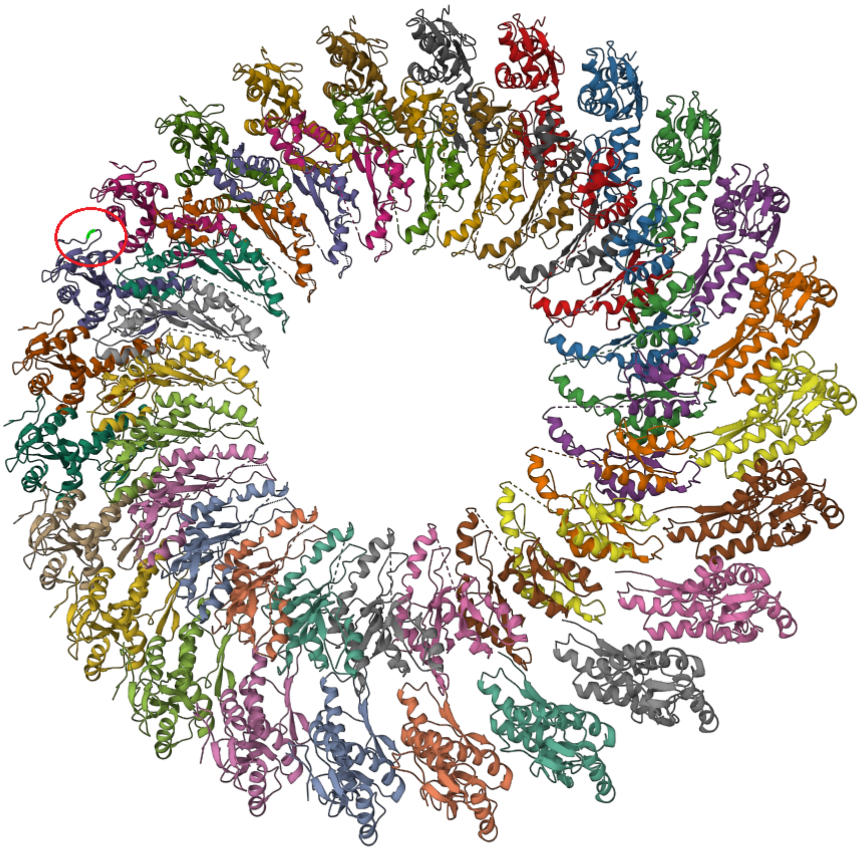


Figure 5.7: PrgH-PrgK complex (PDB ID: 3J6D). The outer ring is built from PrgH subunits, the inner ring from PrgK subunits. The green residue marked with a red circle is the replaced and cross-linked PrgH residue. The linked proline on PrgK is not resolved in this electron microscopy structure.

In the third experiment OpenPepXL identified a link between the pBpa residue at position 36 of SicP and a glutamine at position 75 of SptP. However, this link had a low score of 0.66, whereas an acceptable OpenPepXL score would usually lie above 0.70. We tried to confirm it using pLink2 and interestingly it identified a cross-link to the arginine at position 73 of SptP, but in a different fragment spectrum and also with a score that did not inspire confidence. The pLink score was 0.01294495, with a lower score being better. Usually, acceptable matches have a score that is at least an order of magnitude lower than this. Both tools reporting cross-links to two residues very close to each other suggests that these might be correct identifications, but with only one CSM per cross-link and both with a low match confidence, so far we do not have enough evidence to confirm it.

With these experiments we could show that using the unnatural photo-cross-linking amino acid pBpa is a viable method to confirm very specific hypotheses about protein interactions. However, they also showed that the very low number of expected cross-links can lead to an experiment that yields very little useful information, if the hypothesis can not be confirmed.

Chapter 6

Conclusion and Outlook

The modeling of protein structures is an important endeavor for our understanding of basic biological processes and therefore for essential for many medical and industrial applications. No one approach so far has proved comprehensive enough to be applicable to every protein complex. Integrative structure modeling is becoming more useful as new sources of structural information become available and older methods improve in efficiency, sensitivity and applicability.

This thesis describes the development and benchmarking of an XL-MS identification tool that improves upon the sensitivity of XL-MS identification for experiments targeting purified protein complexes with non-cleavable cross-linkers. The design philosophy of OpenPepXL is to do an exhaustive search without major shortcuts or heuristics that could reduce its sensitivity. Within this constraint it has to be efficient enough to be usable on standard desktop PCs for most applications and this was achieved for experiments studying purified protein complexes. As a consequence of a thorough spectrum matching algorithm that considers relative mass tolerances and ion charge states determined from isotopic patterns or preprocessing of spectra pairs from labeled cross-linkers, OpenPepXL shows a competitive specificity even though it exhaustively scores all candidate peptide pairs for a given spectrum. The combination of the exploration of the entire search space, very strict criteria for matching peaks between theoretical and experimental spectra and efficient data structures and algorithms makes OpenPepXL a sensitive tool with feasible runtime and memory requirements. Additionally it supports CID cleavable cross-linkers that enable proteome wide studies. To facilitate manual inspection and validation of CSMs, as well as the preparation of figures, an interactive visualization of CSMs was implemented into the TOPP tool TOPPView. OpenPepXL supports the open standardized XML format MzIdentML 1.2 for XL-MS identifications, which was developed with members of the Proteomics Standard Initiative. This format is starting to get widespread adoption among XL-MS identification and visualization tools and will hopefully improve the interoperability of software within the XL-MS field.

In comparisons with other state-of-the-art XL-MS identification tools OpenPepXL proved to be competitive in both sensitivity and efficiency. Using multiple datasets with available XRC structures to validate identified cross-links, OpenPepXL reported the highest numbers of structurally validated cross-links. In some cases the results of structurally validated datasets were ambiguous, because of the rigid nature of protein structure models and missing structural information for many residues in those proteins. A synthetic peptide dataset enabled us to study OpenPepXLs performance in more detail with confirmed true positive and false positive identifications. It showed that OpenPepXL has the highest sensitivity without sacrificing specificity compared to the next most sensitive tools. We discovered that a part of that success is due to OpenPepXL assigning target CSMs to more spectra than the other tools before FDR estimation and it might be possible to further improve the performance of OpenPepXL by tweaking the FDR formula. Compared to that, pLink2 assigned the smallest number of CSMs even before FDR estimation. Such heavy use of heuristics and the practice of discarding many candidate matches as early as possible makes the tool very efficient and fast, but that also causes it to discard some true positive candidates even before applying an FDR cut-off.

The implementation of OpenPepXL still has room for improvement and there are still a few ways to make it more efficient without sacrificing its advantage in sensitivity. Some concepts already common to proteomics data analysis like ion indices are already employed by several of the other XL-MS identification tools and could further improve OpenPepXL in the future.

We applied XL-MS to several protein complexes that proved too difficult to tackle with other methods so far. Although in most cases we could not identify the inter-protein cross-links we were hoping for, we still gained a lot of structural information about the proteins and subcomplexes involved. We could also confirm that the photo-cross-linking amino acid pBpa is a viable option for the purpose of testing specific hypotheses about protein interactions. The cross-links it forms can be identified by XL-MS identification software on a residue level resolution. This can be used to probe interactions in a tightly controlled way, without relying on specific residue types being available on the interaction site.

In conclusion, we believe that the work presented in this thesis contributes to the advancement of structural proteomics and will also positively impact the wider field of integrative protein structure modeling.

Bibliography

[1] Smyth M. and Martin J. (2000). X Ray crystallography. *Molecular Pathology*, 53(1):8. 2

[2] Sugiki T., Kobayashi N., and Fujiwara T. (2017). Modern technologies of solution nuclear magnetic resonance spectroscopy for three-dimensional structure determination of proteins open avenues for life scientists. *Computational and structural biotechnology journal*, 15:328–339. 2

[3] Benjin X. and Ling L. (2020). Developments, applications, and prospects of cryo-electron microscopy. *Protein Science*, 29(4):872–882. 2

[4] Hopf T. A., et al. (2019). The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics*, 35(9):1582–1584. 2

[5] Senior A. W., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710. 2

[6] Evans R., et al. (2021). Protein complex prediction with AlphaFold-Multimer. *BioRxiv*, pages 2021–10. 2

[7] Brückner A., Polge C., Lentze N., Auerbach D., and Schlattner U. (2009). Yeast two-hybrid, a powerful tool for systems biology. *International journal of molecular sciences*, 10(6):2763–2788. 3

[8] Li Y. (2011). The tandem affinity purification technology: an overview. *Biotechnology letters*, 33(8):1487–1499. 3

[9] O' Reilly F. J. and Rappsilber J. (2018). Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. *Nature structural & molecular biology*, 25(11):1000–1008. 3

[10] Iacobucci C., Götze M., and Sinz A. (2020). Cross-linking/mass spectrometry to get a closer view on protein interaction networks. *Current opinion in biotechnology*, 63:48–53.

[11] Leitner A., et al. (2010). Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Molecular & Cellular Proteomics*, 9(8):1634–1649. 4

[12] Yu C. and Huang L. (2018). Cross-linking mass spectrometry (XL-MS): An emerging technology for interactomics and structural biology. *Analytical chemistry*, 90(1):144. 3

[13] Liu F., Rijkers D. T., Post H., and Heck A. J. (2015). Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nature methods*, 12(12):1179–1184. 3, 4

[14] Jiao F., et al. (2022). Two-dimensional fractionation method for proteome-wide cross-linking mass spectrometry analysis. *Analytical chemistry*, 94(10):4236–4242.

[15] Rey M., et al. (2021). Advanced in vivo cross-linking mass spectrometry platform to characterize proteome-wide protein interactions. *Analytical Chemistry*, 93(9):4166–4174. 3

[16] Renaud J.-P., et al. (2018). Cryo-EM in drug discovery: achievements, limitations and prospects. *Nature reviews Drug discovery*, 17(7):471–492. 3

[17] Glaeser R. M. (2019). How good can single-particle cryo-EM become? What remains before it approaches its physical limits? *Annual review of biophysics*, 48:45–61. 3

[18] Rantos V., Karius K., and Kosinski J. (2022). Integrative structural modeling of macromolecular complexes using Assembline. *Nature Protocols*, 17(1):152–176. 4

[19] Vallat B., Webb B., Westbrook J., Sali A., and Berman H. M. (2019). Archiving and disseminating integrative structure models. *Journal of biomolecular NMR*, 73:385–398.

[20] Kosinski J., et al. (2016). Molecular architecture of the inner ring scaffold of the human nuclear pore complex. *Science*, 352(6283):363–365.

[21] Zhang Z., et al. (2020). Molecular architecture of the human 17S U2 snRNP. *Nature*, 583(7815):310–313. 4

[22] Maiolica A., et al. (2007). Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. *Molecular & Cellular Proteomics*, 6(12):2200–2211. 4

[23] Sommermeyer V., Béneut C., Chaplais E., Serrentino M. E., and Borde V. (2013). Spp1, a member of the Set1 Complex, promotes meiotic DSB formation in promoters by tethering histone H3K4 methylation sites to chromosome axes. *Molecular cell*, 49(1):43–54. 4, 74

[24] Rousova D., et al. (2021). Novel mechanistic insights into the role of Mer2 as the keystone of meiotic DNA break formation. *Elife*, 10:e72330. 4, 74

[25] Singh N. and Wagner S. (2019). Investigating the assembly of the bacterial type III secretion system injectisome by in vivo photocrosslinking. *International Journal of Medical Microbiology*, 309(6):151331. 5, 80

[26] Fritzsche R., Ihling C. H., Götze M., and Sinz A. (2012). Optimizing the enrichment of cross-linked products for mass spectrometric protein analysis. *Rapid Communications in Mass Spectrometry*, 26(6):653–658. 8

[27] Tayri-Wilk T., et al. (2020). Mass spectrometry reveals the chemistry of formaldehyde cross-linking in structured proteins. *Nature communications*, 11(1):1–9. 8

[28] Rivera-Santiago R. F., Sriswasdi S., Harper S. L., and Speicher D. W. (2015). Probing structures of large protein complexes using zero-length cross-linking. *Methods*, 89:99–111. 8

[29] Ihling C. H., et al. (2019). The isotope-labeled, MS-cleavable cross-linker disuccinimidyl dibutyric urea for improved cross-linking/mass spectrometry studies. *Journal of the American Society for Mass Spectrometry*, 31(2):183–189. 9

[30] Liu F., Lössl P., Scheltema R., Viner R., and Heck A. J. (2017). Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. *Nature communications*, 8(1):15473. 9

[31] Rinner O., et al. (2008). Identification of cross-linked peptides from large sequence databases. *Nature methods*, 5(4):315–318. 16, 29, 57, 58

[32] Leitner A., Walzthoeni T., and Aebersold R. (2014). Lysine-specific chemical cross-linking of protein complexes and identification of cross-linking sites using LC-MS/MS and the xQuest/xProphet software pipeline. *Nature protocols*, 9(1):120–137. 57

[33] Walzthoeni T., et al. (2012). False discovery rate estimation for cross-linked peptides identified by mass spectrometry. *Nature methods*, 9(9):901–903. 16, 17, 41, 57

[34] Kong A. T., Leprevost F. V., Avtonomov D. M., Mellacheruvu D., and Nesvizhskii A. I. (2017). MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nature methods*, 14(5):513–520. 17

[35] Götze M., et al. (2011). StavroX–a software for analyzing crosslinked products in protein interaction studies. *Journal of the American Society for Mass Spectrometry*, 23(1):76–87. 18, 29, 57, 58

[36] Iacobucci C., et al. (2018). A cross-linking/mass spectrometry workflow based on MS-cleavable cross-linkers and the MeroX software for studying protein structures and protein–protein interactions. *Nature protocols*, 13(12):2864–2889. 18, 69, 76, 80

[37] Hoopmann M. R., et al. (2015). Kojak: efficient analysis of chemically cross-linked protein complexes. *Journal of proteome research*, 14(5):2190–2198. 19, 29, 57

[38] Deutsch E. W., et al. (2015). Trans-proteomic pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *PROTEOMICS–Clinical Applications*, 9(7-8):745–754. 19

[39] Eng J. K., Jahan T. A., and Hoopmann M. R. (2013). Comet: an open-source MS/MS sequence database search tool. *Proteomics*, 13(1):22–24. 20

[40] Eng J. K., McCormack A. L., and Yates J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the american society for mass spectrometry*, 5(11):976–989. 20

[41] Keller A., Nesvizhskii A. I., Kolker E., and Aebersold R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical chemistry*, 74(20):5383–5392. 20

[42] Käll L., Canterbury J. D., Weston J., Noble W. S., and MacCoss M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature methods*, 4(11):923–925. 20, 24

[43] The M., MacCoss M. J., Noble W. S., and Käll L. (2016). Fast and accurate protein false discovery rates on large-scale proteomics data sets with percolator 3.0. *Journal of the American Society for Mass Spectrometry*, 27:1719–1727. 20, 24

[44] Fischer L. and Rappsilber J. (2017). Quirks of error estimation in cross-linking/mass spectrometry. *Analytical chemistry*, 89(7):3829–3833. 20, 22, 29, 45, 57

[45] Lenz S., Giese S. H., Fischer L., and Rappsilber J. (2018). In-search assignment of monoisotopic peaks improves the identification of cross-linked peptides. *Journal of proteome research*, 17(11):3923–3931.

[46] Mendes M. L., et al. (2019). An integrated workflow for crosslinking mass spectrometry. *Molecular Systems Biology*, 15(9):e8994. 20

[47] Graham M., Combe C., Kolbowski L., and Rappsilber J. (2019). xiView: A common platform for the downstream analysis of Crosslinking Mass Spectrometry data. *bioRxiv*, page 561829. 20, 45, 60, 79, 81, 82

[48] Chen Z.-L., et al. (2019). A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nature communications*, 10(1):1–12. 22, 29, 57, 58

[49] Yuan Z.-F. e., et al. (2012). pParse: A method for accurate determination of monoisotopic peaks in high-resolution mass spectra. *Proteomics*, 12(2):226–235. 22

[50] Chi H., et al. (2018). Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nature biotechnology*, 36(11):1059–1061. 23

[51] Yang B., et al. (2012). Identification of cross-linked peptides from complex samples. *Nature methods*, 9(9):904–906. 23

[52] Fu Y., et al. (2004). Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. *Bioinformatics*, 20(12):1948–1954. 24

[53] Röst H. L., et al. (2016). OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nature methods*, 13(9):741–748. 24

[54] Kohlbacher O., et al. (2007). TOPP–the OpenMS proteomics pipeline. *Bioinformatics*, 23(2):e191–e197. 24

[55] Karlsson B. *Beyond the C++ standard library: an introduction to boost*. Pearson Education (2005). 25

[56] Chapman B., Jost G., and Van Der Pas R. *Using OpenMP: portable shared memory parallel programming*. MIT press (2008). 25

[57] Martens L., et al. (2011). mzML–a community standard for mass spectrometry data. *Molecular & Cellular Proteomics*, 10(1). 26

[58] Lin S. M., Zhu L., Winter A. Q., Sasinowski M., and Kibbe W. A. (2005). What is mzXML good for? *Expert review of proteomics*, 2(6):839–845. 26

[59] Vizcaíno J. A., et al. (2017). The mzIdentML data standard version 1.2, supporting advances in proteome informatics. *Molecular & cellular proteomics*, 16(7):1275–1285. 27, 45

[60] Griss J., et al. (2014). The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Molecular & Cellular Proteomics*, 13(10):2765–2775. 27

[61] Perez-Riverol Y., et al. (2019). The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic acids research*, 47(D1):D442–D450. 27, 54, 56

[62] Deutsch E. W., et al. (2020). The ProteomeXchange consortium in 2020: enabling 'big data'approaches in proteomics. *Nucleic acids research*, 48(D1):D1145–D1152. 27, 54, 56

[63] Pettersen E. F., et al. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612. 27, 57

[64] Kosinski J., et al. (2015). Xlink Analyzer: software for analysis and visualization of cross-linking data in the context of three-dimensional structures. *Journal of structural biology*, 189(3):177–183. 27, 57

[65] Ferrari A. J., et al. (2019). TopoLink: evaluation of structural models using chemical crosslinking distance constraints. *Bioinformatics*, 35(17):3169–3170. 28, 57

[66] Netz E., et al. (2020). OpenPepXL: An open-source tool for sensitive identification of cross-linked peptides in XL-MS. *Molecular & Cellular Proteomics*, 19(12):2157–2168. 50, 53, 58, 59, 63, 64, 65, 66, 68, 69, 100

[67] Cao X. and Nesvizhskii A. I. (2008). Improved sequence tag generation method for peptide identification in tandem mass spectrometry. *Journal of proteome research*, 7(10):4422–4434. 35

[68] Eng J. K., Fischer B., Grossmann J., and MacCoss M. J. (2008). A fast SEQUEST cross correlation algorithm. *Journal of proteome research*, 7(10):4598–4602. 37

[69] Craig R. and Beavis R. C. (2004). TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467. 37

[70] Beausoleil S. A., Villén J., Gerber S. A., Rush J., and Gygi S. P. (2006). A probability–based approach for high-throughput protein phosphorylation analysis and site localization. *Nature biotechnology*, 24(10):1285–1292. 37

[71] Grünberg R. (2022). Analysis of cross-linking mass spectrometry data with cleavable cross-linkers using OpenMS [Master Thesis]. *University of Tübingen*. 45, 69, 70

[72] Monecke T., et al. (2009). Crystal structure of the nuclear export receptor CRM1 in complex with Snurportin1 and RanGTP. *Science*, 324(5930):1087–1091. 53, 57

[73] Kolbowski L., Mendes M. L., and Rappsilber J. (2017). Optimizing the parameters governing the fragmentation of cross-linked peptides in a tribrid mass spectrometer. *Analytical chemistry*, 89(10):5311–5318. 54

[74] Iacobucci C., et al. (2019). First community-wide, comparative cross-linking mass spectrometry study. *Analytical chemistry*, 91(11):6953–6961. 55

[75] Beveridge R., Stadlmann J., Penninger J. M., and Mechtler K. (2020). A synthetic peptide library for benchmarking crosslinking-mass spectrometry search engines for proteins and protein complexes. *Nature communications*, 11(1):1–9. 55, 56, 63, 65, 66, 67, 68, 69

[76] Chambers M. C., et al. (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nature biotechnology*, 30(10):918–920. 55

[77] Burley S. K., et al. (2021). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic acids research*, 49(D1):D437–D451. 56

[78] Camacho C., et al. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, 10(1):1–9. 57

[79] White S. H. (2004). The progress of membrane protein structure determination. *Protein Science*, 13(7):1948–1949. 73

[80] Das B. B., Park S. H., and Opella S. J. (2015). Membrane protein structure from rotational diffusion. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1848(1):229–245.

[81] Uversky V. N. (2019). Intrinsically disordered proteins and their ”mysterious” (meta) physics. *Frontiers in Physics*, 7:10.

[82] Ruff K. M. and Pappu R. V. (2021). AlphaFold and implications for intrinsically disordered proteins. *Journal of Molecular Biology*, 433(20):167208. 73

[83] Andersen S. L. and Sekelsky J. (2010). Meiotic versus mitotic recombination: Two different routes for double-strand break repair: The different functions of meiotic versus mitotic DSB repair are reflected in different pathway usage and different outcomes. *Bioessays*, 32(12):1058–1066. 74

[84] Adam C., et al. (2018). The PHD finger protein Spp1 has distinct functions in the Set1 and the meiotic DSB formation complexes. *PLoS genetics*, 14(2):e1007223. 74

[85] Jumper J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589. 76

[86] Van Zundert G., et al. (2016). The HADDOCK2. 2 web server: user-friendly integrative modeling of biomolecular complexes. *Journal of molecular biology*, 428(4):720–725. 76

[87] Duroc Y., et al. (2017). Concerted action of the MutLβ heterodimer and Mer3 helicase regulates the global extent of meiotic gene conversion. *Elife*, 6:e21900. 78

[88] Arora K. and Corbett K. D. (2019). The conserved XPF: ERCC1-like Zip2: Spo16 complex controls meiotic crossover formation through structure-specific DNA binding. *Nucleic acids research*, 47(5):2365–2376. 78

[89] Bergeron J. R., et al. (2015). The modular structure of the inner-membrane ring component PrgK facilitates assembly of the type III secretion system basal body. *Structure*, 23(1):161–172. 81

[90] Johnson R., et al. (2017). The type III secretion system effector SptP of Salmonella enterica serovar Typhi. *Journal of bacteriology*, 199(4):e00647–16. 81

[91] "*A schematic overview of the type III secretion system" from https://t3sswiki.science/index.php/File:T3SS_Overview.png, licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/legalcode)* (2023). 83

[92] Wagner S. and Diepold A. (2020). A unified nomenclature for injectisome-type type III secretion systems. *Bacterial Type III Protein Secretion Systems*, pages 1–10. 83

[93] Chin J. W., Martin A. B., King D. S., Wang L., and Schultz P. G. (2002). Addition of a photocrosslinking amino acid to the genetic code of Escherichia coli. *Proceedings of the National Academy of Sciences*, 99(17):11020–11024. 82

[94] Miyazaki R., Akiyama Y., and Mori H. (2020). A photo-cross-linking approach to monitor protein dynamics in living cells. *Biochimica Et Biophysica Acta (BBA)-General Subjects*, 1864(2):129317. 84

[95] Freinkman E., Chng S.-S., and Kahne D. (2011). The complex that inserts lipopolysaccharide into the bacterial outer membrane forms a two-protein plug-and-barrel. *Proceedings of the National Academy of Sciences*, 108(6):2486–2491. 84

Appendix A

Abbreviations

XRC	*X-Ray Crystallography*
NMR	*Nuclear Magnetic Resonance*
EC	*Evolutionary Coupling*
FDR	*False Discovery Rate*
XL-MS	*Cross-linking coupled with mass spectrometry*
LC	*Liquid Chromatography*
MS	*Mass Spectrometry*
RT	*Retention Time*
CID	*Collision induced dissociation*
HCD	*Higher-energy collisional dissociation*
PSM	*Peptide-Spectrum-Match*
CSM	*Cross-link-Spectrum-Match*
URP	*Unique Residue Pair*
MS1	*Precursor spectrum, measurement of full species*
MS2	*tandem MS spectrum, MS/MS spectrum*
DSS	*disuccinimidyl suberate*
BS3	*bis(sulfosuccinimidyl)suberate*
DSBU	*disuccinimidyl dibutyric urea*
pBpa	*p-benzoyl-L-phenylalanine*
PDH	*pimelic acid dihydrazide*
NHS	*N-hydroxysuccinimide*
IWS	*inconsistent with the protein structure*
PPI	*Protein-Protein interaction*
TOPP	*The OpenMS Proteomics Pipeline*
GUI	*Graphical User Interface*

MBP	*Maltose binding protein*
DSB	*Double-strand break*

Appendix B

Contributions

All ideas, approaches and results presented in this work were developed and discussed with my supervisor Prof. Dr. Oliver Kohlbacher (OK). The following co-workers also contributed to the different projects:

Timo Sachsenberg	(TS)
Tjeerd M. H. Dijkstra	(TD)
Ruben Grünberg	(RG)
Mathias Walzer	(MW)
Lukas Zimmermann	(LZ)
Ralf Ficner	(RF)
Thomas Monecke	(TM)
Henning Urlaub	(TM)
Olexandr Dybkov	(TM)
Hadeer Elhabashy	(HE)
John Weir	(JW)
Dorota Rousova	(DR)
Magdalena Firlej	(MF)
Veronika Altmannova	(VA)
Samuel Wagner	(SW)

- **Development of OpenPepXL**

  The project was designed by myself and OK. Development and implementation of XL-MS data analysis algorithms was performed by myself. The implementation of the MzIdentML 1.2 XL-MS identification format was performed by myself and

MW. The scoring function of OpenPepXL was designed by myself and TD. The visualization of labeled spectra in TOPPView was implemented by myself and TS. The XL-MS FDR estimation tool XFDR was initially implemented by LZ and then debugged, improved and maintained by myself. RG implemented the search for cleavable cross-linkers as part of his Master thesis that was supervised by me.

- **Benchmarking of OpenPepXL against other tools**

  The project was designed by myself and OK. The CRM complex experiment was designed by RF, TM, HU and OD. The protein expression and purification of the CRM complex was performed by TM and the cross-linking experiment and MS data acquisition were performed by OD. The XL-MS data analysis and interpretation was done by myself. The structural validation of the ribosomal fraction dataset was performed by myself and HE. The structural validation of all the other datasets was done by myself. The manuscript of the published article that arose from this work was written by myself[66]. RG benchmarked the search for cleavable cross-linkers as part of his Master thesis.

- **Protein complexes involved in meiosis**

  The Mer2-Spp1 complex experiments were designed by JW and DR. The protein expression, purification and cross-linking experiments for the Mer2-Spp1 complex were performed by DR. The experiments involving Mer3 with Mlh1-Mlh2 and Top3-Rmi1-Dmc1 were designed by JW, MF and VA. The protein expression, purification and cross-linking experiments for these complexes were performed by MF and VA. XL-MS data analysis for all these datasets was performed by myself. The Mer2 monomer and dimer models were made by HE, the tetramer model was made by myself.

- **Type III secretion system**

  The experiments were designed by SW and his group. The protein expression, purification and cross-linking experiments for these complexes was performed by SW's group. XL-MS data analysis for all these datasets was performed by myself.

Appendix C

Publications

Accepted manuscripts

Netz E., Dijkstra T. M., Sachsenberg T., Zimmermann L., Walzer M., Monecke T., Ficner R., Dybkov O., Urlaub H., Kohlbacher, O.
"OpenPepXL: An open-source tool for sensitive identification
of cross-linked peptides in XL-MS."
Molecular & Cellular Proteomics, 19(12), 2157-2168. (2020)

Vizcaíno J. A., Mayer G., Perkins S., Barsnes H., Vaudel M., Perez-Riverol Y., Ternent T., Uszkoreit J., Eisenacher M., Fischer L., Rappsilber J., **Netz E.**, Walzer M., Kohlbacher O., Leitner A., Chalkley R.J., Ghali F., Martínez-Bartolomé S.,
Deutsch E.W., Jones A. R.
"The mzIdentML data standard version 1.2,
supporting advances in proteome informatics."
Molecular & Cellular Proteomics, 16(7), 1275-1285. (2017)

Leitner A., Bonvin A. M. J. J., Borchers C., Chalkley R. J.,Chamot-Rooke J., Combe C. W., Cox J. Dong M. Q., Fischer L., Götze M., Gozzo F. C., Heck A. J. R., Hoopman M. R., Huang L., Ishihama Y., Jones A. R., Kalisman N., Kohlbacher O., Mechtler K., Moritz R. L., **Netz E.**, Novak P., Petrotchenko E., Sali A., Scheltema R. A., Schmidt C., Schriemer D., Sinz A., Sobott F., Stengel F., Thalassinos K., Urlaub H., Viner R., Vizcaíno J. A., Wilkins M. R., Rappsilber J.
"Toward Increased Reliability, Transparency, and Accessibility
in Cross-linking Mass Spectrometry."
Structure, 28(11), 1259-1268. (2020)

Zou X., Conrad L. J., Koschinsky K., Schlichthörl G., Preisig-Müller R., **Netz E.**, Krüger J., Daut J., Renigunta V.
"The phosphodiesterase inhibitor IBMX blocks the potassium channel THIK-1 from the extracellular side."
Molecular pharmacology, 98(2), 143-155. (2020)

Rodriguez N., Thomas A., Watanabe L., Vazirabad I. Y., Kofia V., Gómez H. F., Mittag F., Matthes J., Rudolph J., Wrzodek F., **Netz E.**, Diamantikos A., Eichner J., Keller R., Wrzodek C., Fröhlich S., Lewis N. E., Myers C. J., Novère N. L., Palsson B. Ø., Hucka M., Dräger A.
"JSBML 1.0: providing a smorgasbord of options to encode systems biology models."
Bioinformatics, 31(20), 3383-3386. (2015)

Other publications

Ashwood C., Bittremieux W., Deutsch E. W., Doncheva N. T., Dorfer V., Gabriels R., ...**Netz E.**... , Willems S.
"Proceedings of the EuBIC-MS 2020 Developers' Meeting."
EuPA Open Proteomics, 24, 1-6. (2020)

Alka O., Sachsenberg T., Bichmann L., Pfeuffer J., Weisser H., Wein S., ...**Netz E.**... , Röst H.
"OpenMS and KNIME for Mass Spectrometry Data Processing."
In Processing Metabolomics and Proteomics Data with Open Software (pp. 201-231). (2020)

Manuscripts in preparation

Sachsenberg T., Wein S., Bielow C., **Netz E.**, Walter A., ...
"OpenMS 3 expands the frontiers of open-source computational mass spectrometry"

Elhabashy H., **Netz E.**, Kohlbacher O.
"XLEC: Large-scale prediction and modeling of protein-protein interaction
by combining sequence co-evolution and cross-linking data"

Röhl A., **Netz E.**, Kohlbacher O., Elhabashy H.
"CLAUDIO: Software for prediction of protein homo-oligomers from Cross-linking data"

Hirth A., Fatti E., **Netz E.**, Acebron S.P., Papageorgiou D., ... Kohlbacher K., Niehrs C.
"DEAD box RNA helicases are pervasive serine/threonine protein kinase interactors and activators"

Appendix D

Supporting Figures

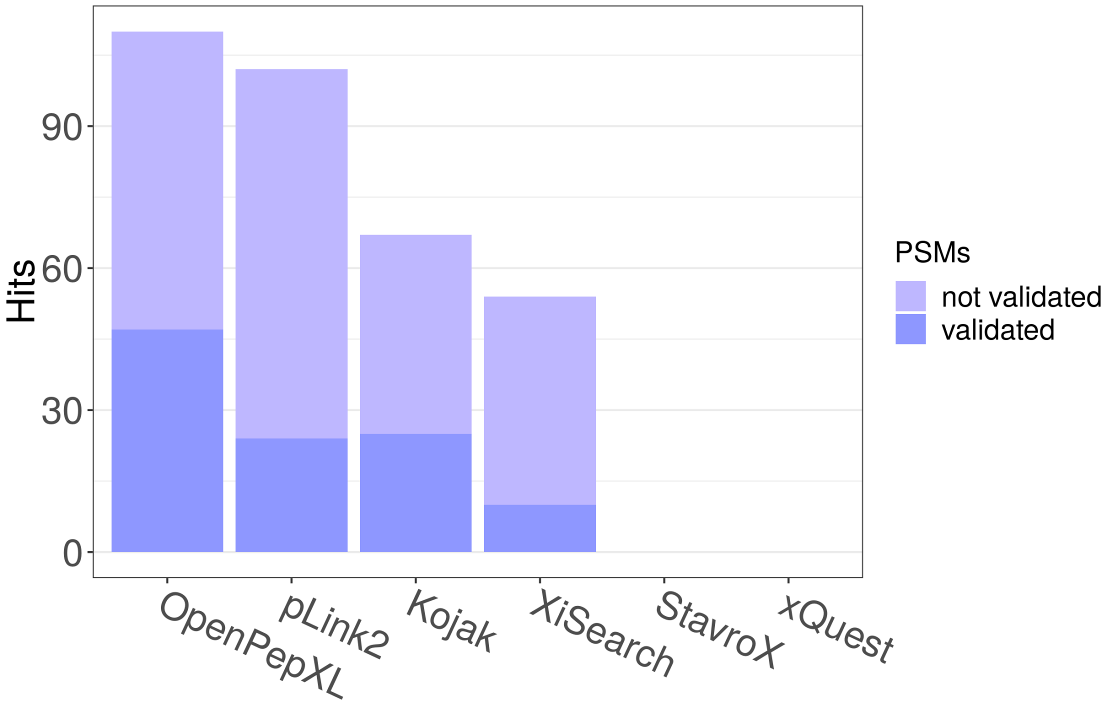


Figure D.1: Structurally validated unique residue pairs (URPs) from the ribosomal fraction data set. The validated URPs were covered by the structures we used for validation and none of them has a solvent accessible surface distance of more than 35 Å between Cβ atoms. The rest of the URPs were not covered by currently published protein structures. Full list of used structures (UniProt sequence ID : PDB ID): P05387:2W1O, P06748:2LLH, P06748:2P1B, P11142:4H5R, P12268:1B3O, P14868:4J15, P19338:2KRR, P19338:2FC9, P21333:3CNK, P23396:5AJ0, P27635:5AJ0, P30050:5AJ0, P39019:5AJ0, P54136:4R3Z, P56192:5GL7, P61353:5AJ0, P62081:5AJ0, P62249:5AJ0, P62277:5AJ0, P62280:5AJ0, P62424:5AJ0, P62701:5A2Q, P62847:5AJ0, P62851:5AJ0, P62888:5AJ0, P62906:5AJ0, P62906:5AJ0, P63173:5AJ0, P63244:5AJ0, P67809:5YTT, P83731:5AJ0, Q00610:2XZG, Q12904:1FL0, Q14152:3J8B, Q53YD7:5JPO, Q5U0F4:5K0Y, Q6PIN5:3J2I, Q7L2H7:3J8B.

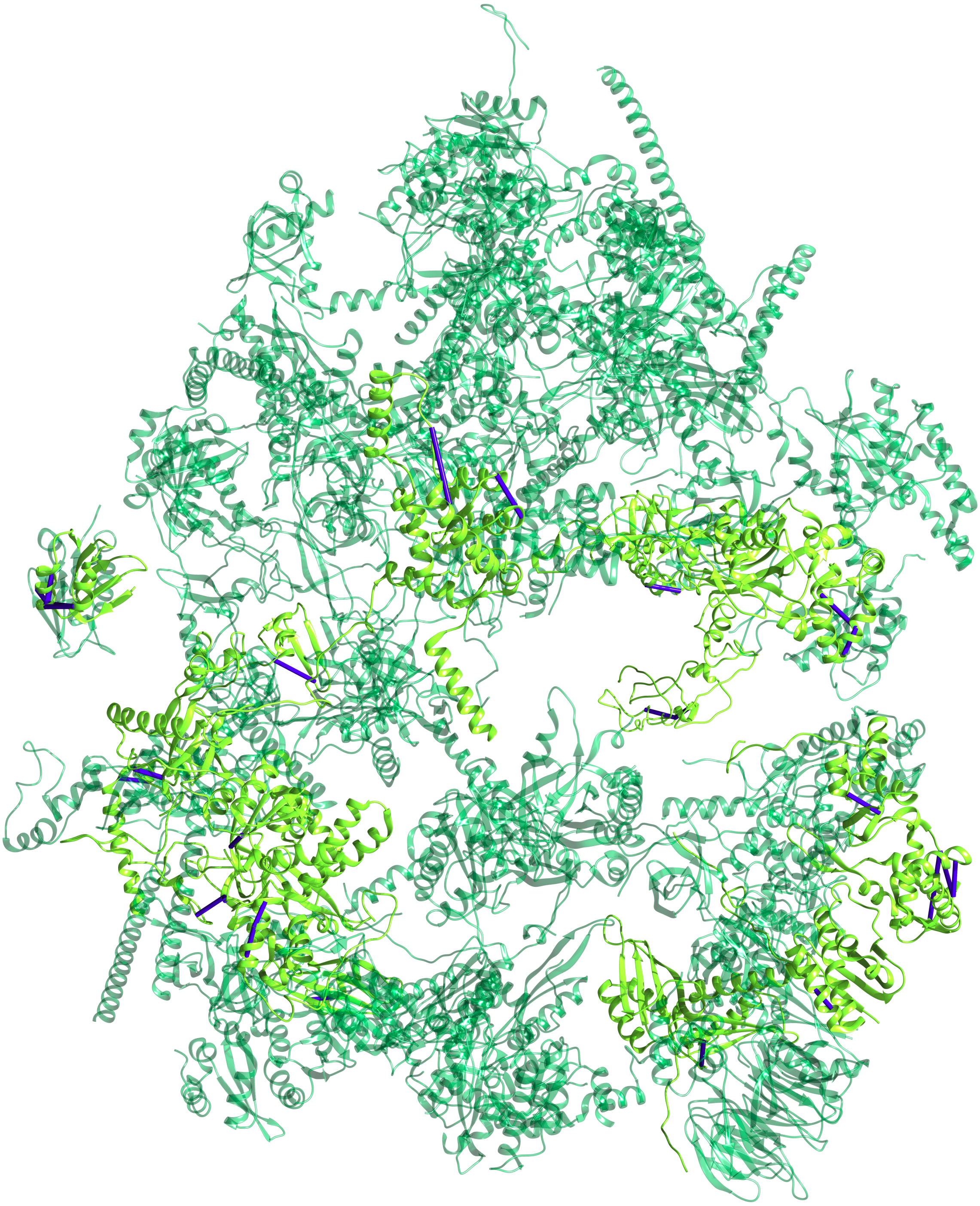


Figure D.2: Cross-links identified by OpenPepXL mapped onto the human ribosome structure (PDB ID: 5AJ0). The RNA was removed from the structure for visual clarity. Proteins without cross-links are shown in dark green, proteins with cross-links are shown in a lighter green. All cross-links are visualized as blue bars.

Appendix E

Supporting Tables

Table E.1: Search parameters for the ribosomal fraction dataset. The symbol '←' means this parameter for this tool was set to the same value as the first column. 'NA' means this setting is not available for this tool and the internal defaults can not be changed. The tools and tool versions were OpenPepXL 1.1, Kojak 1.6.0, pLink 2.3.5, and XiSearch 1.6.731

	OpenPepXL	**Kojak**	**pLink**	**XiSearch**
precursor mass mass tolerance	6 ppm	←	←	←
precursor charges	+3 to +8	NA	NA	NA
precursor monoisotopic peak corrections	0,1,2	0,1,2	NA	NA
fragment mass tolerance	20 ppm	←	←	←
fixed modifications	Carbamidomethyl (C)	←	←	←
variable modifications	Oxidation (M)	←	←	←
max variable mods per peptide	3	←	←	←
enzyme	Trypsin	←	←	←
min peptide length	5	NA	5	NA
max peptide length	NA	NA	600	NA
max missed cleavages	4	←	←	←
min peptide mass	NA	300	300	NA
max peptide mass	NA	50000	50000	NA
cross-linker name	BS3	←	←	←
cross-linker mass	138.0680	←	←	←
mono-link masses	156.0786, 155.0946	156.0786	156.079	156.0786, 155.0946
linked residues	K, N-term	←	←	←

Table E.2: Search parameters for the BSA dataset.The symbol '←' means this parameter for this tool was set to the same value as the first column. 'NA' means this setting is not available for this tool and the internal defaults can not be changed. The tools and tool versions were OpenPepXL 1.1 and xQuest 2.1.3.

	OpenPepXL	**OpenPepXL**	**xQuest**	**xQuest**	**xQuest**	**xQuest**
MS2 type	orbitrap	orbitrap	ion trap	orbitrap	ion trap	orbitrap
precursor mass tolerance	10 ppm	←	←	←	←	←
precursor charges	+3 to +8	←	←	←	←	←
precursor mono-isotopic peak corrections	0,1,2,3,4,5	←	NA	NA	NA	NA
fragment mass tolerance	20 ppm	20 ppm	0.2 Da	20 ppm	0.2 Da	20 ppm
cross-linked fragment mass tolerance	20 ppm	20 ppm	0.3 Da	20 ppm	0.3 Da	20 ppm
fixed modifications	Carbamido-methyl (C)	←	←	←	←	←
variable modifications	Oxidation (M)	←	←	←	←	←
max variable mods per peptide	2	←	←	←	←	←
enzyme	Trypsin	←	←	←	←	←
min peptide length	5	←	←	←	←	←
max missed cleavages	2	←	←	←	←	←
cross-linker name	DSS	PDH	DSS	DSS	PDH	PDH
cross-linker mass	138.0680	152.1061	138.0680	138.0680	152.1061	152.1061
mono-link masses	156.0786, 155.0946	170.1167	156.0786, 155.0946	156.0786, 155.0946	170.1167	170.1167
linked residues	K,S,T,Y, N-term	D,E, C-term	K,S,T,Y, N-term	K,S,T,Y, N-term	D,E, C-term	D,E, C-term
isotopeshift	12.0753	10.0627	12.0753	12.0753	10.0627	10.0627
ntermxlinkable	NA	NA	1	1	0	0
isopair_Mr_tolerance	NA	NA	15 ppm	15 ppm	15 ppm	15 ppm
isopair_Tr_tolerance	NA	NA	3	3	3	3

Table E.3: Search parameters for the CRM dataset.The symbol '←' means this parameter for this tool was set to the same value as the first column. 'NA' means this setting is not available for this tool and the internal defaults can not be changed. The tools and tool versions were OpenPepXL 1.1, Kojak 1.6.0, pLink 2.3.5, XiSearch 1.6.731, StavroX 3.6.6.5, and xQuest 2.1.3

	OpenPepXL	**Kojak**	**pLink2**	**XiSearch**	**StavroX**	**xQuest**
precursor mass tolerance	10 ppm	←	←	←	←	←
precursor charges	+3 to +8	NA	NA	NA	NA	+3 to +8
precursor monoisotopic peak corrections	0,1,2,3,4,5	←	NA	NA	NA	NA
fragment mass tolerance	20 ppm	←	←	←	←	←
fixed modifications	Carbamidomethyl (C)	←	←	←	←	←
variable modifications	Oxidation (M)	←	←	←	←	←
max variable mods per peptide	2	←	←	←	←	←
enzyme	Trypsin	←	←	←	←	←
min peptide length	5	NA	5	5	5	5
min peptide mass	NA	300	300	NA	300	NA
max peptide mass	NA	50000	50000	NA	50000	NA
max missed cleavages	2	←	←	←	←	←
cross-linker name	BS3	←	←	←	←	←
cross-linker mass	138.0680	←	←	←	←	←
mono-link masses	156.0786, 155.0946	156.0786	156.0786	156.0786, 155.0946	156.0786	156.0786, 155.0946
linked residues	K, N-term	←	←	←	←	←