New technologies for cross-linking mass spectrometry

Jared P. Mohr

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

James E. Bruce, Chair

Judit Villen

William Noble

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

New technologies for cross-linking mass spectrometry

Jared P. Mohr

Chair of the Supervisory Committee:
James E. Bruce
Genome Sciences

Cross-linking mass spectrometry is a rapidly evolving technique for obtaining structural information about proteins and protein complexes in their near native state in a high-throughput manner. Cross-linked peptides are challenging to identify due to being low abundance analytes that produce complex fragmentation spectra. While cross-links can provide valuable structural data, these challenges mean that current technologies can sample only a small portion of the protein structures and assemblies that exist in complex systems.

In this work I demonstrate three new technologies I developed to enhance cross-linking mass spectrometry experiments. The first section describes the development of the cross-linking search tool, Mango, which enables identification of cross-links in complex samples generated from a variety of cross-linkers. The next section discusses the development of a tetrameric cross-

linker and its application in studying the mitochondrial interactome from murine hearts. Higher dimensional cross-linkers move experiments from binary interactions to ternary and quaternary interactions, helping to better characterize interfaces composed of many proteins. The final section describes the development of software to facilitate liquid chromatography coupled to tandem mass spectrometry experiments in a Fourier transform ion cyclotron resonance array mass spectrometer by automating ion selection, ion transfer, and analog signal processing for each cell which allows for parallel acquisition of high-resolution mass spectra. While this hardware has been previously described, modifications to the instrument's ion handling and analog signal processing enable cross-linking experiments to be carried out with parallel detection and an enhanced duty cycle.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Graduate school is certainly a journey, and not one I undertook alone. There are many who helped along the way, and there's no way to thank you enough in so few words.

I would start by acknowledging the members of the Bruce lab, past and present, for creating an environment that encouraged innovation. To my advisor, Jim, thanks for taking me on for some summers before graduate school so I could get my feet wet before committing to years of my life doing this, and thanks for providing me with great latitude to experiment and experience everything your lab had to offer. To Juan, thanks for teaching me literally all the practical skills I possess as a mass spectrometrist and for teaching me the most useful skill of all: troubleshooting the NanoAcquity. To Sung-Gun, thanks for suffering through all the software problems I produced for the development instrument, and thanks for driving to Corvallis when Jim sent us on a scrap run. To my fellow graduate students, Martin and Anna, thanks for the conversation over hundreds of lunches and drinks. To Andy, thanks for entertaining my endless queries about different ways to analyze data, even if most of them were bad ideas.

The folks at the UWPR were an invaluable asset during my time in SLU. Priska was an endless font of practical knowledge on instrument maintenance and LC problems. Jimmy was kind enough to add several features to comet that only I will ever use and to bring me snacks he didn't like, but they both made my job a lot easier. Sincerely, thanks to both of you.

I give my thanks to my undergraduate research advisor, Michelle Heck, who introduced me to the world of mass spectrometry proteomics when she offered me a spot in her lab as an

undergraduate. I almost joined an organic chemistry lab, so you saved me from a truly cruel fate. It was through Michelle's lab that I became acquainted with Genome Sciences proteomics, and I would never have made it here without her guidance. It brings me no small joy that I was able to continue to work with you and Mariko on the other side of your collaboration with Jim.

To my cohort and friends within the department, in particular Alberto, Andy, Ian, and Mitchell, I'm glad we could enjoy some of graduate school together. Thank you also to the smaller communities within Genome Sciences, the Noble lab slack, and the proteomics subgroup.

I offer heartfelt appreciation to my family for their support along the way. Jessica, thanks for always checking in and sending me pictures of Kepler. Jordan, thanks for being one of my best friends and helping me procrastinate. Thanks to my parents for supporting me in all my academic endeavors, from kindergarten through graduate school.

Cindy, we've been through the academic wringer together and come out the other side better for it, I'll never be able to adequately express my thanks for your love and support.

Chapter 1. INTRODUCTION

Proteins are the workhorse molecules of cells, carrying out innumerable functions to drive metabolic processes and sustain life. The function of proteins are their structure[1], allowing them to catalyze complex chemical processes by providing an amenable local environment. Beyond proteins' ability to function as individual reactors, their capabilities can be further extended by assembling into complexes, assemblies of proteins sharing a continuous interface. Complex formation enables new biological functions, as catalytic pockets can be formed between two distinct proteins or allosteric effects create new catalytic sites within a single protein. To this end, knowledge of the structures of proteins and protein complexes is a powerful lens through which to understand biological processes.

There are a variety of techniques for obtaining structural information about proteins or protein complexes, and they typically have a trade-off between throughput and structural resolution. For example, while cryo-electron microscopy or x-ray crystallography can provide nearly atomic resolution of dozens of proteins in a complex, these samples can require months to prepare and require specialized instrumentation. In contrast, there are extremely high throughput techniques, including yeast two-hybrid and affinity purification mass spectrometry (AP-MS), which can be done exhaustively on a cell line to identify over 100,000 putative interactions[2] but provide no structural constraints to help define how those protein complexes assemble. Cross-linking mass spectrometry provides a happy medium[3], where hundreds to thousands of interactions can be profiled in a day, while offering modest structural resolution[4] on the order of tens of angstroms.

Cross-linking provides distance constraints can be combined with high-resolution structures (where available) to serve as docking constraints to assemble complexes, discriminate between conformers, or potentially identify new conformers if the data are incompatible with existing structures. These low-resolution constraints can also be combined with protein structure prediction tools, such as Alphafold[5] or RoseTTAFold, as an orthogonal metric to identify potential structures or conformers[6] that could produce the observed cross-links. The structural data gathered from *in vivo* cross-linking experiments are particularly valuable, as they provide insight into protein structures and complexes as they exist in their native environment where they carry out their functions.

Chemical cross-linking coupled with mass spectrometry (XL-MS) has emerged as a powerful technique for studying protein conformations and discovering protein-protein interactions. Cross-linkers are a type of molecule capable of covalently bonding to two distinct residues on a protein or between two proteins, effectively providing a low-resolution snapshot of their relative positions at the time of cross-linking. The output of a cross-linking experiment is a list of pairs of peptides. If these peptides belong to the same protein, an intralink, then they provide information about the protein conformer, as those two residues must have been sufficiently close in some conformer for the cross-link to be physically possible to form. If the peptides belong to different proteins, then they provide not only the identity of two interacting proteins, but once again provide a distance constraint that can be used to evaluate putative models of the interaction. Those two proteins must come together in some way so that the two cross-linked residues are both solvent accessible and proximal enough to be cross-linked, providing a simple metric to filter models.

In recent years, cross-linking has become an increasingly popular and less specialized type of proteomics analysis. A variety of cross-linkers are available from commercial vendors, and hardware support and informatics have expanded to make cross-linking experiments readily attainable to many traditional proteomics labs.

Anatomy of a cross-linker

Cross-linker is an umbrella term covering a vast array of different molecules decorated with various chemical features to enhance their utility. At a basic level, most cross-linkers consist of two reactive groups connected by a spacer arm, a brief selection of different cross-linkers demonstrating some of the chemical diversity can be seen in Figure 1.1.

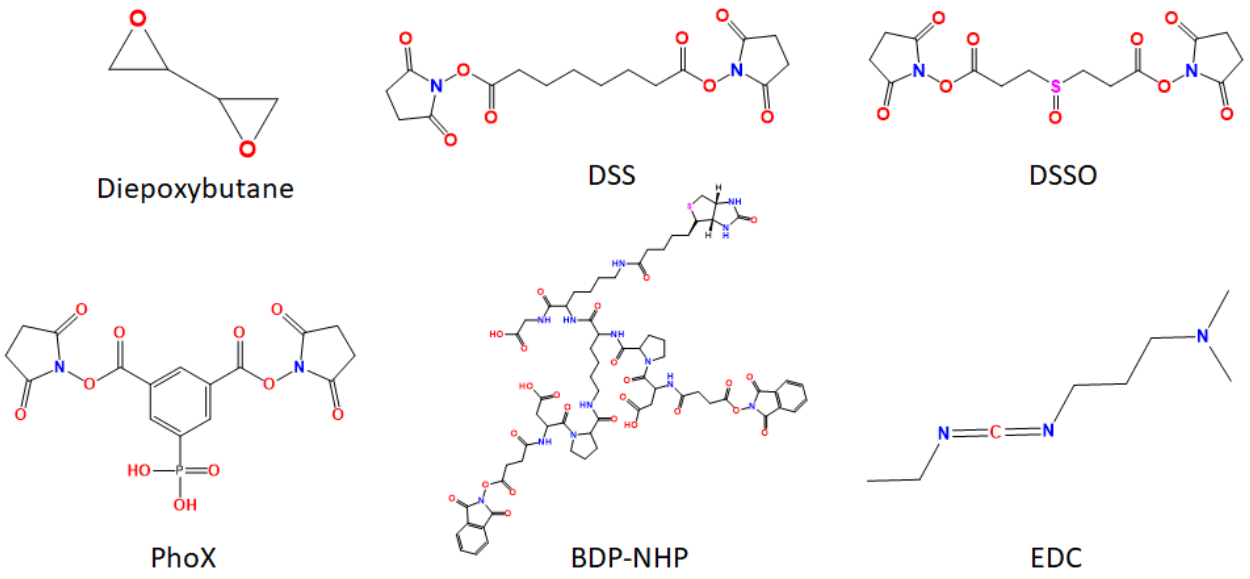


Figure 1.1 A brief selection of various cross-linkers that have been used to study protein structures and protein-protein interactions. Activated ester cross-linkers that react mostly with primary amines (DSS, DSSO, PhoX, BDP-NHP) are the most commonly employed type in recent years, but indiscriminate residue targeting (Diepoxybutane) and heterobifunctional reagents (EDC) have also been used to generate valuable structural data.

The type of reactive sites included in a cross-linker determine what residues it can react with to form a covalent bond. The most common type of reactive group is an amine-reactive ester, such as N-hydroxysuccinimide (NHS)[7–9] or N-hydroxphthalimide (NHP)[10], which readily reacts with primary amines, such as those found in lysine or at the N-termini of proteins. Lysines are an attractive target because they are strong nucleophiles and are sufficiently common on the solvent exposed parts of proteins. However, for other applications, or simply for obtaining a greater diversity of distance constraints[11], targeting alternative residues may be desired. Cross-linkers have been designed to link carboxylic acid groups[12], thiols[13], or react indiscriminately[14] for regions where coverage is not readily attainable with traditional amine-targeting reagents.

The spacer arm can be thought of as the part of the cross-linker in between the reactive groups, and it determines the structural resolution of the cross-linker as it defines the maximum distance between two cross-linked sites. The shorter the spacer arm is, the higher the resolution of any discovered interactions, with the shortest possible distance being a "zero-length" linker[15,16] that contributes no additional atoms to the covalent bond and directly bonds two residues together. However, due to the dynamic nature of proteins, even when using short cross-linkers, a wide distribution of distances is obtained[15] when compared to published high-resolution structures. The cost of using a shorter cross-linker is that there are necessarily fewer possible cross-links in any given system, as some residues will be too far apart to be linked without using a longer cross-linker. Additionally, simulations suggest that in complex samples, the ability to discover inter-links, cross-links between two distinct protein molecules, is correlated to the spacer arm length[17], with longer spacer arms being more capable of producing inter-protein cross-links. Intuitively,

lysine sites within a protein tend to be closer than lysine sites on two different proteins, even when those proteins are part of the same complex[17].

While the spacer arm defines the structural characteristics of the cross-linker, it can be further imbued with additional chemical features to simplify its application. The most common desirable feature in the spacer arm of a cross-linker is the inclusion of mass spectrometry cleavable bonds[10], which are chemical bonds that break at a low activation energy as compared to the peptide backbone. Cross-linkers like the peptide-based Protein Interaction Reporter (PIR) family of cross-linkers[10,18,19], DSSO[20], or DSBU[8] have all included CID-cleavable bonds that fragment to yield some population of intact released peptides, which have the effect of greatly reducing the search space during the assignment of the two cross-linked peptides. Another desirable feature is the inclusion of some type of enrichment tag. Cross-linking reactions tend to be low efficiency, with only a few percent of residues being labeled, and most of the sample material being linear peptides. Consequently, being able to efficiently separate cross-linked peptide from the bulk of the linear peptides greatly enhances the sensitivity of the experiment as less instrument time is spent sampling and accumulating linear peptides. The PIR family of molecules have included biotin that can be pulled down with monomeric avidin[21] to perform enrichment on any molecule bearing the cross-linker. More recently, a cross-linker called PhoX[22] bearing an IMAC-enrichable phosphonic acid group has been developed, allowing for efficient peptide-level enrichment using the well-established IMAC platform developed for phosphoproteomics. The use of the comparatively stable phosphonic acid over phosphoric acid enables an experiment where all the phosphoric acid bearing peptides can be dephosphorylated, leaving PhoX labeled molecules as the primary IMAC

enrichable species. Chemical modifications to the spacer arm allow for deeper interrogations of complex samples by reducing the search space or allowing for enrichment of cross-linked species.

Identification of cross-linked peptide pairs

Many software tools have been developed to facilitate cross-linked peptide identification[23], largely separated by their ability to identify cross-links formed by cleavable cross-linkers or by stable backbone cross-linkers. Stable backbone cross-linkers were the original cross-linkers employed[24], defined by a lack of cleavable bonds in the spacer arm. Consequently, the fragmentation of these peptides produces complex fragmentation spectra in which the precursor masses of either constitutive peptide is unknown. This produces an extremely large search space which is quadratic with the number of peptides in a database, commonly referred to as the n^2 problem, as the primary constraint on the search is that there is some pair of peptide masses in the database that adds up to the observed precursor mass. Due to the computational complexity of these searches, these analyses have typically been carried out in cases where only a small number of proteins need to be considered. Despite this complexity, there are strategies to reduce the number of pairs to computationally feasible levels, with open-searches emerging as a potential solution[25,26]. Open-window searches use wide precursor mass tolerances, typically on the order of hundreds of Daltons, to retrieve candidate peptides to score against a spectrum. This has been shown to be an effective way to identify peptides bearing various post-translational modifications[26] without having to know which mass shifts to consider prior to the search. Functionally, the task of identifying cross-links can be formulated as identifying two peptides, each bearing an unknown modification mass of the other peptide and the cross-linker, meaning they can be identified in open-searches as long as a sufficiently large precursor tolerance is used. There are a variety of

software packages developed for identifying cross-links from stable backbone cross-linkers including StavroX[27], Kojak[28], Search-for-xlinks[29], pLink[30], and many other tools covered in a recent review[23]. For further refinement of results, the semi-supervised machine learning postprocessor, Percolator[31] can be used increase the sensitivity of cross-linking experiments performed using a stable backbone cross-linker.

In recent years, cross-linking experiments on the whole proteome scale have tended to make use of cleavable cross-linkers, which yield a comparatively simple search problem. The use of a cleavable cross-linker enables direct calculation of the two released peptide masses by use of a mass relationship, which can be formulated in a variety of ways for different cross-linkers, but is ultimately the conservation of mass in some form:

$$MS1.Precursor = Xlink.reporter + MS2.mass1 + MS2.mass2 \quad \textbf{(Eq.1.1)}$$

The precursor masses of two constitutive peptides can be determined in a single pass of a spectrum's peak list[32]. This process is fast enough even to perform in real time during an LC-MS experiment[33,34] and has the effect of reducing the n^2 problem to a pair of linear searches, although many pairs of precursors may need to be searched for a single spectrum if many pairs of peaks fulfill equation 1.1. A variety of software tools have been developed to identify cross-linked peptides generated with a cleavable cross-linker. These tools include XlinkX[34,35], MeroX[36], Mango[32], MS-Annika[37], MaXlinker[38], MaxLynx[39], all of which make use of a mass relationship to reduce the search space. Comparisons between different search tools reveal differential performance depending on the cross-linker used[40], without a generally optimal tool emerging.

False discovery rate (FDR) estimation is usually specific to the software package used to do the search, although most approaches rely on some implementation of target-decoy competition[41], potentially with some cross-link specific modifications[42] to handle cross-links between targets and decoys. The Mechtler lab generated a valuable community resource, a ground truth dataset of synthetic cross-linked peptides, and determined that commonly used search tools' have poor FDR control[40], yielding too many false positives at a given cutoff. Proper FDR-control for cross-linked peptide should rely on the minimum evidence score for each peptide pair[43], as imbalanced fragmentation can lead to one peptide having significantly better fragment coverage than its partner. XLinkProphet[44], developed as an additional rescoring step for PeptideProphet[45] results, uses a minimum expect score as a feature and maintains proper FDR control in standard samples as demonstrated by an entrapment strategy. Currently, there is no unified approach or implementation to FDR estimation across tools, but the Rappsilber lab has produced the most rigorous statistical framework for cross-linked peptide FDR control[46].

Applications of chemical cross-linking

Cross-linking experiments have demonstrated significant utility in a variety of applications. In recent years, cross-linking experiments have evolved rapidly from providing qualitative insights into the interactome to identifying novel host-pathogen interactions[47–49], protein drug complexes[50], and characterizing the conformational plasticity of large complexes[51]. Mirroring the progression of traditional proteomics experiments from identifying peptides to quantifying peptides, cross-linking has seen great advancements in quantitative capabilities. With the recent development of the isotopically encoded iqPIR reagents[19,42] or by adapting existing

technologies such as TMT labels[52,53] or SILAC[54], quantitative cross-linking studies have enabled differential systems structural biology measurements.

Quantitative cross-linking can be used to provide information about the relative abundance of different conformers of a protein and simultaneously provide a measurable proxy for the abundance of protein-protein interactions. This is an important added dimension to quantitative proteomics studies as even without protein abundance changing, conformational rearrangements and shifts in the interactome frequently have significant biological consequences. Quantitative cross-linking has provided insight into the conformational shifts caused by various HSP90 inhibitors and provided mechanistic insight into the underpinnings of heart failure in mouse models[55].

With large scale quantitative measurements becoming more common, the ability to model differentially expressed protein complexes or conformers becomes more important. Very recent developments in protein structural prediction techniques, namely Alphafold2[5,56] and RoseTTAfold[57], provide new opportunities for applications of cross-linking data. Alphafold2 has been shown to be a powerful tool for producing models of protein-protein complexes, which is a vital complement to cross-linking experiments that can produce many low-resolution distance constraints to filter new models. The traditional method of producing models of protein-protein interactions from cross-linking data would involve producing a structure of two monomers and then performing rigid body docking using the cross-links as distance constraints. Instead, models can now be produced by co-folding proteins together which has the added benefit of allowing for conformational rearrangement not possible in rigid body docking[57], and then these models can be evaluated based off their agreement with observed cross-links. An initial study looking at large

scale modeling of protein complexes with Alphafold shows that agreement with cross-linking distance constraints correlates with higher model confidence from predicted dimers[58], suggesting that cross-links can help identify biologically relevant models.

Organization of dissertation

The remaining chapters of this dissertation describe three projects that enhance our ability to identify cross-links and obtain structural information about proteins in mass spectrometry-based discovery experiments. The first project describes Mango, a software package that can be used in tandem with Comet[59] and XlinkProphet[44] to efficiently identify cross-links from chimeric MS2 spectra generated from a variety of cross-linkers. The second project describes the development of the first tetrameric cross-linker and a real-time instrument method to enable identification of tetra-linked peptides, enabling the unambiguous identification of up to four proteins at an interface. The final project describes instrument control and signal processing advancements for a developmental FT-ICR mass spectrometer containing multiple mass analyzers to improve the duty cycle in high-resolution experiments through parallel acquisition of spectra. These three projects together all offer improved ways to produce or identify cross-linked peptides.

Chapter 2. MANGO: A GENERAL TOOL FOR CID-CLEAVABLE CROSS-LINKED PEPTIDE IDENTIFICATION

The contents of this chapter are adapted from the following work: Mohr JP, Perumalla P, Chavez JD, Eng JK, Bruce JE. Mango: A general tool for CID_cleavable cross-linked peptide identification. Analytical Chemistry. 2018 90(10) 6028-6034. Copywrite 2022 American Chemical Society.

2.1 ABSTRACT

Chemical cross-linking combined with mass spectrometry provides a method to study protein structures and interactions. The introduction of cleavable bonds in a cross-linker provides an avenue to decouple released peptide masses from their precursor species, greatly simplifying the downstream search, allowing for whole proteome investigations to be performed. Typically, these experiments have been challenging to carry out, often utilizing non-standard methods to fully identify cross-linked peptides. Mango is an open-source software tool that extracts precursor masses from chimeric spectra generated using cleavable cross-linkers, greatly simplifying the downstream search. As it is designed to work with chimeric spectra, Mango can be used on traditional high-resolution MS/MS capable mass spectrometers without the need for additional modifications. When paired with a traditional proteomics search engine, Mango can be used to identify several thousand cross-linked peptide pairs searching against the entire *E.coli* proteome. Mango provides an avenue to perform whole proteome cross-linking experiments without specialized instrumentation or access to non-standard methods.

2.2 INTRODUCTION

Proteins are specialized molecules designed to carry out innumerable functions in a cell. These functions rely not only on the abundance of individual proteins, but are critically dependent upon localization, conformation and interactions between assemblies of proteins. Chemical cross-linking combined with mass spectrometry (XL-MS) is emerging to hold great potential for interrogating conformations and interactions that exist within cells[60–63], and elucidating how these networks of interactions can change under different conditions[64,65] XL-MS experiments have been growing in number and appeal over recent years due to the extensive potential applications of cross-linking data[66]. Cross-links within and between proteins provide physical distance constraints that can be used to predict structures. These constraints can be used alone in *ab initio* folding for improved structures[67], or to filter and refine models from homology templates[68,69]. They can also be combined with complementary structural determination tools, such as cryo-EM[70] or x-ray crystallography[71] to potentially produce higher quality structures by restricting the possible solution space. *In vivo* cross-linking experiments uniquely allow for large-scale quantification and monitoring of protein conformational dynamics[65], which can be used to visualize physical effects on conformations and interactions of drug target proteins, as well as off-target effects in cells. Another unique aspect of cross-linking experiments is the ability to identify direct host-pathogen interactions on a structural level[47,48], providing a lens through which to visualize molecular details of pathogenesis of a variety of bacteria and viruses.

Cross-linking experiments in the past have typically utilized stable-backbone cross-linkers, such as BS3 or DSS, which produce a chimeric spectrum of two peptides with known combined

mass, but unknown individual peptide masses. With only the precursor mass of the pair to limit the search space, a significant number of peptides or combinations of peptides must be considered and scored with proteome-wide searches. The number of these combinations that will need to be evaluated is quadratic with respect to database size, which makes use of stable-backbone cross-linkers difficult in whole proteome analyses. Despite the complexity of data generated by these cross-linkers, many tools have been developed for evaluating such datasets. These tools implement strategies to reduce the number of peptide pairs that need to be considered for any given spectrum[28–30], which partially mitigate the effect of a large database. A promising development has been the implementation of scoring optimizations that facilitate open window searches to find peptides with unknown modifications[26], which is a prominent issue in searching cross-linking data. Regardless of the methodology employed to search these types of cross-linking experiments, statistical power is lost when each spectrum has a large number of candidates.

These problems of quadratic candidate expansion and minimal statistical power are not inherent to all cross-linking experiments, but come about with the use of non-cleavable cross-linkers. Cleavable cross-linkers have provided an alternative to stable backbone cross-linkers by reducing the quadratic search problem to a linear one. These cleavable cross-linkers include molecules such as Protein Interaction Reporter (PIR)[10], DSSO[7], DSBU[8] and others that incorporate bonds that can be cleaved predictably in the gas phase. These molecules are engineered such that cross-linker bond cleavage yields two released peptides, which permits the decoupling of the individual cross-linked peptides from their precursor. Once released peptide masses are determined the search can be formulated as two traditional narrow window linear peptide searches instead of a single quadratic search. Normal peptide searches are linear with database size, so

cleavable cross-linkers perform well in complex biological samples which necessitate a large database. Some commercially available cleavable cross-linkers, such as DSSO[7] and DSBU[8], utilize a characteristic doublet pattern to identify released peptide masses to simplify the search problem and enable the use of a full proteome database. When combined with a tool like XlinkX and an instrument capable of serial fragmentation[34], DSSO can be used to study complex cross-linked lysates. Some newer instruments now have native support[35] for XlinkX, but high throughput cross-linking experiments remain difficult without access to hardware capable of serial fragmentation.

Hardware requirements and limited instrument control have been a long-standing barrier for optimal use of cleavable cross-linkers on a proteome wide scale. Real-time Analysis of Cross-linked peptide Technology[33] (ReACT) is an XL-MS technique developed for dynamic discovery of protein interaction reporter (PIR) cross-linked peptides[18]. ReACT uses high resolution hardware[72] and software modifications on an LTQ-FT to efficiently identify cross-linked peptides. This method incorporates multiple fragmentation and isolation events to produce spectra of the individual cross-linked peptides at the MS3 level. While this method produces individual peptide spectra compatible with traditional search tools and has produced the majority of *in vivo* cross-linked peptides now residing in the database of cross-linked peptides XLinkDB[73,74], it comes at a significant time cost inherent to incorporating multiple scan events. Recently, ReACT has been used to construct libraries that allow for either spectral library searches[75] or Parallel Reaction Monitoring-based quantification of previously identified cross-linked peptides, extending some benefits of the methodology to instruments only capable of MS2. A limitation of spectral libraries

is that no new cross-linked peptide pairs can be identified during an experiment, which restricts the scope of experiments reliant on these libraries. Extension of PIR identification capabilities to MS2 measurements in the absence of required spectral libraries can significantly extend PIR experiments to many other labs.

To help address the need for improved cross-linked peptide identification capabilities, here we present Mango, an open-source search tool for use with CID-cleavable cross-linkers for the identification of novel cross-links. Mango employs logic similar to ReACT to identify cross-linked peptides but requires only the capability to produce high resolution MS2 spectra, making the methodology adaptable for use on many commercially available and commonly used mass spectrometers. Unlike other cross-linking search tools, Mango is capable of efficiently analyzing data from cross-linkers with symmetric fragmentation which lack characteristic doublets, even when a full proteome database is used, and it has a standard output format that is compatible with various traditional peptide search tools.

2.3 METHODS

E.coli cell culture, cross-linking, and digestion conditions

E. coli (K12) was grown to stationary phase in LB media. *E.coli* cells were pelleted at 1500g for 10 minutes and washed with phosphate buffered saline (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4) followed by cross-linking buffer (180 mM Sodium phosphate pH 8.0). The cell pellet was gently resuspended in 500 µL cross-linking buffer, and biotin aspartate

proline-N-hydroxyphtalimide (BDP-NHP), synthesized by solid phase synthesis[33], was added to a final concentration of 10 mM. After one hour at room-temperature any remaining reactive cross-linker was quenched with the addition of 1mL of 100 mM ammonium bicarbonate. After quenching, the cells were again pelleted, the cross-linking buffer was removed, and the cells were resuspended in 100 mM ammonium bicarbonate. Urea was added to 8M and then the cells were lysed by sonication using a GE-130 ultrasonic processor. The lysed samples were reduced with 5 mM Tris(2-carboxyethyl)phosphine for 30 minutes at room temperature, and then alkylated in 10 mM Iodoacetamide for 45 minutes in the dark. The samples were diluted 10-fold with 100 mM ammonium bicarbonate to reduce urea concentration to 0.8M, and the proteins were then digested overnight at 37°C using trypsin (Promega). Digested samples were desalted using C18 sep-pak columns (Waters).

Strong cation exchange fractionation and affinity enrichment

The desalted peptide samples were fractionated by strong cation exchange (SCX) chromatography using an Agilent 1200 HPLC system equipped with a Phenomenex Luna SCX column. A binary linear gradient consisting of buffer A (5 mM KH_2PO_4, pH 2.6, 30% acetonitrile (ACN)) and buffer B (5mM KH_2PO_4, pH 2.6, 30% ACN, 350 mM KCl) was applied at a flow rate of 1.5 mL/min for 97.5 min as follows: 0% B at 0 min, 5% B at 7.5 min, 60% B at 47.5 min, 100% B at 67.5 min, 100% B at 77.5 min, 0% B at 77.51 min to completion. Fractions were taken every 5 min starting at 17.5 min, and fractions were pooled as follows: 1-5, 6-7, 8, 9, 10, 11-14. Fractions 1-5 were not processed any further. The remaining fractions were then reduced to a final volume of 1-2 mL by vacuum centrifugation and pH adjusted to a pH of 8.0 with 1.5M NaOH. After pH adjustment, each sample was incubated for 1 hour with 100 μL of UltraLink monomeric avidin

(Thermofisher) with gentle agitation. The avidin matrix was washed 5 times after this incubation period using 3mL aliquots of 100 mM ammonium bicarbonate, and cross-linked peptides were then eluted off the avidin beads by two additions of 500 µL 70% acetonitrile-0.5% formic acid. The resulting eluent was concentrated by vacuum centrifugation.

LC-MS/MS data acquisition description

Peptides recovered from the avidin matrix were analyzed using an EASY-nLC 1000 coupled to a Q Exactive Plus mass spectrometer. Samples were fractionated over a 60 cm x 75 µm inner diameter fused silica analytical column packed with ReproSil-Pur C8 (5 µm diameter, 120 Å pore size particles) by applying a linear gradient from 90% solvent A (0.1% formic acid in water), 10% solvent B (0.1% formic acid in acetonitrile) to 60% solvent A, 40% solvent B over 240 minutes at a flow rate of 300 nL/min.

The mass spectrometer was operated using a data dependent analysis (DDA) method performing one high resolution (70,000 resolving power (RP) at *m/z* 200) MS1 scan from 400-2000 *m/z* followed by MS2 (17,500 RP) on the 20 most abundant ions with a charge between +4 and +8 inclusive detected in the MS1. Parameters for MS2 scans included an automatic gain control target of 50,000 ions, a maximum ion accumulation time of 100 ms, an isolation window of 3.0 m/z, and a normalized collision energy of 30. A dynamic exclusion window of 30 seconds was used to reduce redundant picking of the same parent ion. MS2 spectra were processed using Mango 2017.01 rev. 0 beta 2, whose output was searched using Comet[59] 2017.01 rev. 2.

The same samples were subsequently analyzed using a Water NanoAcquity UPLC coupled to a Thermo Velos Fourier transform ion cyclotron resonance mass spectrometer[72] (Velos-FT). The chromatography gradient employed is identical to the one previously described. The Velos-FT was operated using the ReACT method[33], where one high resolution (50,000 RP at *m/z* 400) MS1 scan from 400-2000 *m/z* is taken and followed by an MS2 (50,000 RP) on the most abundant ion of at least +4 charge. In real time, if a pair of peaks fulfills the mass relationship (precursor mass = reporter ion mass + peak 1 mass + peak 2 mass) within 20ppm tolerance, then each peak is targeted for 2 additional low resolution MS3 scans in the ion trap. MS3 spectra were searched using Comet 2017.01 rev. 2.

Software description

Mango is an open source tool written in C++ and hosted on GitHub (https://github.com/jpm369/mango) developed for facilitating the identification of cross-linked peptides at the MS2 level. Broadly, Mango is a tool for extracting released peptide masses from MS2 scans of cross-linked peptides. The software outputs a searchable file with multiple precursors for each scan that correspond to candidate released peptide masses.

Mango takes an mzXML[76] file and a Mango parameters file as its inputs. Mango utilizes Hardklör[77] to preprocess experimental spectra, performing charge deconvolution and de-isotoping to reduce spectral complexity. These reduced spectra are then used to identify pairs of peaks that may have been generated by a cross-linked species. Mango loops through all of the peaks in a deconvoluted MS2 spectrum to identify pairs of peaks that fulfill the mass relationship, described

in the results section, for a cleavable cross-linker within some user-specified tolerance (Eq. 2.1). The added requirement of the mass relationship reduces the complexity of the search from quadratic to linear with respect to the number of peptides in a database, analogous to a traditional narrow-window DDA search. This reduction in complexity allows for cross-linking searches to be carried out using an input database containing thousands of proteins without a quadratic increase in search time or loss of statistical power. If a scan contains at least one pair of peaks that fulfills the mass relationship, then each pair of peaks in the spectra that fulfill the mass relationship are written to an ms2 file as a potential precursor mass in place of the MS1 precursor mass isolated. Herein Mango was operated using the following settings: mass_tolerance_relationship = 10.00 ppm, mass_tolerance_peptide = 20.00, reporter_neutral_mass = 751.406080 Da.

Mango outputs files in the .ms2 format[78] that can be directly searched by Comet (Version 2017.01). Comet searches were performed with the default settings and the following changes: mango_search = 1; variable modifications: 15.9949 M, required modifications: 197.032422 at an internal K. Comet loops through the list of released peptide masses in the precursor header of the file and uses each mass identified by Mango as a precursor mass to perform a narrow window search for that spectrum. Comet scores all fragments that could be generated by each linear precursor mass but does not score fragments containing the second peptide (Figure A.5.1). This results in computing a cross-correlation score, E-value, and all other standard Comet metrics for each precursor mass queried, facilitating established downstream analysis. A custom parameter option in Comet 2017.01, mango_search, directs Comet to also provide a unique identifier for each pair of released peptide masses in a scan to facilitate reassembling the individual linear

identifications into a cross-linked identification. This identifier is appended to the spectrum title and contains a pair index and the letter A or B to indicate which identifications should be paired (e.g. 001_A and 001_B). Corresponding linear peptide identifications are paired together according to their unique identifier, and then the results are filtered to a 1% FDR using a target-decoy based filter at the peptide-spectrum match (PSM) level. A PSM for a pair of cross-linked peptides refers to the pair of peptide assignments to a single spectrum. FDR filtering was performed by first assigning a cross-linked pair of peptides an E-value equal to the worse[43] of the two E-values assigned by Comet. Each spectrum is then assigned its best scoring pair, and then selecting an E-value cut-off to limit final results to contain no more than 1% of the pairs that contained one or two decoy hits. This strategy was evaluated by searching cross-linked and non-cross-linked samples using a variety of reporter masses in Mango (Figure A.5.2).

2.4 Results and Discussion

Identifying released peptides masses using Mango.

Cleavable cross-linkers provide an avenue by which the individual masses of a pair of cross-link peptides can be determined. In recent years, a number of search tools have emerged that take advantage of the predictable fragmentation products of cleavable cross-linkers to reduce the number of pairs of peptides that must be scored to identify a cross-linked species[33,34,36]. However, some of these methods require instrumentation capable of MS3, or equipped with other non-standard capabilities, to generate suitable data. Mango provides a tool to facilitate whole-proteome,

in vivo cross-linking experiments using a standard Orbitrap-based or other high resolution mass spectrometer.

Mango makes use of the same mass relationship, originally implemented in ReACT[10], to target candidate cross-linked peptide masses on the fly. Instead of utilizing this information in real-time which requires specialized methodology, Mango utilizes this relationship only in post-processing of spectra, allowing a number of CID/HCD-enabled mass spectrometers to be used in conjunction with Mango. This mass relationship is simply a formulation of the conservation of mass, namely that a crosslinked species will fragment to produce two species whose neutral masses sum together with the reporter mass to match selected precursor ion neutral mass with tight tolerance. For a bi-functional cross-linker, this equation can be generalized to:

$$MS1.Precursor = Xlink.reporter + MS2.mass1 + MS2.mass2 \quad \text{(Eq.2.1)}$$

where the reporter is a constant that corresponds to the cross-linker specific mass offset determined by its specific structure and fragmentation products. For example, in BDP-NHP, a biotinylated peptidic cross-linker, this reporter mass stems from the biotinylated region of the molecule with a mass of 751.4106 (Figure A.5.3). In DSSO this reporter mass does not correspond to a physical species, but rather stems from the mass defect resulting from pairs of the light or heavy stump modifications compared to the mass modification associated with the intact cross-linked pair[34] (Figure A.5.4).

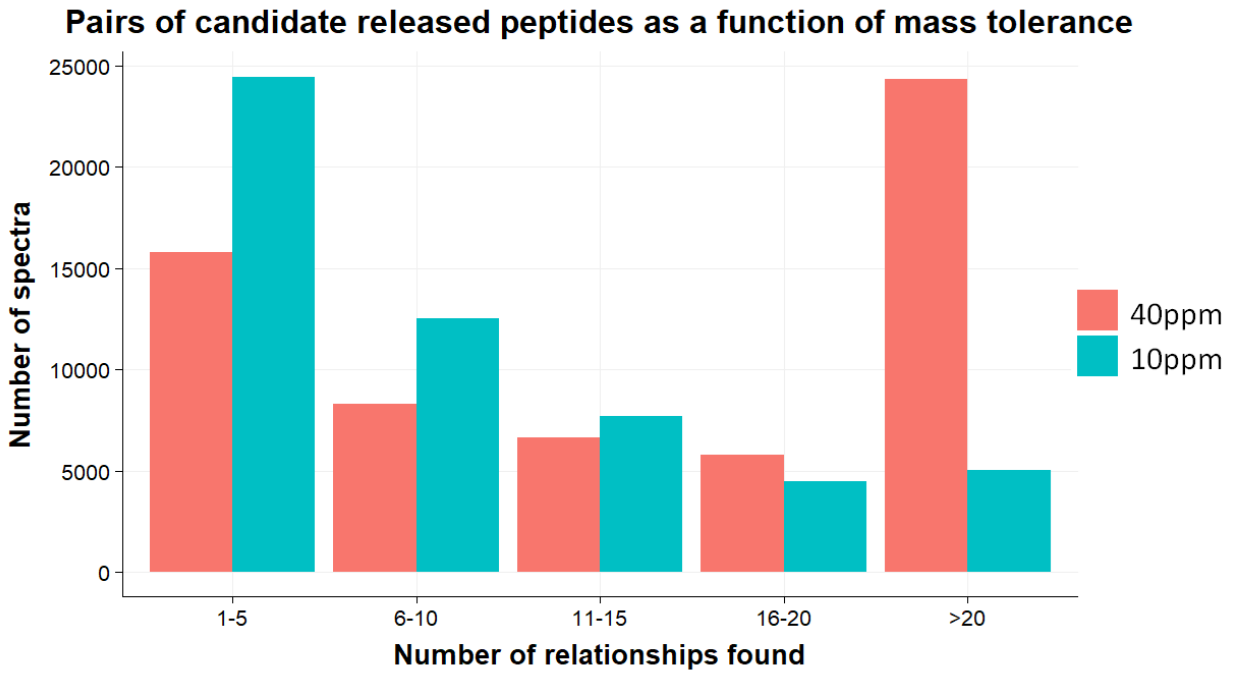


Figure 2.1. The effect of tolerance on the number of putative released peptide pairs found using Mango across an entire dataset. High mass accuracy improves the results obtained with Mango, as it allows for a reduction in the tolerance of the mass relationship utilized for identifying candidate released peptides. A tighter mass tolerance reduces the number of candidates that must be scored for any given spectrum in a downstream search, which improves both search time and statistical power.

Since the reporter mass of a cross-linker can be determined from its structure or measured experimentally and provided as an input to equation 1, the number of candidate released peptide masses is quadratic with respect to the number of peaks in a spectrum. Mango loops through all of the peaks in a tandem mass spectrum, generating all combinations of peaks and evaluates their sum, reporting any that fulfill equation 1 within a user-specified mass tolerance. Empirically, approximately 5 times the expected mass accuracy of the instrument appears to be the optimal

tolerance to limit the number of candidates considered without losing many true or forward hits (Figure A.5.5). Increasing the mass tolerance to higher values results in a large number of spectra having many candidate pairs (Figure 2.1). The increased number of candidates to score increases downstream search time and decreases statistical power, which can lead to an overall lower number of identifications after FDR filtering (Figure A.5).

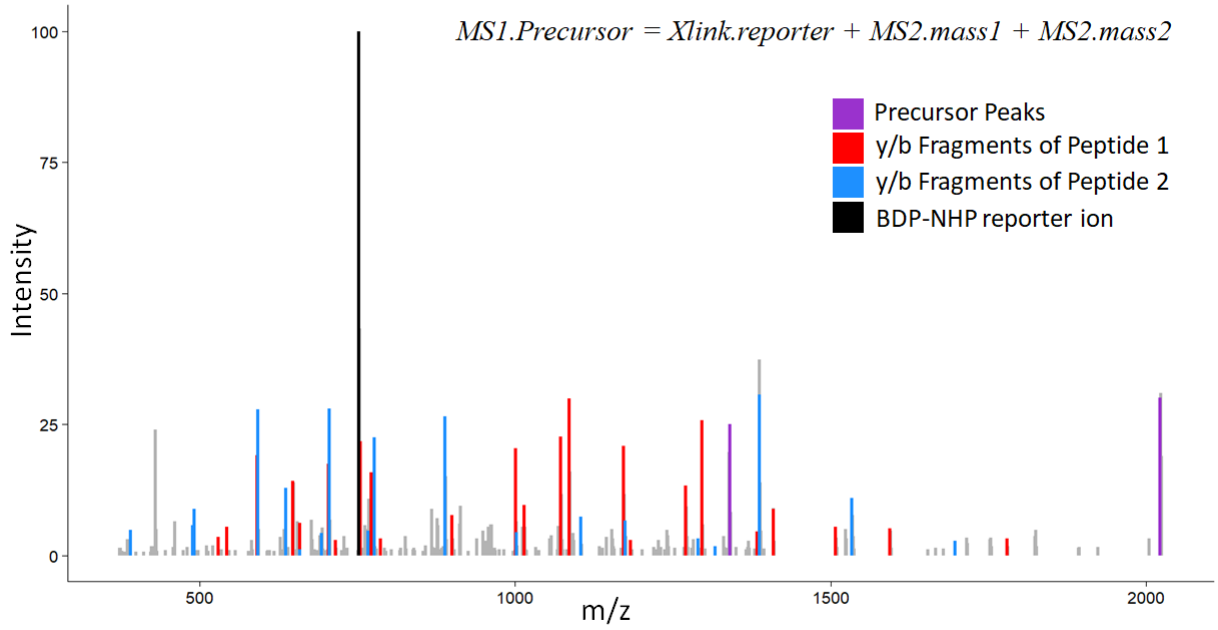


Figure 2.2. Mango identifies pairs of peaks (purple) in an MS2 spectrum that fulfill a mass relationship within some tolerance. These masses can then be used by a peptide search engine, such as Comet, to perform multiple narrow window searches on the same spectrum, twice for each pair of masses output by Mango. These sequential narrow window searches yield a peptide identification for each mass in the pair based off observed fragments for each peptide (red and blue). These identifications can then be paired using a unique identifier assigned by Mango to identify the initial cross-linked species isolated.

A novel feature of Mango among cross-linking analysis tools is that it does not contain a native peptide search engine, and instead outputs an .ms2 file encoding all pairs of masses that

fulfill equation 1 as a precursor mass for the scan from which they were extracted from. A .mgf output file is also available that lists each precursor mass query as a separate scan for search tools that will not handle multiple precursor masses per spectrum, allowing alternative search engines, such as MSFragger[26], to be used (Figure A.5.6). This will facilitate the integration of Mango into existing pipelines without the need for significant change to existing post-processing and analysis methods. Mango can be used as a pre-processing step before peptide scoring, and the appropriate peptides can be paired together at the end of all post-processing steps to identify their progenitor cross-linked species.

Whole proteome cross-linking at the MS2 level

Mango and Comet were used to identify cross-links generated by cross-linking *E. coli in vivo* with BDP-NHP. BDP-NHP is a biotinylated-peptidic cross-linker capable of penetrating cell membranes[79], allowing it to preserve protein-protein interactions in their native context during *in vivo* cross-linking experiments. For this strategy, it is necessary to produce fragment ions for both the intact released peptides, as well backbone fragment ions corresponding to the amino acid sequences of the released peptides (Figure 2.2). The intact fragments are necessary for Mango to be able to identify candidate masses that can effectively constrain the search, while peptide backbone fragment ions are necessary to assign primary sequences for the peptides.

Comet was used to search the output of Mango from an *in vivo* cross-linking experiment. After Mango has extracted candidate released peptide masses, the subsequent database search is identical to a normal peptide search, which is linear with respect to database size. Consequently, this

pipeline can be used to search for cross-links from the whole *E.coli* proteome (4309 target and 4309 decoy proteins).

Across single injections of five SCX fractions, Mango and Comet were able to assign 4170 PSMs, mapping to 2334 non-redundant peptide pairs at less than or equal to 1% FDR (Figure A.5.7). These identifications are not evenly distributed across all fractions, but rather concentrated in the later fractions (8-14) which are enriched for highly charged species by SCX (Figure 2.3). Beyond being able to identify thousands of cross-linked peptide pairs in a single sample, these can also be found with reproducibility similar to traditional DDA runs with duplicate analyses yielding roughly 75% overlap due to the stochastic nature of DDA sampling (Figure 2.3).

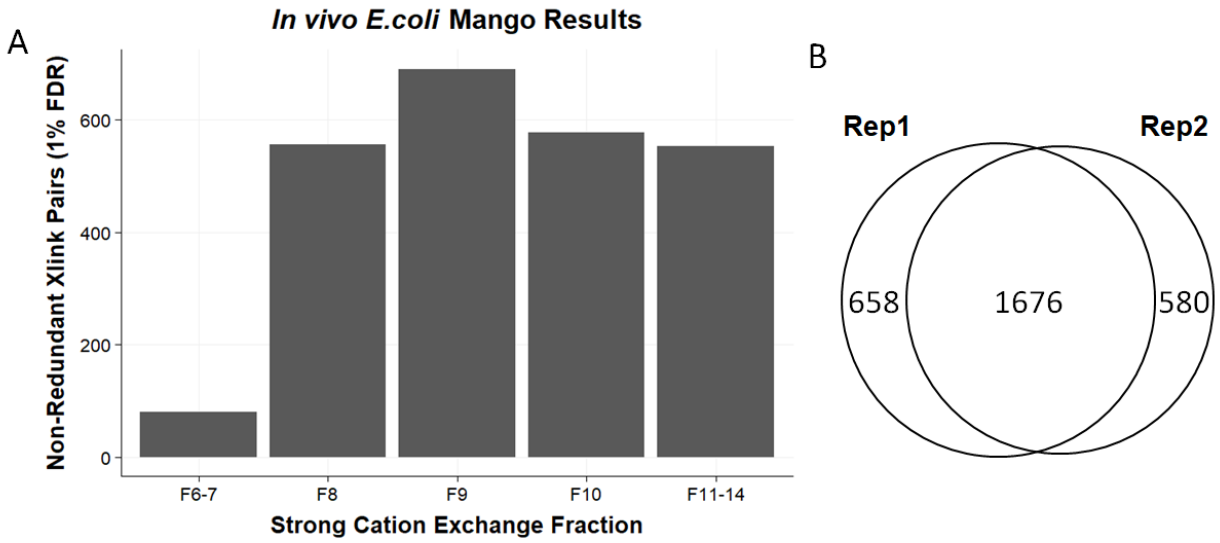


Figure 2.3. Summary of Mango analysis of a cross-linked E.coli sample. (A) Number of non-redundant IDs found in each SCX fraction analyzed after FDR filtering. (B) Overlap of non-

redundant peptide pairs identified post-FDR filtering between technical replicates of the same 5 SCX fractions.

While Mango as presented is applied to a PIR cross-linked sample, it is agnostic to the cross-linker and can work with any cross-linker for which a mass relationship (Equation 1) can be formulated. To test this, Mango was used to identify cross-links from published DSSO cross-linked HeLa lysate data[34] (Figure A.5.8).

Comparison with ReACT

Mango's performance was compared directly to ReACT, a dynamic MS3-based method for identifying cross-links. In contrast to Mango, ReACT is able to dynamically target candidate released peptide masses for MS3[33] increasing the certainty in fragment assignment by avoiding chimeric spectra. However, ReACT requires 2 high resolution scans and 4 low resolution scans totaling to approximately 3 seconds to fully query a cross-linked peptide pair, which limits the depth of coverage and reproducibility in very complex samples. While MS2 chimeric spectra containing fragments from both peptide sequences are more difficult to score than independent MS3 spectra of peptides generated by ReACT, each cross-linked pair can be investigated more quickly by eliminating the MS3 requirement entirely and acquiring lower resolution MS2 spectra. The combination of both these measures produces nearly an order of magnitude more investigations per analytical run, at the cost of increased ambiguity in fragment assignment associated with chimeric spectra.

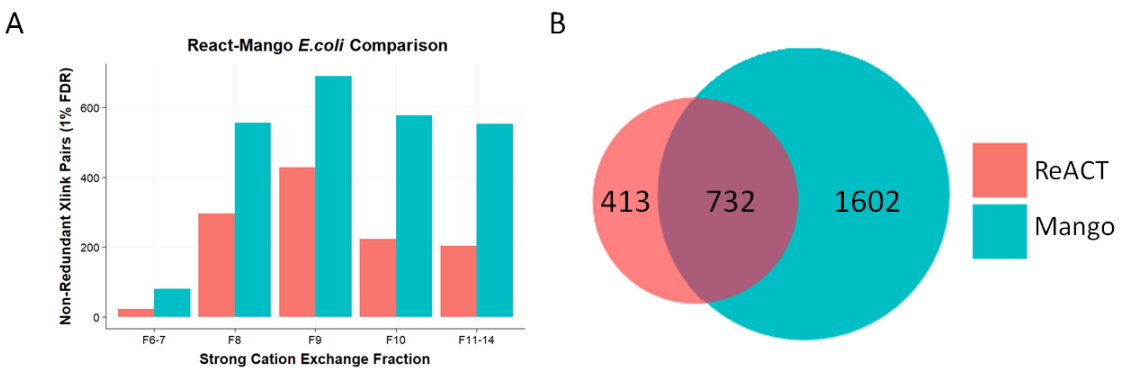


Figure 2.4 Summary of comparison between Mango and ReACT for a whole proteome in-vivo cross-linking experiment. (A) The per fraction comparison between Mango and ReACT. (B) The overall overlap of non-redundant peptides post-FDR filtering identified using Mango and ReACT.

The same fractions of the *E. coli* sample analyzed in the previous section were analyzed using ReACT on the Velos-FT and compared to the results achieved by Mango. This data set consisted of 273886 HCD-MS2 spectra from the Q-Exactive Plus for Mango, and 30916 CID-MS3 spectra from the Velos-FT generated by ReACT. In each fraction, Mango was able to identify at least 50% more non-redundant peptide pairs as compared to ReACT (Figure 2.4). Overall, Mango identifies 2334 non-redundant peptide pairs compared to ReACT's 1135, which are mapped to 1974 and 1002 unique lysine-lysine site interactions respectively. However, a greater fraction of paired ReACT spectra are successfully assigned a high-confidence PSM, indicating that the identification improvements presented by Mango can be largely attributed to having a larger number of queries compared to ReACT.

Comparing the overlap of non-redundant PSMs from the two datasets reveals that each method identifies exclusive sets of cross-linked peptide pairs (Figure 2.4), despite the overall increase in results achieved with Mango. The lysine-lysine site level shows a similar trend as the peptide level, with the two methods sharing 675 unique interactions. While some portion of these missed identifications can be explained by the stochastic sampling employed, differences in the fragmentation strategies between the two methods may further explain some missed identifications. ReACT utilizes multiple isolation and fragmentation events to produce fragments from one released peptide at a time with a traceable relationship to the cross-linked peptide pair first isolated from the MS1. Mango instead relies on a single isolation and fragmentation event to produce fragments that correspond to both species. Intuitively, one would expect that the characteristics of cross-linked species that provide good fragmentation for Mango may not be identical to those that produce good spectra in ReACT. It is likely that some pairs of peptides are difficult to identify from chimeric spectra due to one peptide fragmenting significantly better than its partner[43], a problem which is alleviated in ReACT by isolating and fragmenting each peptide individually. Ultimately this suggests complementarity between Mango and MS3-based methods, where there is a trade-off between the increased sampling rate enabled by producing chimeric spectra versus the time-intensive independent isolation and fragmentation of individual peptides offered by a method like ReACT.

2.5 CONCLUSION

XL-MS provides a unique way of interrogating protein-protein interactions in their native cellular context. Knowledge of the identities of these interacting proteins and the structures of their assemblies can provide insight into a variety of biological systems. Here we describe Mango as a high throughput solution for *in vivo* or *in vitro* cross-linking experiments that requires accurate mass MS/MS capabilities that are widely available in many labs.

Mango has been employed successfully in conjunction with Comet for identifying a large number of cross-linked peptides in a complex *in vivo* cross-linked *E. coli* sample using a Q-Exactive Plus for data acquisition. These results compare favorably to those acquired using an MS3-based cross-linked identification strategy, but the two data sets also show complementarity. The depth of sampling achievable by Mango could likely be improved with dynamic fragmentation whereby each species is fragmented at multiple energies, to expand the cross-linked peptide space Mango is able to sample. While Mango is presented here using BDP-NHP and a Q-Exactive Plus, the software is agnostic to the cross-linker and instrument used, requiring only that a CID-cleavable cross-linker be used on an instrument able to acquire high-resolution MS2 spectra.

Chapter 3. EXTENDING BEYOND BINARY INTERACTIONS WITH A TETRA-REACTIVE CROSS-LINKER

3.1 ABSTRACT

Chemical cross-linking combined with mass spectrometry is a technique to study protein structures and identify protein complexes. Traditionally, chemical cross-linkers contain two reactive groups allowing them to covalently bond a pair of proximal residues, either within a protein or between two proteins. The output of a cross-linking experiment is a list of interacting site pairs that provide structural constraints for modeling new structures and complexes. Due to the binary reactive nature of cross-linking reagents, only pairs of interacting sites can be directly observed, and assembly of higher order structures typically requires prior knowledge of complex composition or iterative docking to produce a putative model. Here we describe a new tetrameric cross-linker bearing four amine-reactive groups, allowing it to covalently link up to four proteins simultaneously, and a real-time instrument method to facilitate the identification of these tetrameric cross-links.We applied this new cross-linker to isolated mitochondria and identified a number of higher-order cross-links in various OXPHOS complexes and ATP synthase, demonstrating its utility in characterizing complex interfaces. We also show that higher-order cross-links can be used to effectively filter models of large protein assemblies generated using Alphafold. Higher-dimensional cross-linking provides a new avenue for characterizing complex protein interfaces even in complex samples such as intact mitochondria.

3.2 INTRODUCTION

There are a variety of analytical techniques for identifying protein-protein interactions, and they typically have a tradeoff between the rate at which interactions can be profiled and what level of structural information is obtained. High-throughput methods of interaction profiling, such as yeast two-hybrid[80] or affinity-purification mass spectrometry, can yield very large numbers of interacting proteins but provide little to no structural information about the interaction. Alternatively, methods such as cryo-EM or x-ray crystallography can provide high-resolution structural information on complexes with hundreds of proteins in a single experiment but are comparatively low-throughput. As a technique, chemical cross-linking coupled with mass spectrometry offers a compromise of throughput and structural information, offering low-resolution distance constraints from many hundreds of pairs of proteins in an experiment.

Chemical cross-linking coupled with mass spectrometry is a broadly applicable technique to obtain spatial information about proteins and protein complexes. The knowledge gained from an identified cross-link includes the identity of two proximal residues that were linked to each other, which indicates the presence of a conformation where the residues are sufficiently close to be cross-linked through a solvent accessible distance. This distance constraint is a function of the cross-linker structure, and has an impact on what cross-links are possible to form[17]. When formed between two different proteins, cross-links provide information about the identity of proteins in complex with each other and can provide sufficient information to assemble complex structures[11,51,81,82]. Within a single protein, intra-links can provide information about conformational states of a protein and provide a proxy by which the relative populations of conformers can be measured[65].

Cross-linking reagents vary dramatically in their structures and the chemistries they employ, as recently described[83]. The most common cross-linkers employed for whole proteome studies contain amine-reactive esters for targeting lysines and incorporate labile bonds in their backbone that break under CID or HCD activation. These labile bonds reduce the problem of cross-link assignment from a quadratic search to a pair of linear searches[32]. Beyond these initial parameters, cross-linkers have been designed to target carboxylic acids[11,14] or thiols[84], or alternatively incorporate a novel affinity tag like phosphonic acid for IMAC enrichment[22]. We explore a new direction in cross-linker design by increasing the number of reactive groups on the cross-linker, allowing it to react with up to four proteins simultaneously. This not only increases the level of structural constraints of model protein structures, but provides unique indication of the existence of interactions among more than two protein partners in a complex sample.

Here we present the synthesis of a tetrameric cross-linker, Bisby, and describe an associated real-time instrument method for automated targeting of released peptides to facilitate their identification. The molecule and method were initially tested by cross-linking bovine serum albumin (BSA), a staple protein for benchmarking cross-linkers[85]. Finally, we applied this new method to mitochondria isolated from mouse hearts and explored the use of Alphafold2[5] to model a trimeric interface in combination with a higher-order cross-link.

3.3 METHODS

Cross-linker Synthesis

The peptide backbone of the tetrameric cross-linker, Bisby, was prepared on 0.05 mmol of low-loading rink amide protide resin (CEM) using Fmoc-based solid phase synthesis on a CEM Liberty

Lite peptide synthesizer, with the sequence Lys-Lys_2-Pro_4-Asp_4-$succinate_4$. Following the succinylation, the resin was transferred to a Poly-prep column (Biorad) and incubated in 12-fold molar excess of TFA-NHP in pyridine for 20 minutes to yield the four terminal NHP-esters. The pyridine was then removed by vacuum filtration, and the resin was washed with 3x10 mL of DMF, allowing the final wash to incubate with the resin for 10 minutes before removal. Three additional washes were performed each with 10 mL of DCM to remove excess DMF. After washing, the esterified product was cleaved from the resin by incubation in a 2 mL cleavage solution comprised of 95% TFA in DCM for 3 hours at room temperature with gentle rocking. The linker was precipitated by adding the cleavage solution to 35 mL of cold diethyl ether and was further washed using 4x20mL of cold ether. The resulting pellet was dried to completion by vacuum centrifugation.

Bisby purification

Reverse phase chromatography was carried out on an Agilent 1200 series HPLC system using a 250mm x 9.4mm Partisil 10µm ODS-3 column (Whatman) equipped with a 20mm x 2.1mm BetaBasic C18 Javelin guard column (Thermo Scientific). The mobile phases consist of 0.1% TFA in water (Solvent A) and 0.1% TFA in acetonitrile (Solvent B). Crude product was resuspended at a concentration of 20 mg/mL in 0.1% TFA in 50% acetonitrile, and 10 mg were loaded per injection. Product was separated at a flowrate of 0.5 mL/min for 15 min with the following gradient: 50%B at 0 min, 70%B at 10 minutes, 50%B at 10.01 min until completion. Fractions were collected at 1-minute intervals starting at 4 minutes, and fraction 7 was retained (Figure B.5.9). Fraction 7 was pooled across all injections and dried to completion by vacuum centrifugation before being resuspended in DMSO for use. The purity of the stock solution was

estimated by direct infusion (Figure 3.1) to be 35% for fully intact Bisby, increasing to 67% when considering the first and second hydrolysis products which can still generate productive cross-links.

BSA Cross-linking

One mg of bovine serum albumin (BSA) was dissolved in 1mL of 170 mM Na_2HPO_4 pH 8.0. Bisby was then added to a final concentration of 0.1 mM, and the reaction was allowed to proceed for 30 minutes at 27°C with constant shaking at 600 rpm in a thermomixer. After cross-linking, urea was added to a final concentration of 8M, and then tris(2-carboxyethyl)phosphine (TCEP) was added to a final concentration of 5 mM to reduce disulfides, allowing the reduction to continue for 30 minutes at 27°C at 600 rpm on a thermomixer. Following reduction, iodoacetamide (IAA) was added to a final concentration of 15 mM, and cysteines were alkylated by mixing for 30 minutes at 27°C at 600 rpm on the thermomixer. After alkylation, the solution was diluted 8-fold with 100mM ammonium bicarbonate pH 8.0, trypsin was added at a 1:200 ratio of trypsin to protein, and digestion was carried out overnight at 37°C shaking at 600 rpm on a thermomixer. The resulting digest was desalted using Waters C18 sep-paks on a vacuum manifold, and the desalted peptides were then fractionated using peptide size exclusion chromatography.

Animal model

All protocols concerning animal use were approved by the Institutional Animal Care and Use Committee at the University of Washington. This study utilized six wild-type mice, strain C67Bl6/NCrl (IMSR_CRL:27). Adult (>10 weeks old, weighing 22-24g) male mice were

chosen randomly. Mice were housed in a vivarium with a 12-hr light/dark cycle at 22°C. Mice were maintained on ad libitum standard rodent diet and water.

Mitochondria isolation from cardiac tissue

Hearts were excised from mice and the aortas and atria were removed. Heart tissues were rinsed briefly in ice-cold mitochondria isolation medium (MIM: 70mM sucrose, 220mM mannitol, 5mM MOPS, 1.6mM carnitine hydrochloride, 1mM EDTA, 0.025% fatty acid-free BSA, pH 7.4 with 5M KOH), to remove residual blood. Tissues were minced on ice and resuspended in fresh MIM, followed by trypsin digestion (10□g/ml) and incubated on ice for 10 min. Trypsin digestion was stopped by the addition of trypsin inhibitor (0.5mg/ml) and additional BSA (1 mg/ml) to MIM. The suspension was centrifuged for 1 min at 1,500 x g at 4°C, and the supernatant was discarded. The tissue pellets were resuspended in fresh MIM containing 1mg/ml BSA, and transferred to a Teflon-glass tube and homogenized on ice with a Teflon pestle. The homogenates were centrifuged for 10 min at 800 x g at 4°C. The supernatants were collected and centrifuged for 10 min at 8,000 x g at 4°C. The supernatant was discarded, and the mitochondrial pellets were resuspended in MIM to wash. The resuspension was centrifuged for 10 min at 8,000 x g at 4°C, and the supernatant discarded. The mitochondrial pellet was used for the cross-linking reaction.

Mitochondria Cross-linking

Six total mice were used to generate four mitochondrial samples for cross-linking. The first sample comprises half the mitochondria gathered from a pool of three mice hearts, while the remaining three samples are each generated from a single mouse. Each sample was resuspended in 100 µl of 170 mM Na_2HPO_4 pH 8.0, and Bisby was added to a final concentration of 10 mM.

The solution was mixed at 600 rpm in a thermomixer at 27°C for 45 minutes. Samples were centrifuged at 8000g to pellet the mitochondria, and the supernatant was removed. Each sample was resuspended in 500 µl of 100 mM ammonium bicarbonate pH 8.0 and pelleted again at 8000g to wash. The samples were then resuspended in 100 µl of ammonium bicarbonate pH 8.0 containing 48 mg of urea. The mitochondria are then lysed by sonication using a GE-130 ultrasonic processor. The samples were reduced with 5 mM TCEP for 30 min at 600 rpm in a 27°C thermomixer and then alkylated with a final concentration of 15 mM iodoacetamide for 45 minutes in the dark. The samples were diluted to a final volume of 800 µl with 100 mM ammonium bicarbonate pH 8.0 and digested overnight at 37°C with 1:200 ratio of trypsin. Following digestion, samples were desalted with a Waters C18 sep-pak, dried to completion by vacuum centrifugation, and subjected to peptide size exclusion chromatography.

Size Exclusion Chromatography

Cross-linked peptides from both the BSA and mitochondrial samples were fractionated by peptide size-exclusion chromatography (SEC) using an AKTA Pure system (GE) equipped with a Superdex Peptide 10/300 GL column (GE). Desalted peptides were resuspended in 0.5 mL of 30% ACN/0.1% TFA in water. Peptides were fractionated using an isocratic flow at a rate of 0.5 mL/min consisting of 70% solvent A (0.1% TFA in water) and 30% solvent B (0.1% TFA in ACN) for 1.1 column volumes (CVs). 1 mL fractions were collected during elution, starting at 0.2 CV. For BSA, only fraction A5 was carried forward for LC-MS analysis. For the mitochondrial samples, fraction A5 and 200 µl of A6 were carried forward for LC-MS analysis, while the remaining 800 µl of fraction A6 were further fractionated by SCX.

Strong Cation Exchange Chromatography

Fraction 6 from the peptide SEC from each of the mitochondrial samples was further fractionated by SCX, as it contains a significant number of binary linked peptides that separate efficiently by SCX. SCX was carried out on an Agilent 1200 series HPLC system equipped with a 250 x 10 mm column packed with Luna 5 µm 100 Å particles (Phenomenex). The mobile phases consisted of 7 mM KH_2PO_4, 30% acetonitrile pH 2.8 (Solvent A) and 7 mM KH_2PO_4, 350 mM KCl, 30% ACN pH 2.8 (Solvent B). Fraction 6 was resuspended in 0.5 mL of solvent A prior to injection. Peptides were fractionated at a flow rate of 1.5 mL/min for 97.5 min with the following gradient: 0% B at 0 min, 5% B at 7.5 min, 60% B at 47.5 min, 100% B at 67.5 min, 100% B at 77.5 min, 0% B at 77.51 min to completion. Fractions were collected at 5-minute intervals starting at 17.5 minutes and were combined into 5 final pools consisting of fractions 6-7,8,9,10, and 11-14. Fractions were concentrated to approximately 1 mL final volume by vacuum centrifugation to remove acetonitrile, and then were desalted using Waters C18 sep-pak cartridges

LC-MS Analysis

Peptides from all fractions were analyzed using a Waters NanoAcquity UPLC coupled to a Thermo Velos Fourier-transform ion cyclotron resonance mass spectrometer[72] (Velos-FT). Samples were fractionated over a 60 cm x 75 µm inner diameter fused silica analytical column, maintained at 45°C and packed with ReprosSil-Pur C8 (5 µm diameter, 120 A pore size), by applying a linear gradient 12% to 30% solvent B (0.1% formic acid in acetonitrile) in solvent A (0.1% formic acid in water) at a flow rate of 300 nL/min. The Velos-FT was operated using a real-time strategy developed for the analysis of tetra-linked peptides. A top 1 DDA method was used

in which a high resolution (50000 mass resolving power at 400 m/z) MS1 scan from 400 to 2000 m/z is taken, followed by a high resolution (50000 mass resolving power at 400 m/z) MS2 scan on the most abundant ion of charge 4+ or greater not currently in dynamic exclusion. Each MS2 scan was processed in real-time to determine if a set of 4 peaks fulfilling a mass relationship were present, and if they were then 2 low resolution ion trap scans were taken for each valid target. MS1 scans used an AGC target of $5x10^5$, MS2 scans used an AGC target of $2x10^5$, and MS3 scans used an AGC target of $1x10^5$. MS1 and MS3 scans used a maximum injection time of 500 ms, while MS2 scans used a maximum injection time of 1500 ms.

Real-time Instrument method

ReACT4 is an updated version of the original ReACT[33] implementation for real-time targeting and sequencing of cross-linked peptides, and was implemented in ion trap control language (ITCL), the native control language used on LTQ series Thermo Scientific mass spectrometers. Whenever an MS2 spectrum was acquired, its full peak list was pulled from the acquisition system. First, this peak list was scanned to determine if any peak corresponds to the precursor mass – 215.042987, as this suggests that at least one arm of the cross-linker was hydrolyzed. This peak list was then de-isotoped to remove redundant peaks from downstream consideration. In an initial pass of the de-isotoped list, the peak list was analyzed to annotate all pairs of complement ions, any pair of neutral masses that add up to the precursor mass. Following the detection of complement ions, all peaks with charges greater than 3 were discarded. A peak of neutral mass = 215.042987 was prepended to the peak list, which corresponds to the mass of a hydrolyzed arm. This reduced peak list was then used to construct an array of all pairwise sums of peaks in the spectrum. The array of pairwise sums was sorted by the summed neutral mass of its

two constitutive peaks. Due to memory limitations on the instrument this procedure is limited to the 62 most abundant peaks in the reduced peak list, and the sort was performed by sorting up to four 500 element chunks which were then merged to produce the final sorted array. After the array was sorted, all solutions that fulfill the mass relationship

$$m_{precursor} \cong m_{reporter} + \sum_{i=1}^{4} m_{peak(i)} \text{ (eq. 3.1)}$$

were identified in a single pass of the array. In **equation 3.1**, $m_{precursor}$ is the neutral mass of the precursor, $m_{reporter}$ is the neutral mass of the reporter ion (789.525 for Bisby), and $m_{peak(i)}$ is the neutral mass of some peak in the spectrum. Solutions to the mass relationship were found with a tolerance of 20 ppm. Once all solutions were determined they were ranked first by the number of unique complement ions observed for the set of four peaks, and then the solution with the highest summed intensity of the set of four peaks was selected for scheduling. If a peak corresponding to the precursor mass-215.042987 was detected, then solutions with at least one hydrolyzed arm were prioritized. For each peak, two ion-trap MS3 scans were scheduled, one targeting the observed peak that triggered the solution, and a second that targets either the same peak again or the calculated 2+ charge state of the ion if a 1+ ion was targeted for the first scan. Peaks corresponding to the mass of a hydrolyzed arm had no scans scheduled. If three or more peaks in the selected solution corresponded to hydrolyzed arms, then no scans were scheduled.

Data searching

The MS3 spectra were searched using Comet[59] version 2019.01.5 using the default parameters for analysis of low-resolution spectra with a high-resolution precursor mass with the following modifications: isotope_error=5, allowed_missed_cleavage=5, ms_level=3; variable

modifications: 15.9949 at M, 42.010565 at protein N-termini; required modification: 197.032422 at an internal K. The BSA search additionally included a variable modification of -17.02655 at peptide N-terminal C. BSA samples were searched against the uniprot bovine database, while mitochondria samples were searched against the MitoCarta 2.0 database[86]. Both databases were supplemented with reverse protein sequences for all entries to serve as decoys during the search. After searching with comet, PSMs were processed using PeptideProphet[45] and iProphet[87] to assign probabilities to all PSMs, as well as ProteinProphet[88] to perform protein inference. PSMs were filtered to a 1% FDR based on their iProphet probabilities, and then they were assembled into their appropriate grouping of tetra-/tri-/binary-links based on their parent MS2 scan. To account for the accumulation of decoys and corresponding increase in FDR associated with grouping, we performed an entrapment search (Supplemental Methods B) to estimate the FDR at the cross-link level, which is estimated to be 4.4% (Figure B.5.10).

Structural Modeling

Colabfold[89] was used to generate models of the ATPA/ATPG/ATPE hetero-trimer. Uniprot mouse sequences (Q03265, Q91VR2, P56382 for ATPA, ATPG, ATPE respectively) with the annotated mitochondrial signaling peptide removed from ATPA and ATPG were used as input. To generate models, we used 32 random seeds with 5 models each to generate a pool of 160 models. These models were clustered using Calibur[90], and clusters with more than 10 members were retained yielding 5 final clusters. The representative model from each cluster was then compared against an identified inter-protein tri-link between the three modeled proteins to evaluate model consistency with an observed tri-link.

Data Availability

Raw files for all samples are available on PRIDE[91] with the dataset identifier PXD032222 with username reviewer_pxd032222@ebi.ac.uk and password gFGy1m6P.

3.4 Results and Discussion

Linker Synthesis

Bisby is a tetrameric cross-linker prepared using solid-phase peptide synthesis (SPPS) employing traditional fmoc-based protection chemistry. SPPS provides the advantage of fast prototyping of molecules while enabling diverse functionalities to be incorporated and has been previously used in the synthesis of various protein interaction reporter[10] (PIR) cross-linkers. These previous PIR cross-linkers have included photocleavable groups[92], affinity enrichment tags[33], and recently isobaric reagent sets for quantitative interactome studies[19,42]. The primary distinguishing feature of Bisby from previous cross-linkers is the inclusion of four reactive NHP-esters, allowing it to react with up to four proximal primary amines during a cross-linking experiment. As with other members of the peptide-based Protein Interaction Reporter (PIR) family of cross-linkers, Bisby contains a CID-cleavable aspartyl-prolyl (DP) bonds in each arm that preferentially fragment during CID-activation (Figure 3.1). The inclusion of CID cleavable bonds enables facile MS3 targeting of released peptides by making use of a mass relationship connecting the precursor to its released peptides (Equation 3.1).

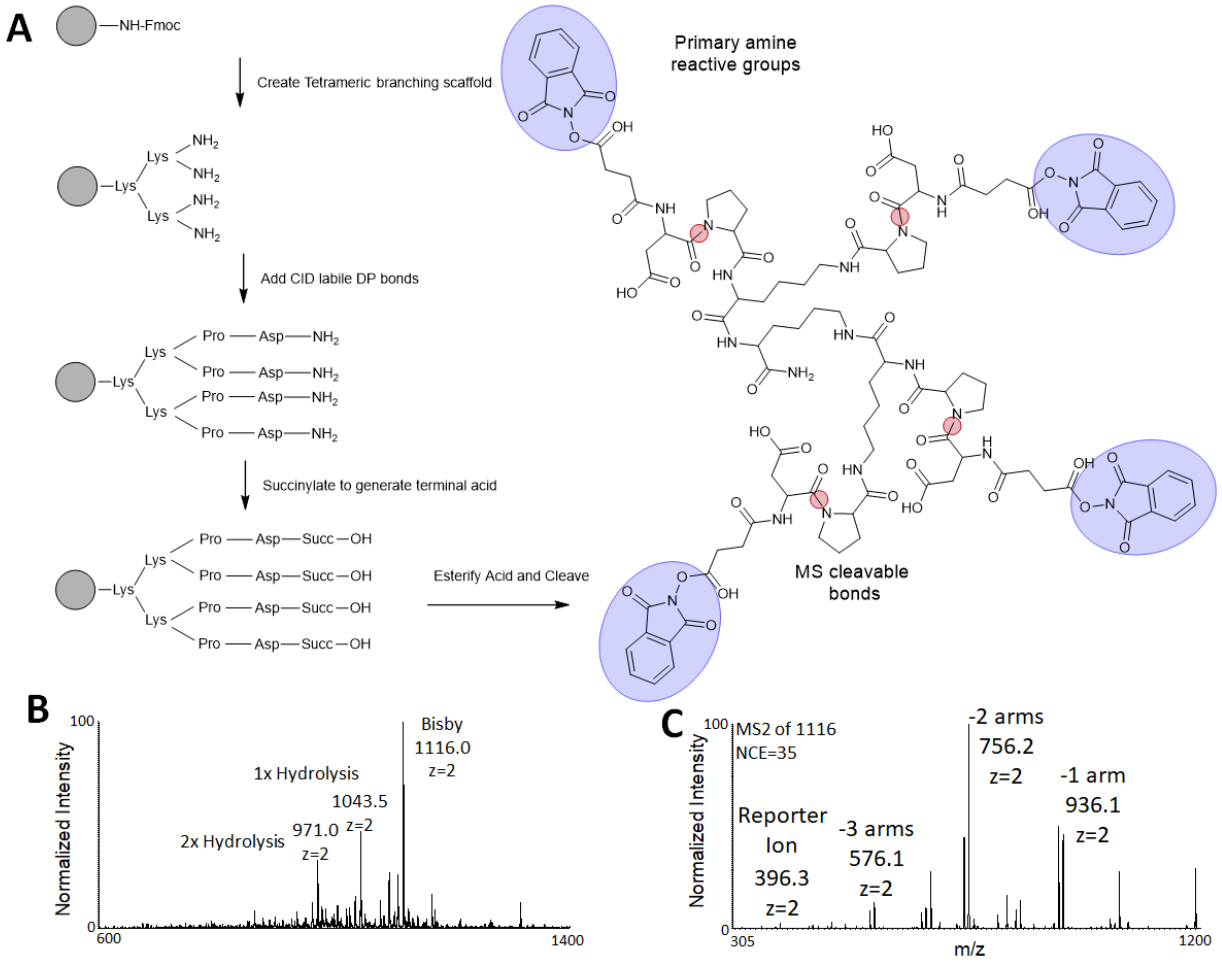


Figure 3.1. (A) Outline of synthesis of Bisby. Primary amine-reactive NHP esters are highlighted in blue and mass spectrometry cleavable DP bonds are highlighted in red. (B) Direct infusion spectrum of Bisby showing the intact species and partial hydrolysis products. Additional peaks corresponding to water loss (-18 Da) or TFA adducts (+97 Da) are also present. (C) CID fragmentation spectrum of Bisby (m/z = 1116) showing distribution of product ions formed by fragmentation of one to four DP bonds. CID activation predominantly leads to the fragmentation of one or two DP bonds.

Following low-PH reverse phase purification, we obtained Bisby and some partial hydrolysis products. CID-activation (NCE=35) of the 2+ charge state of Bisby yields a fragmentation spectrum with a distribution of DP bond cleavage events. The most abundant

fragmentation product is formed by cleavage of two DP bonds, but fragments resulting from the cleavage of one, three, or all four DP-bonds are also observed. For each of these major fragmentation products, we also observe a second peak corresponding to water loss, likely from one of the aspartic acids. In general, we expect to observe a distribution of DP-bond cleavage events under CID activation, as ions fall out of resonance after fragmentation, causing some labile bonds to go unbroken.

Instrument method

Real-time mass spectrometry methods have become more sophisticated in recent years, as they facilitate deeper and more efficient ways to interrogate samples. Real-time Comet searches enable improved quantitation with shorter gradients for enhanced[93] throughput, while augmented deconvolution and multiple precursor targeting[94] for top-down studies has been shown to improve sequence coverage. ReACT4 is a modified version of the original ReACT[33] platform for real-time targeting of cross-linked peptides generated by cleavable cross-linkers and is the first implementation of a cross-linking strategy to identify cross-links composed of more than two peptides. The overview of the targeting strategy of ReACT4 is outlined in Figure 3.2. First, a high-resolution MS1 spectrum is acquired. The high-resolution scan is required to resolve the high mass and charge that tetra-linked peptide quartets typically carry. From this spectrum, an MS2 target of charge 4+ or higher is selected using traditional DDA logic and obeying dynamic exclusion, and a high-resolution MS2 spectrum is acquired. Charge state generally increases with the number of attached peptides, and a charge of 4 or higher facilitates sampling of both binary links and higher

order links during a single run. This MS2 spectrum is used as input to ReACT4, wherein a set of 4 peaks fulfilling the mass relationship are determined using equation 3.1.

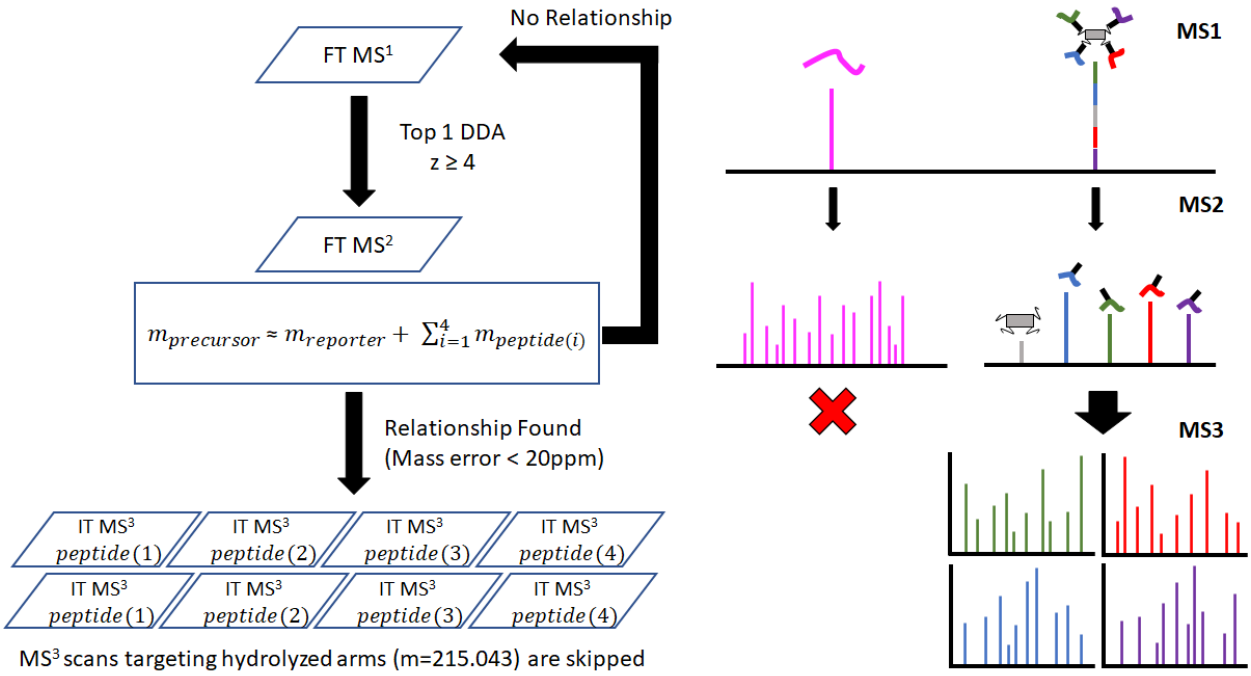


Figure 3.2. Schematic representation of ReACT4 for real-time targeting of tetra-linked peptides. MS2 spectra are processed in real-time, and MS3 scans targeting each released peptide are scheduled only if a valid mass relationship was detected. Hydrolyzed arms are added to the peak list for all MS2 spectra to enable detection of binary, ternary, and quaternary cross-links during a single experiment, but are skipped for MS3 targeting.

Beyond the ability to detect tetra-linked peptides, ReACT4 includes several additional features to improve runtime and accuracy. Input peaks from the MS2 spectrum are de-isotoped prior to conversion to neutral mass, and peaks with charges greater than three are not considered for the relationship detection. Both heuristics reduce the number of candidate peaks, and due to

sort-related memory limitations on the instrument only 62 candidate peaks can be considered for each spectrum. Reducing the number of candidate peaks makes the most of limited memory while also improving runtime, as the number of solutions to the mass relationship to evaluate is quadratic with respect to the number of input peaks as solutions are found by iterating through a sorted list of all possible pairs of ions. Additionally, ReACT4 considers the presence of complement ions in the MS2 spectrum when selecting a peak set to target for MS3. Bisby contains 4 MS-cleavable bonds, and during CID activation, it is very common that only one of the bonds breaks to yield a released peptide but also yields the corresponding complement ion. As such, when determining the final set of peaks to target, solutions for which the greatest number of unique observed complement ions are preferentially selected. Finally, ReACT4 can detect relationships resulting from partially hydrolyzed cross-links containing zero to four waters, enabling simultaneous identification of 2-, 3-, or 4-armed cross-links from a single run.

An MS3 based identification strategy was selected over a chimeric spectrum identification strategy for characterizing tetra-links. Chimeric spectrum identification strategies[34,36,95] are often favored over MS3 based strategies due to the faster duty cycle allowing a greater number of precursors to be sampled during an LC-MS run, while also not requiring MS3 capable hardware or real-time targeting methods. The primary downside to chimeric spectrum identification is that MS2 spectra resulting from the co-fragmentation of released peptides can yield spectra dominated by fragments from one of the peptides, making identification of the cross-linked species difficult or impossible[96]. Intuitively, this downside is exacerbated with up to four co-fragmenting peptides in a single cross-link, as fragments need to be observed from all four peptides to fully identify a tetra-link. An MS3 based strategy avoids this difficulty, as only the released peptides and not their

fragments need to be observed in the MS2 spectrum to enable MS3 targeting. MS3 isolation and fragmentation of each released peptide occurs serially and independently, improving the ability to obtain fragments from all four released peptides, at the cost of a slower duty cycle.

Cross-linking a purified protein

For initial testing of the new cross-linker, we cross-linked bovine serum albumin (BSA), a staple protein standard for testing cross-linkers and cross-linking methods[85]. We utilized peptide size exclusion chromatography for enrichment of quaternary and ternary cross-links, which has been demonstrated previously as an efficient strategy for enriching binary cross-links[97]. A detailed example of an identified tetra-link is shown in Figure 3.3. A high-mass high-charge ion was selected from the MS1 (Figure 3.3A), which generates a rich fragmentation spectrum after activation (Figure 3.3B). Candidate released peptides for this spectrum were detected at a sub-3ppm mass error according to equation 1. All four of these released peptides also have a complement ion corresponding to the precursor minus one of the released peptides (Figure 3.3). These complement ions are used as the primary method to rank solutions when multiple mass relationships are detected in a single spectrum, with solutions containing the most unique complement ions being selected. Additional complement ions resulting from the loss of any combination of two released peptides can also be observed (Figure 3.3), but they are not currently used for ranking potential solutions.

Following the successful detection of a relationship from this MS2 spectrum, 8 MS3 scans are scheduled, 2 for each target with the first scan targeting the observed peak fulfilling the mass

relationship and the second targeting at least a 2+ charge of the same species, to increase the chance of obtaining an identifiable fragmentation spectrum as 1+ ions often fragment poorly. The higher scoring of each of those pairs is shown in Figure 3C, and it is this set of spectra that is used to fully identify the tetra-linked peptide species. This tetra-link can be mapped onto a crystal structure of BSA (pdb: 3V03, Figure 3D), which yields C_α-C_α distance ranging from 13.5 to 33.0 Å, within the expected maximum span of the linker of 45 Å.

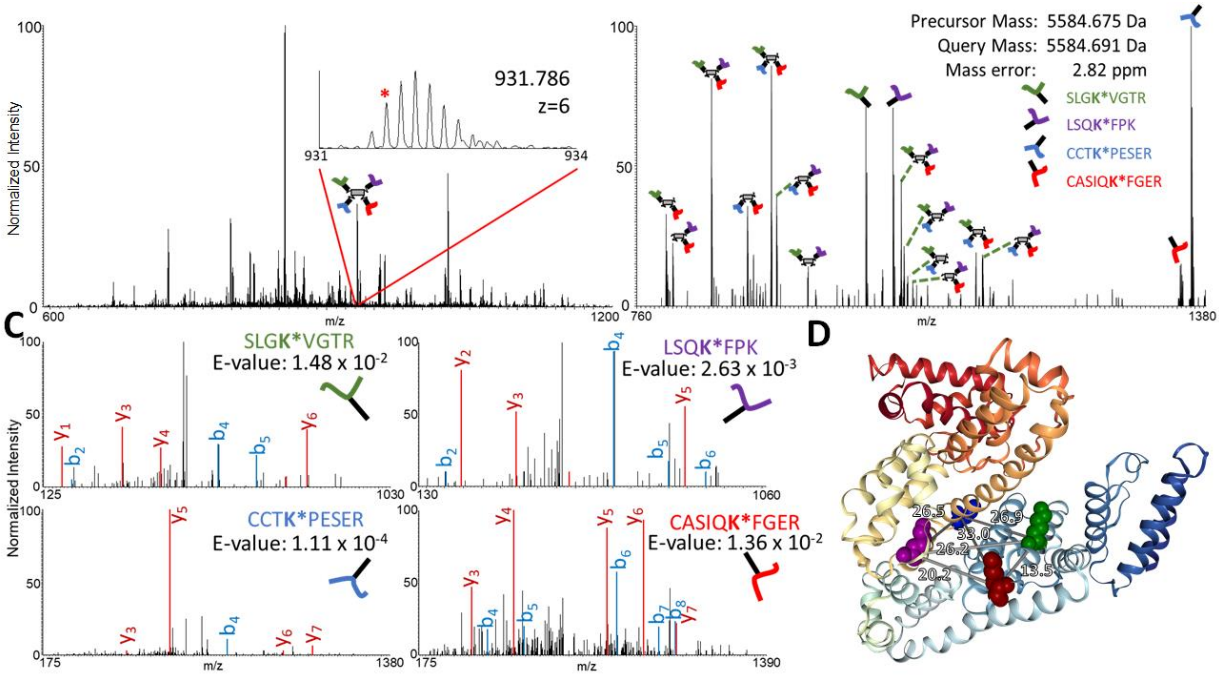


Figure 3.3. (A) MS1 signal with inset showing isotope envelope selected for MS2 analysis. (B) Fragmentation spectrum of selected tetra-linked species with released peptides and complement ions formed from one and two peptide losses annotated. (C) Highest scoring MS3 spectrum from each released peptide annotated in panel B. (D) Cross-linked residues corresponding to this identified tetra-link mapped onto BSA (pdb: 3V03) with all pairwise C_α-C_α distances annotated.

Beyond this tetra-link, we additionally identified 30 other quaternary links, 105 ternary links, and 91 binary links. In total, these links can be represented as 103 unique lysine-lysine site linkages within a monomer of BSA. The number of unique lysine pairs identified is comparable to the average of 78 found across a diverse number of cross-linkers and workflows[85], albeit with an additional technical replicate contributing additional unique pairs. While these cross-links can be predominantly mapped to BSA monomers, there are several links that contain unambiguous homo-dimer associations, and a small number of links that contain unambiguous homo-trimers of a peptide. Interestingly, in all cases where a homo-trimeric link is detected, the same peptide (SLGKVGTR) appears as the homo-trimeric peptide, suggesting some consistent assembly of a higher order homo-oligomer. BSA is known to form dimers and potentially trimers[98] depending on its concentration, so the observation of unambiguous homo-trimers seems physically plausible. All these links provide information about the spatial relationship of at least two lysines, yielding 2 to 6 distance constraints depending on the number of peptides identified in the cross-link. In contrast to traditional binary cross-linkers, not all arms of Bisby need to be assigned to a peptide sequence to provide structural information. In cases with hydrolysis of some arms or poor spectral evidence for some released peptides, structural information is still obtained if at least two arms can have a peptide sequence assigned.

Cross-linking isolated mitochondria

Bisby was further applied to cross-link mitochondria isolated from mouse hearts to explore its performance in a complex system. Mitochondria serve as an ideal system due to their many well characterized complexes and super-complexes. In our initial application of Bisby to mitochondria, we identified 39 unique quaternary links, 186 unique ternary links, and 801 binary

links after FDR filtering. In total, these can be flattened to 691 unique lysine-lysine site pairs involving 249 protein pairs and can be viewed in the **MitoBXP_mouse_mixed_Bruce** table on XLinkDB[74]. The comparison of $C_\alpha - C_\alpha$ in various mitochondrial cross-linking experiments (Supplemental Methods) reveals a correlation between the observed distance distribution and cross-linker spacer-arm length, with Bisby forming links between lysines that are farther apart on average compared to a smaller cross-linker like DSSO (Figure B.5.11). These links are concentrated heavily in electron transport chain complexes, large chaperones, and members of the TCA cycle. For example, we identify an unambiguous homo-tetrameric link between copies of CH10 (Figure B.5.12), consistent with its known heptameric assembly (PDB: 4PJ1). Overall, the depth of site pair coverage obtained by Bisby is less than commonly used binary cross-linkers, but unlike binary cross-linkers Bisby provides the benefit of unambiguous identification of three and four protein interfaces. Greater depth could likely be obtained through more extensive fractionation or the inclusion of an affinity tag such as a biotin tag[33], azide group[99], or phosphonic acid[22] as with previous cross-linkers, at the tradeoff of a more difficult synthesis.

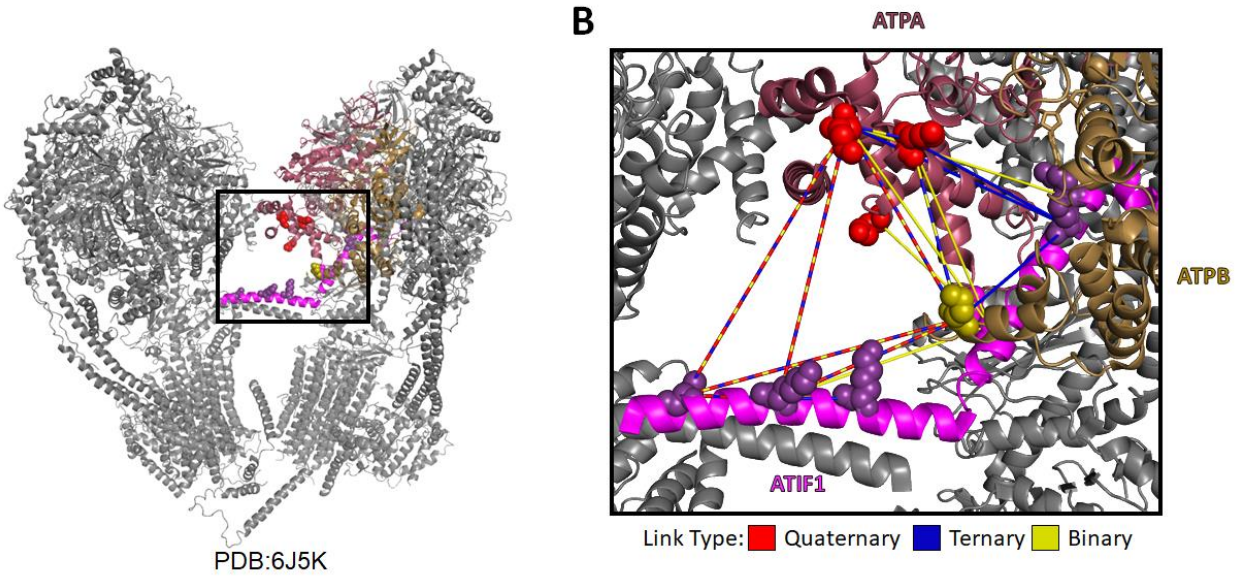


Figure 3.4. (A) Positional context for interaction between ATIF1 and ATP Synthase from half of the tetrameric structure. (B) Zoomed in inset highlighting an identified tetra-link (red) between two residues of ATIF1, and one residue each from ATPA and ATPB. Also shown are all the ternary (blue) and binary links (yellow) identified between ATIF1 and ATPA or ATPB. The tetra-link identified here is supported by all four possible ternary links and all six possible binary links between the four linked lysines.

One of the tetra-links identified forms a link between ATPA, ATPB, and ATIF1. ATPA and ATPB are catalytic subunits of ATP synthase (CV), and the ATPA-ATPB interfaces form the catalytic pockets for ATP synthesis. This quaternary link represents the first unambiguous identification of an interface formed from three different proteins from a single cross-link. ATIF1 is a known CV inhibitor that functions by halting rotation of the central stalk thought to occur during conditions with low matrix pH to prevent reverse rotation and consumption of ATP by CV. This tetra-link, as well as all the other links between ATIF1 and ATP synthase, are shown in Figure 3.4, consistent with the cryo-EM structure of ATIF1 in complex with ATP synthase. Further reinforcing confidence in this tetra-link assignment, all four potential constitutive ternary links and

all six potential binary links were also observed. Although already known from 6J5K, the tetra-links and tri-links in Fig 4B conclusively identified the ternary ATPA-ATPB-ATIF1 interaction that was present in mitochondria at the time of cross-linker application. This conclusion is not generally possible with consideration of binary links only and offers new opportunities for mapping multi-subunit complexes in mitochondria or other complex systems.

One of the new functionalities enabled by Bisby is the unambiguous identification of protein complex interfaces comprised of more than two proteins. Traditionally, multi-protein interfaces are assembled from binary cross-linking data by using observed cross-links combined with prior knowledge of complex composition and iterative docking methods to assemble complexes protein by protein. Iterative docking to assemble complexes of more than two proteins using constraints provided by multiple binary cross-links can be problematic, as some complex members could be mutually exclusive[100,101]. Binary interaction data connecting three different proteins does not provide the ability to discriminate, for example, between a set of three heterodimers or a single heterotrimer without additional information. Bisby helps resolve this ambiguity by facilitating identification of up to four proteins at an interface, which can only occur if none of those identified proteins are mutually exclusive with each other. These multi-protein interfaces can now be modeled in a single step using Alphafold[5], which enables simultaneous modeling of an arbitrary number of proteins in a complex[56]. This method of modeling has the added benefit of allowing conformational rearrangements associated with complex formation which is not captured by rigid-body docking, and may improve accuracy of the resulting complex[57].

To explore this combination of a tetrameric cross-linker and Alphafold2 modeling of protein complexes, a single tri-link to model the three-protein interface between ATPA, ATPG, and ATPE from ATP synthase was used. Preliminary results suggest that agreement with cross-linking distance constraints correlates with higher confidence dimer models from Alphafold2[58]. A potential challenge for this assembly for traditional iterative docking is that ATPA and ATPE share no interface in any rotational state, making it challenging to determine their relative orientation in the absence of additional constraints. A single wholly inter-protein tri-link provides two important pieces of information. It provides unambiguous evidence that three proteins were in close-proximity as well as supplying a set of three distance constraints that must be fulfilled in some conformation. These two pieces of information provide both a target to model, and information by which to filter resulting models for accuracy.

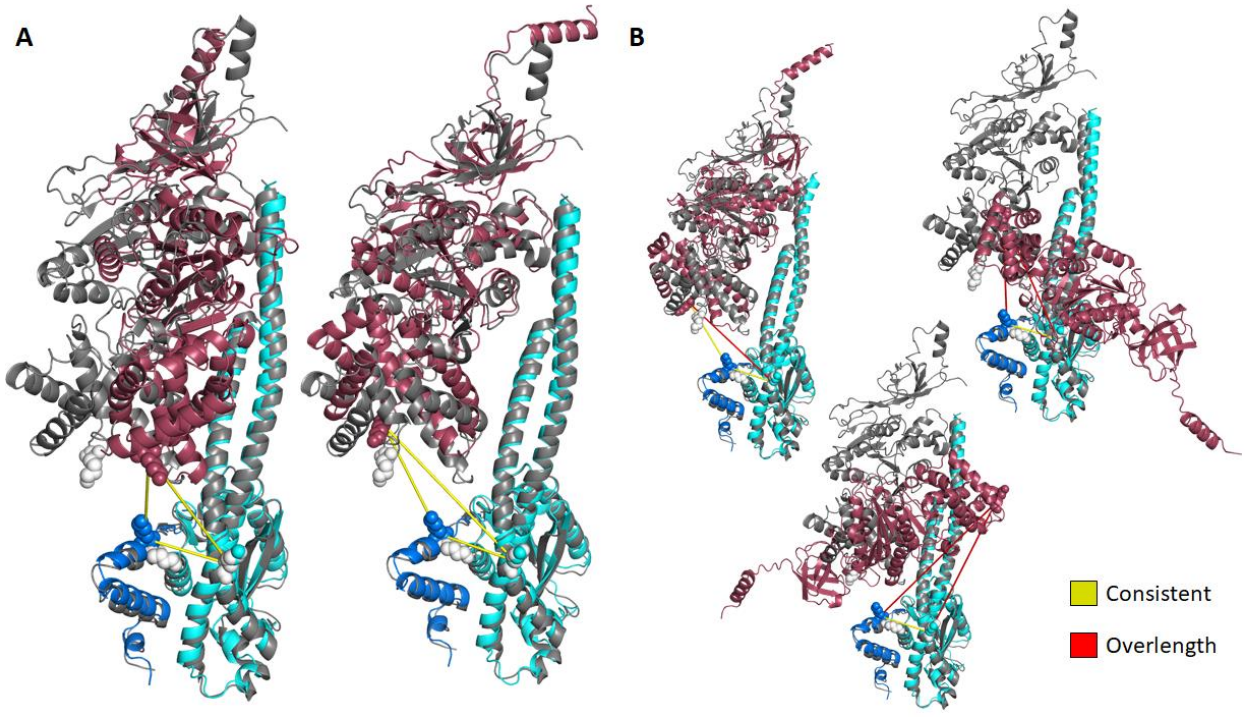


Figure 3.5. (A) Representative models from clusters that are consistent with an observed ternary link between ATPA (red), ATPG (cyan), and ATPE (blue), aligned against the same proteins from one rotamer of ATP synthase (grey; pdb: 6J5I; chains: A,G,I). A ternary link yields 3 pairwise distance constraints that must be consistent with some assembly. (B) Representative models from clusters that can be rejected due to the presence of at least one unfulfilled distance constraint derived from the same ternary link.

Colabfold[89] was used to generate 160 models of these three proteins in complex to capture a breadth of potential complex conformations. These models were then clustered using Calibur to find representative models, and clusters with more than 5 members were evaluated against the tri-link. We identified, two clusters that had representative models consistent with a tri-link between ATPA:K454, ATPG:K39, and ATPE:K37 (Figure 3.5) were identified. These models are very similar to cryo-EM structures obtained for ATP synthase (PDB: 6J5K, Figure 5), and differ with varying degrees of rotation of ATPA about the central axis of ATPG. In contrast, models that are inconsistent with this tri-link have ATPA rotated perpendicular to ATPG, or are rotated too far around the central axis with respect to ATPG. In these cases, one or more distance constraints supplied by the trimeric cross-link are violated, allowing us to reject the model. The models that have all three constraints satisfied have excellent agreement with known structures, especially considering the small interface shared by ATPA and ATPG, and the lack of interface between ATPE and ATPA. Currently, distance constraints from cross-linking experiments can only be used to filter Alphafold2 models and are not yet used to constrain the initial solution space. The ability to consider these constraints during modeling would likely improve Alphafold2 applications to multiprotein complex structure predictions.

3.5 CONCLUSION

XL-MS provides a method for identifying protein-protein interactions from complex systems, which can provide new biological insights. Previously, XL-MS has provided information about numerous binary interactions between or within proteins, which can be used to identify and potentially assemble complex structures. Here, traditional cross-linking chemistry and methodologies were extended to facilitate the linking and identification of up to four proteins simultaneously with the synthesis of the tetra-reactive cross-linker, Bisby. A real-time strategy for the analysis of tetra-linked products based on the original ReACT strategy was developed and used to facilitate multi-linked peptide identification, where a mass relationship was used to determine the constitutive released peptides and then schedule MS3 targeting. As an initial application, Bisby was used to cross-link BSA to demonstrate that tetra-linked peptides were formed and ReACT4 was useful for identification. Next, Bisby was applied to a more complex system of isolated mitochondria, where generation and identification of tetra-links and tri-links from chaperones and OXPHOS complexes in their native environment was demonstrated. While the initial implementation of ReACT4 was performed within ITCL on a Velos-FTICR, it could be similarly implemented using the API on Orbitrap-tribrid instruments for broader use.

Chapter 4. PARALLEL SIGNAL ACQUISITION ON AN FT-ICR ARRAY CELL

4.1 ABSTRACT

The duty cycle of Fourier transform (FT)-based mass spectrometers is determined primarily by how long of a data acquisition period is required to produce a mass spectrum of the targeted resolution. To improve the duty cycle it is necessary to shorten the time domain signal required to achieve the target resolution, or to acquire multiple spectra simultaneously. One approach to improve the duty cycle of Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometers is parallel signal acquisition in an ICR array cell, which reduces the average time required per spectrum by producing multiple spectra per detection event. While an array mass analyzer has been previously demonstrated[102,103], it was not suitable for liquid chromatography mass spectrometry (LC-MS) experiments due to the long re-arm time of the data acquisition system and manually specified ion injections. Here we show modifications to a Velos-Pro FT-ICR's instrument code to support multiple serial injections using a data dependent acquisition (DDA) peak picker, as well as a multi-channel digitizer configuration for high-throughput analog signal processing of a two-cell ICR array. We demonstrate that our acquisition strategy produces spectra of similar quality to the instrument's native workflow, and that we can dynamically select ions for injection during an LC-MS experiment. This procedure is tested using cross-linked bovine serum albumin, which confirms a significantly enhanced duty cycle, albeit with reduced spectral quality and fewer identified cross-linked peptides compared to commercial ICR cells. These results

lay the groundwork for future generations of array ICR cells that can be quickly integrated and characterized in the existing set-up.

4.2 INTRODUCTION

Mass spectrometry proteomics experiments are typically focused on the analysis of complex mixtures of peptides containing many millions of different analytes. Due to the limited time scale of liquid chromatography experiments and limited dynamic range of mass analyzers, only a small fraction of relatively abundant precursors, on the order of 10^4-10^5 different species, can be sampled during an experiment. At the same time, to efficiently identify peptides in complex samples, high mass measurement accuracy is beneficial, as the precursor mass of a peptide greatly restricts the number of candidates to be considered during a database search[104]. The primary limitation in sampling speed during an experiment is the duty cycle of FT-based mass spectrometers, where high resolution spectra suitable for peptide assignment take 10s to 100s of milliseconds of data acquisition time to produce spectra of sufficient resolution. The benefit of high-resolution spectra for identifying peptides coupled with the inherently slow duty cycle required to produce them suggests that proteomics experiments could benefit from high-resolution mass analyzers with improved duty cycles[105].

A prime candidate mass analyzer for achieving an improved duty cycle is the ion cyclotron resonance (ICR) cell due to various synergistic methods through which its acquisition rate can be scaled[106]. ICR cells are mass analyzers that measure the resulting ion motion produced when ions are moved off the central axis of a strong magnetic field, causing a circular orbit at a frequency dependent on their mass-to-charge ratio[107]. The resolution of a peak in a mass spectrum produced

by an ICR cell is directly proportional to the observed ion frequency of a peak, which is determined primarily by the strength of the magnetic field employed. Trivially, the duty cycle of a Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometer can be improved by utilizing a stronger magnetic field. For example, moving from a 7T to a 21T field[108] reduces the transient required to achieve any target resolution for any ion by two-thirds, tripling the duty cycle when ion fill times are fast compared to acquisition periods. However, producing high field magnets with a bore size large enough to hold an ICR cell is still an active area of research, and very high-field magnets like 21T systems are currently only employed at national labs.

While increasing the strength of the magnetic field is one such aspect of the ICR experiment that can be scaled, there exist several other promising avenues of improving the duty cycle. One such method is commonly referred to as harmonic or frequency-multiple detection, wherein an alternative geometry of electrodes is used to measure overtones of fundamental ion motion[109,110], increasing the resolution by the order of symmetry in the detectors. A traditional ICR cell uses two detection electrodes that are 180° out of phase to measure the fundamental ion motion, which produces a single sine wave for each complete revolution about the central magnetic field axis. A second harmonic detector uses four detection electrodes that are 90° out of phase and measure the first overtone of ion motion, where each complete revolution within the cell now produces two full sine waves, with twice the frequency of the fundamental ion motion[111]. Upon transforming this signal to the m/z domain, the resulting peaks will have twice the resolution, mimicking some of the effect of having a stronger magnetic field. This is a well scaling technique, as up to the 6^{th} harmonic has been demonstrated[112]. The primary downsides of this technique are the difficulty in producing high signal at the desired harmonic and the spectral complexity caused

by observing peaks at a variety of off-target harmonics associated with ion clouds being centered off the magnetic field axis[113].

Another promising dimension to scale FT-ICR duty cycle is by increasing the number of mass analyzers employed within a single instrument. While this improvement is not specific to ICR cells as it would be technically feasible to include multiple orbitraps or other analyzers in a single instrument, FT-ICR instruments benefit from a significantly improved ease of implementation. Including multiple orbitraps in an instrument is expensive as orbitrap analyzers are difficult to machine and consequently expensive, and each additional orbitrap requires an additional dedicated C-trap and ion transfer optics due to the requirement of synchronized ion injection into the orbitrap. In contrast, ICR cells derive their mass resolving power primarily from the magnetic field rather than the detector geometry, and as such are limited mostly by the available sensing volume within the magnetic field where the field is homogeneous. Multiple ICR cells can be placed in a line of the central axis[102], or even placed orthogonal to the central axis[103] without requiring additional dedicated optics for each ICR cell. This new type of array ICR-cell has been demonstrated with direct infusion experiments and shown to be capable of producing up to five high resolution spectra in parallel, a five-fold increase in duty cycle compared to a single cell[102].

As compared to a stronger magnetic field or harmonic detection, parallel detection in an ICR cell poses new technical challenges, as the duty cycle improvement is not derived from measuring higher frequency ion motion, but rather from measuring fundamental ion motion in parallel from several cells. In LC-MS experiments, data processing time needs to be approximately as fast as spectral acquisition, so that time-domain signals can be transformed into mass spectra and data dependent decisions can be informed by the resulting spectrum before the next acquisition

period begins. To address these concerns, we have integrated a high-performance multi-channel digitizer and developed a full analog signal processing pipeline complete with instrument communications to support LC-MS acquisitions with an array ICR cell. Furthermore, we have modified the instrument's code to facilitate automated filling of each cell in the array using traditional data dependent acquisition (DDA) logic. We demonstrate this new data acquisition system by analyzing standard solutions as well as by analyzing a cross-linked standard of bovine serum albumin.

4.3 METHODS

LTQ FT-ICR Array MS

A modified LTQ-FT (Thermo Scientific) with a Velos pro front-end and equipped with a 7 T actively shielded superconducting magnet (Jastec Japan Superconductor Technology, Tokyo, Japan) was used to acquire all spectra. The instrument was further modified with a printed circuit board array ICR cell containing a one-inch cell in the back and a two-inch cell closer to the ion source. This cell was close-coupled to a multi-channel in-vacuum preamplifier (GAA Engineering) with each cell occupying one channel of the preamplifier. The in-vacuum preamplifier is fed to an external second stage amplifier (GAA Engineering), whose output is forwarded to the external data acquisition system (DAQ, PCIE-1840L-AE, Advantech) for channel two, and is split to both the external digitizer and the instrument's digitizer for channel one. A modular intelligent power supply (MIPS, GAA Engineering) was used to provide additional DC voltages to the cell, triggered by the beginning of the ion trap ejection event to prepare for trapping in the appropriate cell.

BSA Cross-linking

One mg of bovine serum albumin (BSA) was dissolved in 1mL of 170 mM Na_2HPO_4 pH 8.0. BDP-NHP was then added to a final concentration of 1 mM, and the reaction was allowed to proceed for 30 minutes at 27°C with constant shaking at 600 rpm in a thermomixer. After cross-linking, urea was added to a final concentration of 8M, and then tris(2-carboxyethyl)phosphine (TCEP) was added to a final concentration of 5 mM to reduce disulfides, allowing the reduction to continue for 30 minutes at 27°C at 600 rpm on a thermomixer. Following reduction, iodoacetamide (IAA) was added to a final concentration of 15 mM, and cysteines were alkylated by mixing for 30 minutes at 27°C at 600 rpm on the thermomixer. After alkylation, the solution was diluted 8-fold with 100mM ammonium bicarbonate pH 8.0, trypsin was added at a 1:200 ratio of trypsin to protein, and digestion was carried out overnight at 37°C shaking at 600 rpm on a thermomixer. The resulting digest was desalted using Waters C18 sep-paks on a vacuum manifold, and the desalted peptides reconstituted at a concentration of 1 mg/mL in 2% acetonitrile/0.1% formic acid.

Multi-cell ion injection

A new ion injection routine was implemented in ion trap control language (ITCL) to automate filling of the array ICR cell, using the back cell to detect MS1 ion packets while the front cell was used to detect MS2 ion packets. The injection code was modified so that each MS1 injection was immediately followed by a second MS2 injection event prior to a detection event. In LC-MS experiments, the MS2 m/z was selected using the instrument's native DDA peak picker and obeying dynamic exclusion using the split channel coupled to the instrument's digitizer to supply the peak picker with MS1 spectra. Once an m/z had been selected, a new isolation waveform for that m/z was generated and applied, the ion packet was fragmented, and the

fragments were ejected from the trap so they could be trapped in the front cell in the array. By default, the maximum injection time for the MS2 scan is set to 500 ms but is reduced to 20 ms if no valid precursors are returned from the DDA peak picker. Both scans use the same AGC value by obligation, as this value is loaded into instrument firmware when the event list is initialized and cannot be adjusted until after a detection event. Ultimately this is summarized as a scan cycle in which an MS1 ion packet is injected and trapped in the back cell, an MS2 ion packet is injected and trapped in the front cell, and then ions in both cells are excited with a single excite waveform and detected in parallel using two channels of the external DAQ.

Parallel digitizing and analog signal processing

The instrument's analog output from two cells was coupled to the inputs of a PCIE-1840L-AE digitizer using BNC cables. A digital trigger line was connected from MIPS to the DAQ's trigger input to mark the start of the data acquisition routine coupled to the ejection of the MS2 ion packet from the ion trap. All portions of the digitizer's processing code were implemented in C++ and made use of the DAQ's API to control hardware events, which include the following parameters: sampling rate = 10 MHz, number of samples = 1960000, number of channels = 2. Effectively, after the sampling period, this returns two arrays of doubles of length equal to the number of samples that are used as input into the signal processing pipeline.

The signal processing pipeline first spawns a worker thread for each channel to be digitized, set as two channels in all the experiments described here, but supports up to four channels. Each array of analog signal data is apodized with a hann window, zero-filled to 2097152 points, and subject to real fast Fouier transform (FFT) using a precomputed FFT plan initialized before the

first scan. The frequency data is converted to a magnitude mode spectrum and is trimmed to the relevant frequency band for the ions being detected depending on the m/z range, and the frequency spectrum is subjected to peak picking using a simple rolling Z-score calculation to detect bins more than 3.5 standard deviations above the local signal. The resulting frequency peak list was converted to the m/z domain using a two-term mass calibration and peaks were assigned a charge using a previously described scoring scheme for high-resolution mass spectra[114]. The resulting m/z, charge, and intensity were fed forward into a ReACT analysis for cross-linked samples for all channels except the first channel, and then all channels were written to disc after each acquisition.

Feedback from digitizer to instrument control

Real-time communications were established between the external digitizer and the instrument control software to facilitate ReACT analysis. For each acquired MS2 spectrum, relevant scan metadata including the MS2 precursor m/z and charge are written to the instruments LTQ log file when the scan is scheduled. When the DAQ's acquisition routine is triggered, it generates the appropriate log file name based off the current date (e.g. LTQ_20220402.log, changing at midnight), and opens a connection to read the log file backwards until it finds a line containing metadata for a scan, corresponding to the ions it is currently processing. The MS2 mass spectrum returned from the signal processing pipeline is fed into a mass relationship, finding two peaks that sum to the neutral mass of the precursor ion[33] obtained from the log file. If a mass relationship is detected within 500ppm, the digitizer writes an ITCL file over the network containing a list of parameters specifying four MS3 m/z and charges. Whenever the instrument finished the MS2 injection from the serial MS1-MS2 injection, it would recompile and then execute the ITCL written by the external DAQ to load variables. These constants were then used

to schedule 4 MS3 scans of the precursors specified from the DAQ, which will schedule those scans to occur after the next MS2 scan, rather than immediately after the triggering scan. To prevent errant MS3s when no targets are found, all four MS3 precursors are set to 0, causing them to be skipped by the instrument's event builder. Whenever the DAQ detects a mass relationship, it additionally writes the full time domain signal for both channels to disc.

Analysis of calibration samples

Angiotensin II and Neurotensin, two standard peptides, were purchased from Sigma and spiked into LTQ Velos ESI Positive ion Calibration Solution (Calmix, Thermo Scientific) at a concentration of 5 µM. This sample was flowed at 3.0 µL/min by direct infusion with a spray voltage of 2.8 kV applied through a metal union to generate ions.

LC-MS

Cross-linked BSA was analyzed using a Waters NanoAcquity UPLC coupled to a Thermo Velos Fourier-transform ion cyclotron resonance mass spectrometer[72] (Velos Pro-FT). Samples were fractionated over a 60cm x 75 µm inner diameter fused silica analytical column packed with ReprosSil-Pur C8 (5 um diameter, 120 A pore size) by applying a linear gradient from 90% solvent A (0.1% formic acid in water), 10% solvent B (0.1% formic acid in acetonitrile) to 60% solvent A, 40% solvent B over 120 minutes at a flow rate of 300 nL/min. The Velos Pro-FT was operated using serial injection of MS1 and MS2 ion packets for each FT scan event, using the external DAQ to analyze both channels, while the MS1 channel was also fed back into the instrument's native digitizer. A top 1 DDA method was used in which a high resolution (25000 mass resolving power at 400 m/z) MS1 scan from 400 to 2000 m/z is taken in parallel with an MS2 scan of the same

resolution on the most abundant ion of charge 4+ or greater not currently in dynamic exclusion. Each MS2 scan was processed in real-time on the external DAQ to determine if a pair of peaks fulfilling a mass relationship are present, and if they were then 2 low resolution ion trap scans were scheduled for each valid target. MS1 and MS2 scans used an automatic gain control (AGC) target of $5x10^5$ while MS3 scans used an AGC target of $1x10^5$. MS1 and MS3 scans used a maximum injection time of 500 ms, while MS2 scans use a maximum of 500 ms for new targets and 20 ms for repeated targets.

4.4 RESULTS

Serial injection of MS1 and MS2 ion packets

The first step in creating a functional array FT-ICR instrument for LC-MS experiments is to create an ion injection routine capable of filling multiple cells before a detection event occurs. Intuitively, the general setup of an FT scan event involves the accumulation of ions in the ion trap, transferring the ions to and trapping them in ICR cell, and finally exciting and detecting to produce a mass spectrum. The primary modification made to accommodate an array ICR cell is inserting additional injection events between the first ion injection event and the excitation event.

Spectral quality for both cells was initially optimized using Calmix and the instrument's native digitizer to selectively digitize one channel at a time from either cell. An example of paired spectra is shown in Figure 4.1, in which a full scan MS1 ion packet is trapped in the back cell of the ICR cell, and the ion 1422 is isolated but not fragmented, and ejected from the ion trap and trapped in the front cell. While these detection events are not parallel, they are both recorded with the corresponding ion packet in the other cell. These spectra demonstrate that different ion

populations have been injected and trapped in each cell, and the cross-talk from the MS1 cell to the MS2 is minimal.

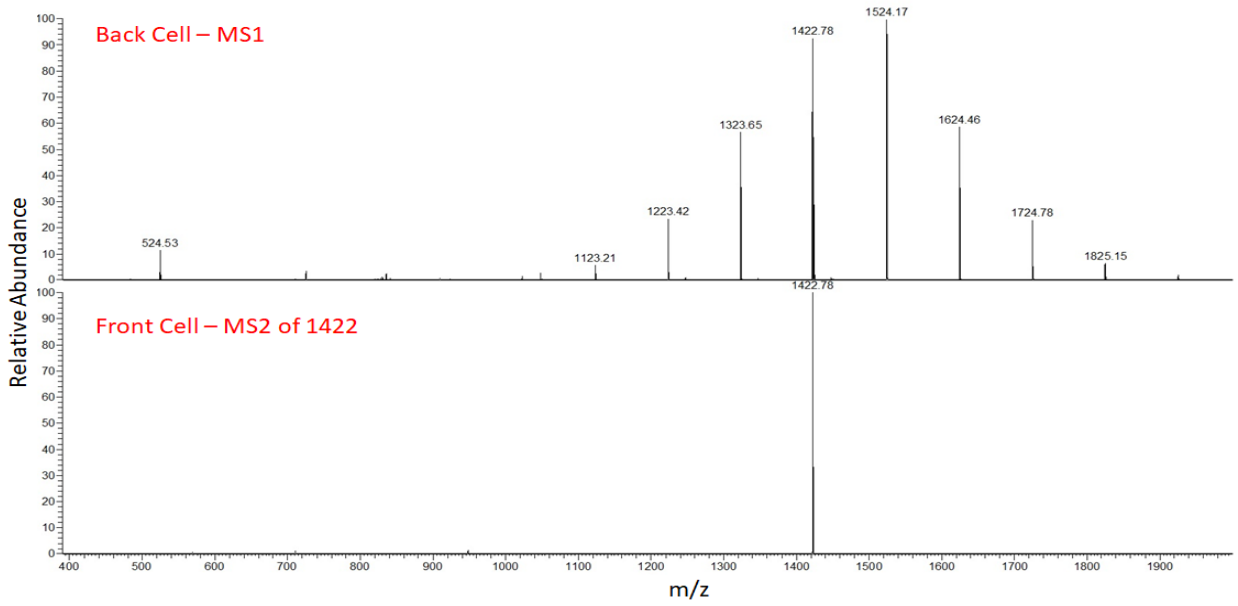


Figure 4.1. Example of spectra resulting from serial injection of different ion packets to each cell. An MS1 ion packet of m/z 400-2000 is injected and trapped into the back cell (Top), and the ultramark ion 1422 is isolated and injected into the front cell (Bottom). Both cells are excited and detected simultaneously.

The injection routine was extended to additionally fragment the MS2 ions prior to ejection, as would be done in a traditional proteomics experiment for the analysis of peptides. Initially, the resulting fragmentation spectra had extremely long ion accumulation times for relatively abundant analytes, and poor signal to noise for many fragment ions (Figure 4.2). We identified this as a problem with the MS2 isolation event which applied inappropriate isolation waveforms for the selected precursor ion of interest, as the broad band MS1 isolation waveform from the immediately

preceding injection was applied instead. This is corrected by regenerating the isolation waveform after the MS2 parameters were set, causing an appropriate waveform to be applied, which yielded a spectrum with more than a 6-fold increase in intensity, with a 200-fold decrease in accumulation time (Figure 4.2).

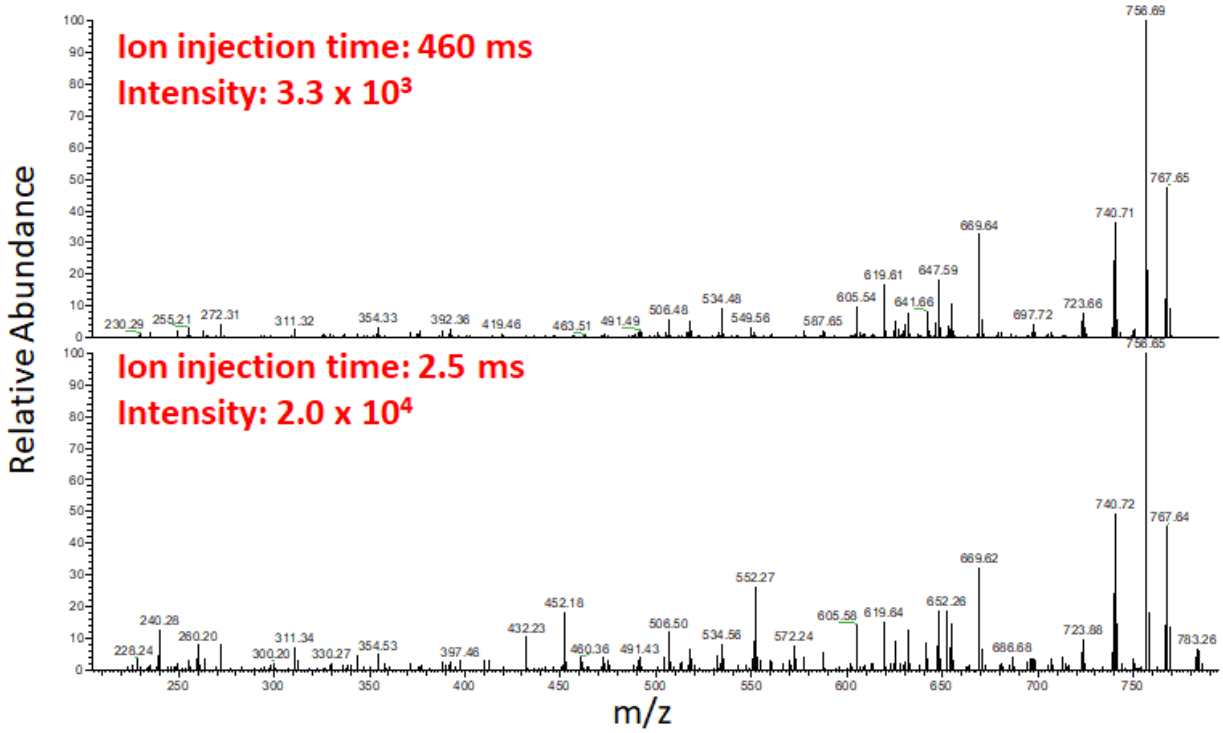


Figure 4.2. Example spectra resulting from the fragmentation of Neurotensin (m/z = 784, z=2+) from the dedicated MS2 cell. In the top spectrum, an improper isolation waveform is applied, resulting in an extremely long injection period with poor intensity for low m/z fragments, while in the bottom spectrum the correct waveform is calculated between the two injections, resulting in improve spectral quality and intensity with a shorter injection period.

Analog signal processing

The primary goals in designing an analog signal processing pipeline were to be able to record one to four mass spectra in parallel and do so with signal quality equivalent to the instrument's native digitizer. As feedback and communication between the instruments control software and the external digitizer were paramount to the operation of the array cell, it was necessary that the resulting spectra from the same input signal were nearly identical. A mirror plot is shown in Figure 4.3, demonstrating the extremely high degree of agreement between the

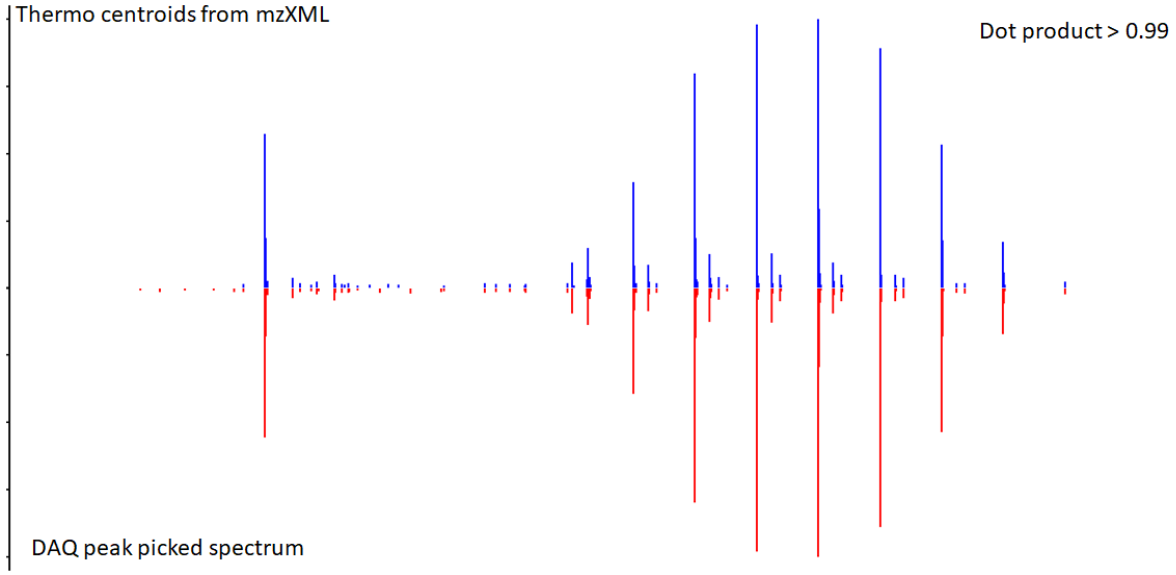


Figure 4.3. Mirror plot showing the agreement between the stock instrument acquisition system (Top) and the external multi-channel digitizer and signal processing pipeline (Bottom).

two processing pipelines on split analog signal generated from Calmix. All major peaks and their intensities are recapitulated in the spectrum produced by the external DAQ. There are several peak centroids represented in only one of the two spectra, but these centroids are all less than 2% of the base peak and likely result from differences in thresholding for peak calling. It is also possible that

they are noise peaks that are unique to the physical connections used to connect the external amplifier to each digitizer.

Automated filling of the MS2 cell

To enable the use of an array cell in a DDA LC-MS proteomics experiment, it is necessary that the MS2 injections can be selected dynamically during the experiment based on the most recent MS1 spectrum. As the instrument control software records and processes the MS1

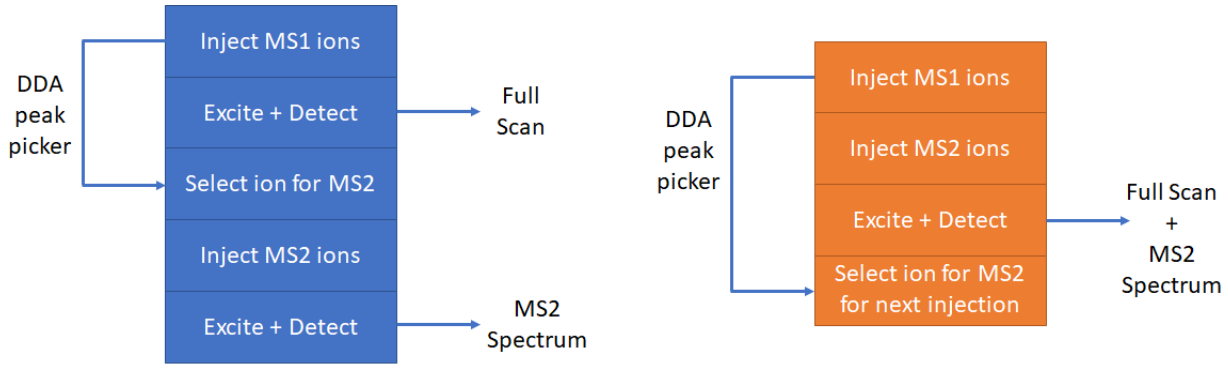


Figure 4.4. Schematic showing the differences between the stock DDA implementation on the instrument (left) versus the new implementation designed (right) to support serial injection with the DDA peak picker during LC-MS acquisition.

spectrum as that cell is split to the native digitizer and the external DAQ, the DDA peak picker can be used largely as normal, but with some modifications. In the simplest DDA experiment, a top 1 method, an MS1 ion packet is accumulated, transferred to the ICR cell, excited and detected, and a mass spectrum is produced. This spectrum is analyzed to select the most intense peak not currently in dynamic exclusion, and that ion is then isolated, fragmented, and subjected to an excitation and detection event to produce an MS2 spectrum (Figure 4.4). We condense this process and call the DDA peak picker whenever an MS1 ion packet is injected to the ICR cell. This uses the peak list generated by the previous MS1 scan event rather than the one just injected to select a

precursor for the MS2 injection, which then occurs immediately (Figure 4.4). While not a major modification, this enables a normal DDA workflow, scalable to any number of cells in an array based off how many precursors are requested from the DDA peak picker, without having to encode or schedule any dependent events. The caveat to this approach is that the very first injection in a run will use whatever the last MS2 mass injected by the instrument as it hasn't seen an MS1 spectrum yet, but this is a minor cost in an LC-MS run where tens of thousands of spectra are recorded in one run.

Analysis of cross-linked BSA

The main objective of an ICR array is to produce an instrument with a fast duty cycle capable of producing high-resolution spectra. Analysis of cross-linked samples is a particularly appealing application of an array cell, due to the large number of unique analytes generated by cross-linking samples and the requirement of high-resolution MS1 and MS2 spectra[106]. These high-resolution spectra are required both to resolve the potentially high-charge fragments resulting from the high charge (4+) precursors, but also to obtain high mass accuracy peaks to solve a mass relationship. This creates a situation in which the samples are not only more complex than traditional linear peptide samples, but also require higher resolution, leading to a slowed duty cycle that leaves many precursors unsampled.

For an initial test, we analyzed cross-linked BSA, which has become a staple standard protein for cross-linking workflows[85]. We compared these results against the same instrument equipped with the native Ultracell mass analyzer running its stock injection code operated using ReACT analyzing the same sample. From this test we did not identify as many unique cross-linked

peptide pairs as the stock set-up (Figure 4.5), despite our greatly enhanced sampling rate (Figure 4.5). Overall, the array cell produces more than twice as many total FT spectra, but produces a small number of MS3 spectra, thus resulting in a fraction of the cross-linked peptide identifications. There are a variety of reasons contributing to this performance gap that are currently under investigation. Perhaps the greatest contributing factor currently is the sensitivity of the PCB array cells is lower than the Ultracell, causing many precursors to not ever rise above the noise, and thus they cannot be sampled or identified.

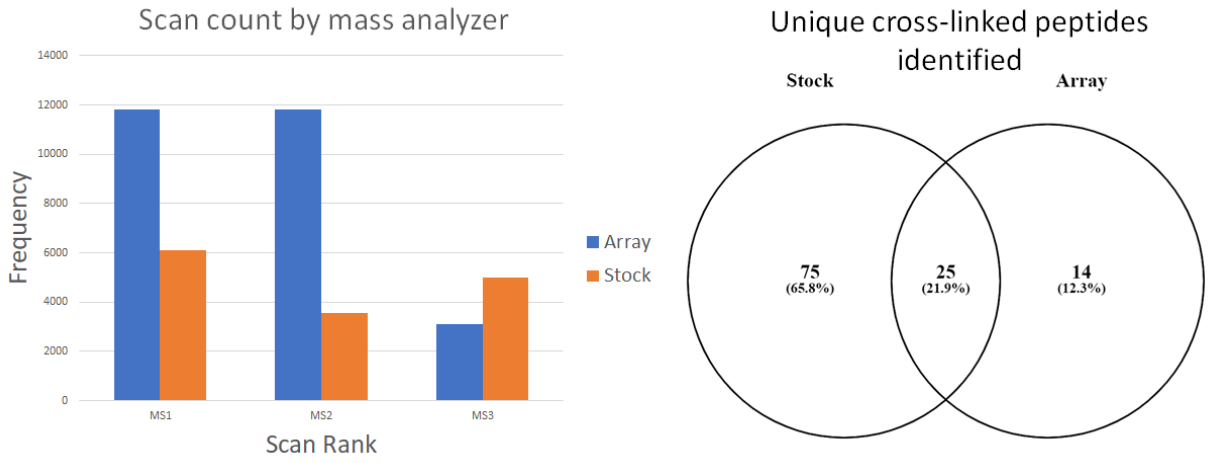


Figure 4.5. Summary of the LC-MS analysis of cross-linked BSA on a developmental FT-ICR-MS equipped with an array ICR cell or the stock Ultracell ICR cell. The array cell produces more than twice as many FT spectra as compared to the stock ICR cell, and produces an equal number of MS1 and MS2 spectra (Left). Overall, the array cell triggers ReACT less frequently, resulting in a small number of MS3s and subsequent cross-linked peptide identifications (Right).

4.5 CONCLUSIONS

Attaining high mass accuracy requires high mass resolving power which in turn requires long data acquisition periods that reduce the duty cycle of a mass spectrometer. In experiments

where a complex mixture is being analyzed, a high duty cycle is required to get representative sampling of precursors, but it is generally always desirable to have a fast duty cycle. In other experiments, such as the analysis of cross-linked peptides, high mass accuracy is required to properly characterize the analytes under investigation. ICR cells are a promising platform for fast duty cycle, high-resolution spectral analyses due to various aspects that can be scaled multiplicatively. While higher magnetic fields and harmonic cells have been investigated in experiments with separations, array ICR cells have so far only been used in direct infusion experiments due to a lack of support for serial ion injection and high-throughput multichannel digitizing.

Here we have demonstrated several significant technological advances in the development of an FT-ICR array mass spectrometer. In contrast to the previous demonstrations of array cells which used a physical relay switch to suppress the excitation event between injections[102], we have developed a true serial injection pipeline in ITCL, to perform multiple injections of different ion packets for a single acquisition event. Additionally, we have shown that the serial injection pipeline can be used in tandem with the instrument's native DDA peak picker to perform intelligent filling of the second cell with new analytes each scan, enabling its use in LC-MS experiments where the analytes to be investigated are not known *a priori*. Furthermore, to enable the use of an array cell in a traditional experiment where there is only a short time on the order of hundreds of milliseconds between acquisitions, we developed a signal processing pipeline on a multi-channel digitizer to process the high-frequency analog signal simultaneously generated from many cells in parallel as fast as it can be acquired. As part of this pipeline, we also created a two-way communication framework between the instrument control software and the external digitizer,

where the control software can push relevant data to a log file readable by the external DAQ, and the DAQ can generate select ITCL files to affect the instrument operation.

These technological advances taken together allowed us to analyze cross-linked BSA in an LC-MS experiment. Analysis of cross-linked peptides is a strong motivating factor for the development of a high duty cycle instrument, as these samples are extremely complex mixtures of low abundance, high mass, high charge products that generally require analysis at high resolution. Current instruments are only able to barely scrape the surface of characterizing cross-linked peptides generated from complex samples of hundreds of proteins[115]. While we do not identify as many unique cross-linked peptide pairs as the stock set-up, this appears to be due to spectral quality and sensitivity issues from the PCB array cell, which can likely be improved in future hardware iterations of the array cell. We did achieve our goal of producing a high-resolution, high-duty cycle instrument, as evidenced by the significantly greater number of FT spectra acquired compared to the stock cell. Furthermore, we did trigger productive MS3s that led to the identification of cross-linked peptides, indicating that the signal processing pipeline and two-way instrument communications are working as intended during an LC-MS experiment. This is the first demonstration of a complex mixture with LC separations on an array ICR cell, laying the framework to allow for more efficient characterization and evaluation of future generations of array cells, with greater numbers of cells and improved spectral quality and sensitivity.

Chapter 5. CONCLUDING REMARKS

5.1 SUMMARY

Cross-linking mass spectrometry is a continually evolving technique in the field of structural proteomics. New reagents, software, and hardware come together to make each generation of experiments better than those that preceded it. In the above chapters I have described my contributions to cross-linking technologies to help dig into the vast network of the protein interactome.

Mango was one of the earliest tools developed that allowed analysis of chimeric spectra generated from cleavable cross-linkers, although many such tools exist now. The existing search tools[34,36] at the time relied on using cross-linkers that produce prominent doublets, such as DSSO or DSBU, to determine candidate peptides. The fragmentation of PIR cross-linkers infrequently produced such doublets, but ReACT demonstrated that the conservation of mass is sufficient for identifying candidate released peptides. Mango was designed as a lightweight tool to leverage application of a mass relationship to convert the identification of chimeric spectra into a series of paired linear peptide searches that can be handled by any traditional search tools. By separating detection from peptide assignment, we can take advantage of new search tools without having to make any significant changes. Mango shows significant improvement over the MS3 based ReACT within our lab, typically yielding on the order of 50% more unique identified cross-links for the same amount of instrument time. Despite yielding improved results, Mango does comparatively worse on a per spectrum basis, with only a few percent of MS2 spectra with mass relationships

being identified and the overall improvement stemming from a faster duty cycle. This result seems to be echoed within the field, as MS3 based cross-linked identification is readily available on Thermo tribrid instruments, but most publications still make use of MS2 chimeric spectrum identification.

Cross-linker design is still a ripe area of research, owing in part to the simple constraints of what a functional cross-linker requires. Any molecule that can react with two distinct residues on a protein to form a covalent bond can serve as a cross-linking reagent, which yields a large potential design space of chemical probes that is largely unexplored. Peptide synthesis is a powerful technique for rapid prototyping of novel cross-linkers, requiring about one day to produce a testable molecule and being able to take advantage of a variety of chemical building blocks in the form of derivatized amino acids. While I synthesized a variety of new cross-linkers, including an amidated form of BDP-NHP, two different sets of isotopically encoded iqPIR cross-linkers[19,42], and various others, the largest departure from existing cross-linkers is described in chapter 3, the tetrameric cross-linker Bisby. While there has been exploration of reactive groups, enrichable tags, and spacer arm size, I chose to examine the number of reactive groups as a new design direction. This was a particularly exciting direction, as it would allow for direct observation of up to four proteins at an interface, something that would take six binary cross-links and an assumption that all six are contemporaneous, could be shown unambiguously for the first time.

The initial design and synthesis of Bisby using Fmoc-based solid phase synthesis was straightforward by successive coupling of branching lysine to generate 4-branch points upon which to build the rest of the molecule in a manner analogous to BDP-NHP. Identification of tetra-linked species was a more difficult question, as chimeric spectrum identification for binary cross-links is

already inefficient due to splitting fragments across the two constitutive peptides. Naturally, MS3-based identification lends itself well to identification of tetra-linked species because we need only produce sufficient populations of the released peptides to identify candidates without having to balance production of fragment ions for identification. ReACT[33], built in ITCL on a Velos-FT-ICR for sequencing binary cross-links served as an excellent basis for real-time MS2 processing and on-the-fly MS3 scheduling, although required significant development to efficiently solve the n^2 problem presented by the four-term mass-relationship on the LC-MS time scale on the instrument computer. I was able to identify a variety of tetra-linked proteins from intact mitochondria isolated from murine hearts, representing a variety of previously characterized OXPHOS complexes to support that our identifications are biologically relevant. Alphafold2 allowed for an arbitrary number of proteins to be co-folded together, which was a synergistic development as it provided a novel pipeline for rapid modeling and evaluation of three-protein complexes, not previously possible without iterative binary docking procedures. We demonstrated this utility by recapitulating the interface between ATPA, ATPG, and ATPE using a single ternary link. While Bisby does not yet produce the network breadth achievable with binary cross-linkers, it represents an important step in our ability to accurately characterize and identify novel protein complexes, mixtures of heterodimers identified by cross-linking could potentially be identified unambiguously as hetero-trimers and hetero-tetramers.

The final main chapter looks at the development of new hardware tools and the software required to use them on an LC-MS time scale. Cross-linked peptides are hard to identify as require high resolution MS1 and MS2 spectra to make use of a mass relationship, and frequently exist in samples that are orders of magnitude more complex that traditional linear peptide samples due to

the combinatorial space of cross-linkable peptides. This presents a problem that stresses the duty cycle of existing instruments, leaving many precursors unsampled in an experiment. FT-ICR mass spectrometers have a variety of ways that can scale their acquisition rate, making them a promising platform for a high-duty cycle instrument.

Our lab had previously developed array ICR cells and showed that it was possible to inject five distinct ion packets into a 5-cell array, and simultaneously acquire spectra from all five cells for the time cost of one detection event, greatly improving the theoretical duty cycle of the instrument. However, the setup of this experiment was not applicable to LC-MS time scale experiments, as it required prior knowledge of what ions to inject into each cell and made use of a digitizer that required several seconds of processing to re-arm between acquisitions. My primary contribution to our instrument development project was helping to address these concerns. I redeveloped the native ion injection code on the instrument to perform multiple injections between detection events, as well as connecting the DDA peak picker to the injection script to dynamically pick ions from MS1 data to fill the additional cells. The second main development was to develop a signal processing platform on a PCIe multi-channel digitizer with sufficient throughput to match the instrument's duty cycle while producing mass spectra of similar quality to the native digitizer. These two advances in tandem allow for general use of an array ICR cell for LC-MS applications in a manner similar to a traditional ICR cell, allowing for better characterization and faster feedback for future generations of array cells.

The technologies described in this dissertation all improve the ways in which cross-linked samples can be produced or interrogated. Mango enables the identification of chimeric spectra from any cleavable cross-linker, allowing for more cross-links to be identified from samples by

leveraging faster MS2-only duty cycles. Tetralinking allows for more information-dense cross-link identifications, yielding six distance constraints for a fully identified species compared to a binary cross-linker's single distance constraint, and also allows for unambiguous identification of multi-protein interfaces, avoiding problematic inferences about potentially mutually exclusive interactions. The improvements made to an ICR array mass spectrometer enable a greatly enhanced duty cycle to be applied to LC-MS experiments, which may yield improved characterization of complex samples as the mass analyzer design improves. Cross-linking is a complicated problem spanning many experimental dimensions, and here I have described several new developments to help dig deeper into the protein interactome by improving our ability to produce and identify cross-linked peptides.

5.2 Looking Forward

Cross-linking experiments offer an important new dimension to the field of proteomics as they enable us to make measurements of protein conformers and complexes. Proteomics has largely been focused on improving the throughput and accuracy of quantitation in complex samples to gain biological insight by measuring changes in expression between experimental conditions. It is important to keep in mind that protein abundances tell only a part of the biological story, and it is possible to have changes in conformers and complexes without any associated change in protein abundance. The developments over the past decade, including instruments, search tools, new cross-linkers, and accompanying commercial adoption from vendors have helped bring cross-linking mass spectrometry to a broader audience in the

proteomics community. Naturally, this is an exciting time in the field as broader adoption brings new applications and expertise.

5.2.1 *Applications*

Cross-linking shines as a relatively simple method to obtain low-resolution structural information about some of the proteins in a sample. While this usage has its own benefits, it becomes more powerful when combined with high-resolution structural techniques. Cryo-electron microscopy (Cryo-EM) combined with cross-linking experiments are an excellent example of the value of low resolution structural constraints coupled with high-resolution structures, and cross-linking has been used to help assemble density maps for a variety of cryo-EM structures as summarized in a recent review[116]. These types of applications I foresee becoming more commonplace as cross-linking methods become routine experiments that can be performed in tandem with a core facility.

Beyond the interplay with current protein structural determination techniques, I think the most exciting developments for cross-linking in the near term will be centered around combining cross-linking with improved protein structure prediction. Currently, the primary drawback of cross-linking experiments is that they provide low-resolution distance constraints which can be difficult to interpret. Protein structure prediction tools like I-TASSER[67] or Phyre2[117] and docking tools like PatchDock[118] have been staples of cross-linking analysis pipelines, but fail to produce high-quality models for many targets. Alphafold[5] and RoseTTAFold[119] both bring significant improvement not only to monomer prediction, but also to co-folding of monomers to produce complexes that allow for conformational shifts in individual subunits. As protein monomer and

complex modeling improves, the value cross-linking experiments increases as their primary drawback is mitigated. For example, if it is possible to accurately model a heterodimer complex from sequence alone, then identifying a interlink, which provides the identity of two interacting proteins, provides sufficient information to produce a high-quality model of a complex. Furthermore, heterodimers are not locked to a single conformation, and cross-linking can also help identify cases where an alternate assembly must exist even if a high-confidence model can be produced. These cases can be tackled by modeling with less refinement, which tends to produce more diverse models of lower confidence. This combination of cross-linking and improved protein structure prediction I expect to become routine, particularly as the hardware requirements for modeling are reduced, making an experiment in which one produces a predicted model for each observed link feasible for labs without access to vast computational resources.

Another application that I expect cross-linking to become valuable in is the study of interactions between proteins and small molecules, which has significant relevance in the field of drug discovery. Proteins and protein complexes are prime targets for small molecule therapeutics and there are pipelines for doing high-throughput virtual screenings to identify ligands that could bind to a target protein[120]. While putative ligands can be screened computationally, the validation of binding targets requires some type of structural experiment such as nuclear magnetic resonance (NMR) or X-ray crystallography. These methods both have the drawbacks of requiring a highly purified sample and removing the target protein from its native cellular environment which might remove functional interactors. In many cases, quantitative cross-linking could be used to validate targets by showing that a drug has an impact on protein conformation or that a drug stabilizes or disrupts some target protein complex in live cells.

Beyond the impact cross-linking can have for validating ligand binding, it has further potential for identifying alternative druggable targets. If there are no viable ligands found for a target protein known to be important in some disease, another way to produce the desired therapeutic effect could be by targeting an interacting partner of that protein. Cross-linking provides the means to identify interaction partners of a protein, and then quantify how those interactions change as with the addition of some ligand. This is a particularly compelling future application, as it could offer new paths towards previously undruggable targets.

5.2.2 *Experimental hurdles*

Independent of what new applications arise for cross-linking mass spectrometry, experimental methods will continue to improve. Efficiency in the field has increased significantly over the past twenty years, with datasets improving from dozens of cross-links[121] to hundreds then thousands[122] for a similar amount of instrument time. This improvement in identifying cross-links from complex samples is caused by a combination of improved instrumentation, real-time identification strategies, better sample processing and enrichment, and new cross-linking reagents. At this time, I think the aspect with the most potential for improvement is increasing the fraction of spectra that are identified.

Currently, only a small fraction of spectra are identified in most cross-linking experiments, with runs of tens of thousands of spectra yielding less than a thousand identifications. This inefficiency is identification is not explicitly discussed in the literature, but even with good enrichment of cross-linked peptides, the identification rate seems to max out at approximately a thousand unique links per hour of instrument time for complex samples[123], with

the median being closer to one-hundred unique links per hour of instrument time. This is one of the most pressing experimental issues in the field now, as it severely limits the throughput of discovery cross-linking experiments.

Some fraction of these unidentified spectra results from the scenario in which only one peptide can be assigned with confidence, resulting in a failed identification. These spectra have the potential to be rescued with a more sophisticated experiment using real-time searches on MS3 enabled instruments. If a confident identification can be obtained in real time for one of the peptides, then the mass of the second peptide is known explicitly and can be targeted for MS3 to potentially recover an identification. Real-time searches can offer significant duty cycle improvements in this experiment, as MS3s are not scheduled when both peptides or neither peptide can already be identified, allowing a duty cycle closer to MS2-only methods to be realized. This triggered MS3 is largely analogous to the real-time search application described as part of Orbiter[93] to improve the duty cycle of experiments that make use of synchronous precursor selection (SPS), and might offer similar duty cycle improvements by reducing the number of unproductive MS3 scans scheduled during an experiment. Improved ease of access to vendor APIs for new instruments make this a technically feasible experiment in an increasing number of labs.

The other major issue that contributes the bulk of unidentified spectra is simply that most spectra contain few to no identifiable fragments from any peptide. This is a problem that can only be solved by generating higher quality spectra, which I think will be a topic of major focus in the coming years. I think a likely explanation for many of these spectra is that cross-linked samples are so complex that it is difficult to produce good spectra for most low abundance

precursors due to many similarly low abundance coeluting species. Offline fractionation is the canonical solution to sample complexity, and there have been many experiments looking at how best to perform offline fractionation for improved enrichment of cross-links and total number of identifications[38,97,124], but none of these approaches significantly increase the throughput of identifying cross-links. Ion mobility seems like a promising direction for producing more high-quality spectra, but its initial applications in cross-linking have yet to show significant gains in complex samples[39,125,126]. Labs continue to make incremental progress on separations and enrichment of cross-linked peptides, but it seems that developing an experiment that produces a high rate of identified spectra will require further technological development.

The final aspect that I think will see significant advances in the coming years will be the development of better cross-linkers. I described in detail earlier about the large chemical design space of cross-linkers, and what small fraction has been explored thus far. The curious part of cross-linker design is that most cross-linkers perform similarly in terms of producing identifications; there is not yet a clear standout reagent in the field. While it is possible that the field has arrived at the set of maximally effective cross-linkers, instead I think that there are dozens of incremental improvements that have been developed across various molecules that need to be brought together to produce the next generation of cross-linking reagents.

For example, the general use cleavable cross-linkers of the current generation of reagents, commonly DSBU or DSSO, share the feature of producing doublets upon activation by virtue of containing two cleavable bonds that cause asymmetric fragmentation. While doublets can be a useful feature for detecting released peptides[34,36], they come at a cost of signal to noise caused by dividing the released peptide signal across two peaks per charge state. Furthermore, this problem

is compounded when identifying released peptides, as fragment ions containing the cross-linker mod site also appear as doublets with reduced signal-to-noise. From a sensitivity standpoint, it seems like the ideal cleavable cross-linker would produce only a single peak for each released peptide to improve signal-to-noise. Unfortunately, this is a difficult design problem as it requires generating a symmetric molecule containing a single CID-cleavable bond at its center.

Beyond fragmentation properties, there are a variety of other desirable characteristics a cleavable cross-linker could contain. For example, it is useful if a cross-linker modification site could retain the positive charge on the lysine side-chain, which has been implemented in the non-cleavable cross-linker, diethyl suberthioimidate (DEST)[127], but no cleavable cross-linkers to date. Cleavable cross-linkers rely on generating fragments from released peptides, which are more likely to fragment if they can stabilize more charges by retaining a basic site at the lysine side-chain. Additionally, most cross-linkers currently make use of activated esters for their reactive groups, which have the downside of readily reacting with water. It is possible to design a cross-linker that leverages the reactivity difference between primary amines and water to generate a cross-linker that has a long half-life in water to better form cross-links in live cells, similar to NNP9 with its NHS-carbamate groups[128], but few cross-linkers make use of this reactivity difference. None of these are trivial modifications, and any new designs for *in vivo* applications must be evaluated for solubility and their ability to penetrate cell membranes, further increasing the difficulty of designing new cross-linkers. These design problems are not inherently unsurmountable, but they will likely require more organic synthesis capabilities than are possessed by the typical cross-linking or proteomics lab. However, the proliferation of cross-linking methods and increased industry involvement in supporting commercial cross-linkers

offers some hope that better reagents may be coming in the near term by bringing more synthetic expertise to the space of cross-linker design.

5.3 Final Remarks

Proteins carry out their functions by means of physical interactions. They might interact with a metabolite, another protein, DNA, or anything in between, but most proteins function by reaching out and touching another molecule. Proteins derive their functional diversity from their structure, which can be changed by post-translational modifications, allosteric effects, or through complex formation. *In vivo* cross-linking is a powerful tool because it enables mass spectrometers to make measurements relating to protein structures as they existed in a cell. Cross-linking is a generally difficult problem, but it is one worth pursuing as it provides a unique avenue to survey protein structures in living cells, where proteins are carrying out whatever suite of functions are required to keep a cell functioning normally with whatever local environment they require to function.

BIBLIOGRAPHY

(1) Hegyi, H.; Gerstein, M. The Relationship between Protein Structure and Function: A Comprehensive Survey with Application to the Yeast Genome 11Edited by G. von Heijne. *J. Mol. Biol.* **1999**, *288* (1), 147–164.

(2) Huttlin, E. L.; Bruckner, R. J.; Navarrete-Perea, J.; Cannon, J. R.; Baltier, K.; Gebreab, F.; Gygi, M. P.; Thornock, A.; Zarraga, G.; Tam, S.; et al. Dual Proteome-Scale Networks Reveal Cell-Specific Remodeling of the Human Interactome. *Cell* **2021**, *184* (11), 3022-3040.e28.

(3) Sinz, A. Cross-linking/Mass Spectrometry for Studying Protein Structures and Protein–Protein Interactions: Where Are We Now and Where Should We Go from Here? *Angew. Chemie Int. Ed.* **2018**, *57* (22), 6390–6396.

(4) Singh, P.; Panchaud, A.; Goodlett, D. R. Chemical Cross-Linking and Mass Spectrometry as a Low-Resolution Protein Structure Determination Technique. *Anal. Chem.* **2010**, *82* (7), 2636–2642.

(5) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589.

(6) del Alamo, D.; Sala, D.; Mchaourab, H. S.; Meiler, J. Sampling Alternative Conformational States of Transporters and Receptors with AlphaFold2. *Elife* **2022**, *11*, e75751.

(7) Kao, A.; Chiu, C.; Vellucci, D.; Yang, Y.; Patel, V. R.; Guan, S.; Randall, A.; Baldi, P.; Rychnovsky, S. D.; Huang, L. Development of a Novel Cross-Linking Strategy for Fast and Accurate Identification of Cross-Linked Peptides of Protein Complexes. *Mol. Cell. Proteomics* **2011**, *10* (1), M110.002212.

(8) Müller, M. Q.; Dreiocker, F.; Ihling, C. H.; Schäfer, M.; Sinz, A. Cleavable Cross-Linker for Protein Structure Analysis: Reliable Identification of Cross-Linking Products by Tandem MS. *Anal. Chem.* **2010**, *82* (16), 6958–6968.

(9) Staros, J. V. N-Hydroxysulfosuccinimide Active Esters: Bis(N-Hydroxysulfosuccinimide) Esters of Two Dicarboxylic Acids Are Hydrophilic, Membrane-Impermeant, Protein Cross-Linkers. *Biochemistry* **1982**, *21* (17), 3950–3955.

(10) Tang, X.; Munske, G. R.; Siems, W. F.; Bruce, J. E. Mass Spectrometry Identifiable Cross-Linking Strategy for Studying Protein-Protein Interactions. *Anal. Chem.* **2005**, *77* (1), 311–318.

(11) Mintseris, J.; Gygi, S. P. High-Density Chemical Cross-Linking for Modeling Protein

Interactions. *Proc. Natl. Acad. Sci.* **2020**, *117* (1), 93 LP – 102.

(12) Leitner, A.; Joachimiak, L. A.; Unverdorben, P.; Walzthoeni, T.; Frydman, J.; Förster, F.; Aebersold, R. Chemical Cross-Linking/Mass Spectrometry Targeting Acidic Residues in Proteins and Protein Complexes. *Proc. Natl. Acad. Sci.* **2014**, *111* (26), 9455–9460.

(13) Giron-Monzon, L.; Manelyte, L.; Ahrends, R.; Kirsch, D.; Spengler, B.; Friedhoff, P. Mapping Protein-Protein Interactions between MutL and MutH by Cross-Linking. *J. Biol. Chem.* **2004**, *279* (47), 49338–49345.

(14) Iacobucci, C.; Götze, M.; Piotrowski, C.; Arlt, C.; Rehkamp, A.; Ihling, C.; Hage, C.; Sinz, A. Carboxyl-Photo-Reactive MS-Cleavable Cross-Linkers: Unveiling a Hidden Aspect of Diazirine-Based Reagents. *Anal. Chem.* **2018**, *90* (4), 2805–2809.

(15) Hage, C.; Iacobucci, C.; Rehkamp, A.; Arlt, C.; Sinz, A. The First "Zero-Length" Mass Spectrometry-Cleavable Cross-Linker for Protein Structure Analysis. *Angew. Chemie Int. Ed.* **2017**, 1–6.

(16) Kalkhof, S.; Ihling, C.; Mechtler, K.; Sinz, A. Chemical Cross-Linking and High-Performance Fourier Transform Ion Cyclotron Resonance Mass Spectrometry for Protein Interaction Analysis: Application to a Calmodulin/Target Peptide Complex. *Anal. Chem.* **2005**, *77* (2), 495–503.

(17) Keller, A.; Chavez, J. D.; Felt, K. C.; Bruce, J. E. Prediction of an Upper Limit for the Fraction of Interprotein Cross-Links in Large-Scale In Vivo Cross-Linking Studies. *J. Proteome Res.* **2019**, *18* (8), 3077–3085.

(18) Tang, X.; Bruce, J. E. A New Cross-Linking Strategy: Protein Interaction Reporter (PIR) Technology for Protein-Protein Interaction Studies. *Mol. Biosyst.* **2010**, *6* (6), 939–947.

(19) Chavez, J. D.; Keller, A.; Mohr, J. P.; Bruce, J. E. Isobaric Quantitative Protein Interaction Reporter Technology for Comparative Interactome Studies. *Anal. Chem.* **2020**, *92* (20), 14094–14102.

(20) Kao, A.; Chiu, C.; Vellucci, D.; Yang, Y.; Patel, V. R.; Guan, S.; Randall, A.; Baldi, P.; Rychnovsky, S. D.; Huang, L. Development of a Novel Cross-Linking Strategy for Fast and Accurate Identification of Cross-Linked Peptides of Protein Complexes. *Mol. Cell. Proteomics* **2011**, *10* (1).

(21) Chavez, J. D.; Mohr, J. P.; Mathay, M.; Zhong, X.; Keller, A.; Bruce, J. E. Systems Structural Biology Measurements by in Vivo Cross-Linking with Mass Spectrometry. *Nat. Protoc.* **2019**, *14* (8).

(22) Steigenberger, B.; Pieters, R. J.; Heck, A. J. R.; Scheltema, R. A. PhoX: An IMAC-Enrichable Cross-Linking Reagent. *ACS Cent. Sci.* **2019**, *5* (9), 1514–1522.

(23) Yılmaz, Ş.; Shiferaw, G. A.; Rayo, J.; Economou, A.; Martens, L.; Vandermarliere, E. Cross-Linked Peptide Identification: A Computational Forest of Algorithms. *Mass Spectrom. Rev.* **2018**, *37* (6), 738–749.

(24) Fasold, H.; Klappenberger, J.; Meyer, C.; Remold, H. Bifunctional Reagents for the Crosslinking of Proteins. *Angew. Chemie Int. Ed. English* **1971**, *10* (11), 795–801.

(25) Chen, Z.-L.; Meng, J.-M.; Cao, Y.; Yin, J.-L.; Fang, R.-Q.; Fan, S.-B.; Liu, C.; Zeng, W.-F.; Ding, Y.-H.; Tan, D. A High-Speed Search Engine PLink 2 with Systematic Evaluation for Proteome-Scale Identification of Cross-Linked Peptides. *Nat. Commun.* **2019**, *10* (1), 1–12.

(26) Kong, A. T.; Leprevost, F. V; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry–Based Proteomics. *Nat. Methods* **2017**, *14* (5), 513–520.

(27) Götze, M.; Pettelkau, J.; Schaks, S.; Bosse, K.; Ihling, C. H.; Krauth, F.; Fritzsche, R.; Kühn, U.; Sinz, A. StavroX—a Software for Analyzing Crosslinked Products in Protein Interaction Studies. *J. Am. Soc. Mass Spectrom.* **2011**, *23* (1), 76–87.

(28) Hoopmann, M. R.; Zelter, A.; Johnson, R. S.; Riffle, M.; Maccoss, M. J.; Davis, T. N.; Moritz, R. L. Kojak: Efficient Analysis of Chemically Cross-Linked Protein Complexes HHS Public Access. *J Proteome Res. J Proteome Res. May* **2015**, *1* (145), 2190–2198.

(29) McIlwain, S.; Draghicescu, P.; Singh, P.; Goodlett, D. R.; Noble, W. S. Detecting Cross-Linked Peptides by Searching against a Database of Cross-Linked Peptide Pairs. *J. Proteome Res.* **2010**, *9* (5), 2488–2495.

(30) Yang, B.; Wu, Y.-J.; Zhu, M.; Fan, S.-B.; Lin, J.; Zhang, K.; Li, S.; Chi, H.; Li, Y.-X.; Chen, H.-F.; et al. Identification of Cross-Linked Peptides from Complex Samples. *Nat. Methods* **2012**, *9* (9), 904–906.

(31) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-Supervised Learning for Peptide Identification from Shotgun Proteomics Datasets. *Nat. Methods* **2007**, *4* (11), 923–925.

(32) Mohr, J. P.; Perumalla, P.; Chavez, J. D.; Eng, J. K.; Bruce, J. E. Mango: A General Tool for CID-Cleavable Cross-Linked Peptide Identification. *Anal. Chem.* **2018**, acs.analchem.7b04991.

(33) Weisbrod, C. R.; Chavez, J. D.; Eng, J. K.; Yang, L.; Zheng, C.; Bruce, J. E. In Vivo Protein Interaction Network Identified with a Novel Real- Time Cross-Linked Peptide Identification Strategy. *J. Proteome Res.* **2013**, *12* (4), 1569–1579.

(34) Liu, F.; Rijkers, D. T. S.; Post, H.; Heck, A. J. R. Proteome-Wide Profiling of Protein Assemblies by Cross-Linking Mass Spectrometry. *Nat. Methods* **2015**, *12* (12), 1179–1184.

(35) Liu, F.; Lössl, P.; Scheltema, R.; Viner, R.; Heck, A. J. R. Optimized Fragmentation Schemes and Data Analysis Strategies for Proteome-Wide Cross-Link Identification. *Nat. Commun.* **2017**, *8* (May), 15473.

(36) Götze, M.; Pettelkau, J.; Fritzsche, R.; Ihling, C. H.; Schäfer, M.; Sinz, A. Automated

Assignment of MS/MS Cleavable Cross-Links in Protein 3d-Structure Analysis. *J. Am. Soc. Mass Spectrom.* **2014**, *26* (1), 83–97.

(37) Pirklbauer, G. J.; Stieger, C. E.; Matzinger, M.; Winkler, S.; Mechtler, K.; Dorfer, V. MS Annika: A New Cross-Linking Search Engine. *J. Proteome Res.* **2021**, *20* (5), 2560–2569.

(38) Yugandhar, K.; Wang, T.-Y.; Leung, A. K.-Y.; Lanz, M. C.; Motorykin, I.; Liang, J.; Shayhidin, E. E.; Smolka, M. B.; Zhang, S.; Yu, H. MaXLinker: Proteome-Wide Cross-Link Identifications with High Specificity and Sensitivity. *Mol. Cell. Proteomics* **2020**, *19* (3), 554–568.

(39) Yılmaz, S.; Busch, F.; Nagaraj, N.; Cox, J. Accurate and Automated High-Coverage Identification of Chemically Cross-Linked Peptides with MaxLynx. *Anal. Chem.* **2022**, *94* (3), 1608–1617.

(40) Beveridge, R.; Stadlmann, J.; Penninger, J. M.; Mechtler, K. A Synthetic Peptide Library for Benchmarking Crosslinking-Mass Spectrometry Search Engines for Proteins and Protein Complexes. *Nat. Commun.* **2020**, *11* (1), 1–9.

(41) Elias, J. E.; Gygi, S. P. Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. *Nat. Methods* **2007**, *4* (3), 207–214.

(42) Chavez, J. D.; Keller, A.; Wippel, H. H.; Mohr, J. P.; Bruce, J. E. Multiplexed Cross-Linking with Isobaric Quantitative Protein Interaction Reporter Technology. *Anal. Chem.* **2021**, *93* (50), 16759–16768.

(43) Trnka, M. J.; Baker, P. R.; Robinson, P. J. J.; Burlingame, A. L.; Chalkley, R. J. Matching Cross-Linked Peptide Spectra: Only as Good as the Worse Identification. *Mol. Cell. Proteomics* **2014**, *13* (2), 420–434.

(44) Keller, A.; Chavez, J. D.; Bruce, J. E.; Kelso, J. Increased Sensitivity with Automated Validation of XL-MS Cleavable Peptide Crosslinks. *Bioinformatics* **2018**, No. August, 1–3.

(45) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem.* **2002**, *74* (20), 5383–5392.

(46) Fischer, L.; Rappsilber, J. Quirks of Error Estimation in Cross-Linking/Mass Spectrometry. *Anal. Chem.* **2017**, *89* (7), 3829–3833.

(47) Schweppe, D. K.; Harding, C.; Chavez, J. D.; Wu, X.; Ramage, E.; Singh, P. K.; Manoil, C.; Bruce, J. E. Host-Microbe Protein Interactions during Bacterial Infection. *Chem. Biol.* **2015**, *22* (11), 1521–1530.

(48) Deblasio, S. L.; Chavez, J. D.; Alexander, M. M.; Ramsey, J.; Eng, J. K.; Mahoney, J.; Gray, S. M.; Bruce, J. E.; Cilia, M. Visualization of Host-Polerovirus Interaction Topologies Using Protein Interaction Reporter Technology. *J. Virol.* **2015**, *90* (December), JVI.01706-15.

(49) Alexander, M. M.; Mohr, J. P.; DeBlasio, S. L.; Chavez, J. D.; Ziegler-Graff, V.; Brault, V.; Bruce, J. E.; Heck, M. C. Insights in Luteovirid Structural Biology Guided by Chemical Cross-Linking and High Resolution Mass Spectrometry. *Virus Res.* **2017**, *241*.

(50) Chavez, J. D.; Tang, X.; Campbell, M. D.; Reyes, G.; Kramer, P. A.; Stuppard, R.; Keller, A.; Zhang, H.; Rabinovitch, P. S.; Marcinek, D. J. Mitochondrial Protein Interaction Landscape of SS-31. *Proc. Natl. Acad. Sci.* **2020**, *117* (26), 15363–15373.

(51) Tüting, C.; Iacobucci, C.; Ihling, C. H.; Kastritis, P. L.; Sinz, A. Structural Analysis of 70S Ribosomes by Cross-Linking/Mass Spectrometry Reveals Conformational Plasticity. *Sci. Rep.* **2020**, *10* (1), 12618.

(52) Ruwolt, M.; Schnirch, L.; Borges Lima, D.; Nadler-Holly, M.; Viner, R.; Liu, F. Optimized TMT-Based Quantitative Cross-Linking Mass Spectrometry Strategy for Large-Scale Interactomic Studies. *Anal. Chem.* **2022**.

(53) Yu, C.; Huszagh, A.; Viner, R.; Novitsky, E. J.; Rychnovsky, S. D.; Huang, L. Developing a Multiplexed Quantitative Cross-Linking Mass Spectrometry Platform for Comparative Structural Analysis of Protein Complexes. *Anal. Chem.* **2016**, *88* (20), 10301–10308.

(54) Chavez, J. D.; Keller, A.; Zhou, B.; Tian, R.; Bruce, J. E. Cellular Interactome Dynamics during Paclitaxel Treatment. *Cell Rep.* **2019**, *29* (8), 2371–2383.

(55) Caudal, A.; Tang, X.; Chavez, J. D.; Keller, A.; Villet, O.; Zhou, B.; Walker, M. A.; Tian, R.; Bruce, J. E. Mitochondrial Interactome Quantitation Reveals Structural Changes in Metabolic Machinery in Failing Murine Heart. *bioRxiv* **2021**, 2021.08.13.456027.

(56) Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J.; et al. Protein Complex Prediction with AlphaFold-Multimer. *bioRxiv* **2021**, 2021.10.04.463034.

(57) R., H. I.; Jimin, P.; Minkyung, B.; Aditya, K.; Ivan, A.; Sergey, O.; Jing, Z.; J., N. T.; Sudeep, B.; R., B. S.; et al. Computed Structures of Core Eukaryotic Protein Complexes. *Science (80-.).* **2022**, *374* (6573), eabm4805.

(58) Burke, D. F.; Bryant, P.; Barrio-Hernandez, I.; Memon, D.; Pozzati, G.; Shenoy, A.; Zhu, W.; Dunham, A. S.; Albanese, P.; Keller, A.; et al. Towards a Structurally Resolved Human Protein Interaction Network. *bioRxiv* **2021**, 2021.11.08.467664.

(59) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: An Open-Source MS/MS Sequence Database Search Tool. *Proteomics* **2013**, *13* (1), 22–24.

(60) Zheng, C.; Yang, L.; Hoopmann, M. R.; Eng, J. K.; Tang, X.; Weisbrod, C. R.; Bruce, J. E. Cross-Linking Measurements of in Vivo Protein Complex Topologies. *Mol Cell Proteomics* **2011**, *10* (10), M110.006841.

(61) Navare, A. T.; Chavez, J. D.; Zheng, C.; Weisbrod, C. R.; Eng, J. K.; Siehnel, R.; Singh, P. K.; Manoil, C.; Bruce, J. E. Probing the Protein Interaction Network of Pseudomonas Aeruginosa Cells by Chemical Cross-Linking Mass Spectrometry. *Structure* **2015**, *23* (4),

762–773.

(62) Wu, X.; Chavez, J. D.; Schweppe, D. K.; Zheng, C.; Weisbrod, C. R.; Eng, J. K.; Murali, A.; Lee, S. A.; Ramage, E.; Gallagher, L. A.; et al. In Vivo Protein Interaction Network Analysis Reveals Porin-Localized Antibiotic Inactivation in Acinetobacter Baumannii Strain AB5075. *Nat. Commun.* **2016**, *7*, 1–14.

(63) Guerrero, C.; Tagwerker, C.; Kaiser, P.; Huang, L. An Integrated Mass Spectrometry-Based Proteomic Approach. *Mol. Cell. Proteomics* **2006**, *5* (2), 366–378.

(64) Chavez, J. D.; Schweppe, D. K.; Eng, J. K.; Zheng, C.; Taipale, A.; Zhang, Y.; Takara, K.; Bruce, J. E. Quantitative Interactome Analysis Reveals a Chemoresistant Edgotype. *Nat. Commun.* **2015**, *6*, 1–12.

(65) Chavez, J. D.; Schweppe, D. K.; Eng, J. K.; Bruce, J. E. In Vivo Conformational Dynamics of Hsp90 and Its Interactors. *Cell Chem. Biol.* **2016**, *23* (6), 716–726.

(66) Leitner, A.; Faini, M.; Stengel, F.; Aebersold, R. Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. *Trends Biochem. Sci.* **2016**, *41* (1), 20–32.

(67) Yang, J.; Zhang, Y. I-TASSER Server: New Development for Protein Structure and Function Predictions. *Nucleic Acids Res.* **2015**, *43* (W1), W174–W181.

(68) Young, M. M.; Tang, N.; Hempel, J. C.; Oshiro, C. M.; Taylor, E. W.; Kuntz, I. D.; Gibson, B. W.; Dollinger, G. High Throughput Protein Fold Identification by Using Experimental Constraints Derived from Intramolecular Cross-Links and Mass Spectrometry. *Proc. Natl. Acad. Sci.* **2000**, *97* (11), 5802–5806.

(69) Rozbesky, D.; Sovova, Z.; Marcoux, J.; Man, P.; Ettrich, R.; Robinson, C. V.; Novak, P. Structural Model of Lymphocyte Receptor NKR-P1C Revealed by Mass Spectrometry and Molecular Modeling. *Anal. Chem.* **2013**, *85* (3), 1597–1604.

(70) Shi, Y.; Fernandez-Martinez, J.; Tjioe, E.; Pellarin, R.; Kim, S. J.; Williams, R.; Schneidman-Duhovny, D.; Sali, A.; Rout, M. P.; Chait, B. T. Structural Characterization by Cross-Linking Reveals the Detailed Architecture of a Coatomer-Related Heptameric Module from the Nuclear Pore Complex. *Mol. Cell. Proteomics* **2014**, *13* (11), 2927–2943.

(71) Erzberger, J. P.; Stengel, F.; Pellarin, R.; Zhang, S.; Schaefer, T.; Aylett, C. H. S.; Cimermančič, P.; Boehringer, D.; Sali, A.; Aebersold, R.; et al. Molecular Architecture of the 40S · EIF1 · EIF3 Translation Initiation Complex. *Cell* **2014**, *158* (5), 1123–1135.

(72) Weisbrod, C. R.; Hoopmann, M. R.; Senko, M. W.; Bruce, J. E. Performance Evaluation of a Dual Linear Ion Trap-Fourier Transform Ion Cyclotron Resonance Mass Spectrometer for Proteomics Research. *J. Proteomics* **2013**, *88*, 109–119.

(73) Zheng, C.; Weisbrod, C. R.; Chavez, J. D.; Eng, J. K.; Sharma, V.; Wu, X.; Bruce, J. E. XLink-DB: Database and Software Tools for Storing and Visualizing Protein Interaction

Topology Data. *J. Proteome Res.* **2013**, *12* (4), 1989–1995.

(74) Schweppe, D. K.; Zheng, C.; Chavez, J. D.; Navare, A. T.; Wu, X.; Eng, J. K.; Bruce, J. E. XLinkDB 2.0: Integrated, Large-Scale Structural Analysis of Protein Crosslinking Data. *Bioinformatics* **2016**, *32* (17), 2716–2718.

(75) Schweppe, D. K.; Chavez, J. D.; Navare, A. T.; Wu, X.; Ruiz, B.; Eng, J. K.; Lam, H.; Bruce, J. E. Spectral Library Searching to Identify Cross-Linked Peptides. *J. Proteome Res.* **2016**, *15* (5), 1725–1731.

(76) Pedrioli, P. G. A.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; et al. A Common Open Representation of Mass Spectrometry Data and Its Application to Proteomics Research. *Nat. Biotechnol.* **2004**, *22* (11), 1459–1466.

(77) Hoopmann, M. R.; Finney, G. L.; Maccoss, M. J. Spectrum Quality Assessment of Shotgun Proteomics Datasets. **2008**, *79* (15), 5620–5632.

(78) McDonald, W. H.; Tabb, D. L.; Sadygov, R. G.; MacCoss, M. J.; Venable, J.; Graumann, J.; Johnson, J. R.; Cociorva, D.; Yates III, J. R. MS1, MS2, and SQT—Three Unified, Compact, and Easily Parsed File Formats for the Storage of Shotgun Proteomic Spectra and Identifications. *Rapid Commun. Mass Spectrom.* **2004**, *18* (18), 2162–2168.

(79) Chavez, J. D.; Weisbrod, C. R.; Zheng, C.; Eng, J. K.; Bruce, J. E. Protein Interactions, Post-Translational Modifications and Topologies in Human Cells. *Mol Cell Proteomics* **2013**, *12* (5), 1451–1467.

(80) Fields, S.; Song, O. A Novel Genetic System to Detect Protein–Protein Interactions. *Nature* **1989**, *340* (6230), 245–246.

(81) Schweppe, D. K.; Chavez, J. D.; Lee, C. F.; Caudal, A.; Kruse, S. E.; Stuppard, R.; Marcinek, D. J.; Shadel, G. S.; Tian, R.; Bruce, J. E. Mitochondrial Protein Interactome Elucidated by Chemical Cross-Linking Mass Spectrometry. *Proc. Natl. Acad. Sci.* **2017**, *114* (7), 1732–1737.

(82) Hevler, J. F.; Lukassen, M. V; Cabrera-Orefice, A.; Arnold, S.; Pronker, M. F.; Franc, V.; Heck, A. J. R. Selective Cross-Linking of Coinciding Protein Assemblies by in-Gel Cross-Linking Mass Spectrometry. *EMBO J.* **2021**, *40* (4), e106174.

(83) Piersimoni, L.; Kastritis, P. L.; Arlt, C.; Sinz, A. Cross-Linking Mass Spectrometry for Investigating Protein Conformations and Protein–Protein Interactions─A Method for All Seasons. *Chem. Rev.* **2021**.

(84) Iacobucci, C.; Piotrowski, C.; Rehkamp, A.; Ihling, C. H.; Sinz, A. The First MS-Cleavable, Photo-Thiol-Reactive Cross-Linker for Protein Structural Studies. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (1), 139–148.

(85) Iacobucci, C.; Piotrowski, C.; Aebersold, R.; Amaral, B. C.; Andrews, P.; Bernfur, K.; Borchers, C.; Brodie, N. I.; Bruce, J. E.; Cao, Y.; et al. First Community-Wide,

Comparative Cross-Linking Mass Spectrometry Study. *Anal. Chem.* **2019**, *91* (11), 6953–6961.

(86) Calvo, S. E.; Clauser, K. R.; Mootha, V. K. MitoCarta2.0: An Updated Inventory of Mammalian Mitochondrial Proteins. *Nucleic Acids Res.* **2016**, *44* (D1), D1251–D1257.

(87) Shteynberg, D.; Deutsch, E. W.; Lam, H.; Eng, J. K.; Sun, Z.; Tasman, N.; Mendoza, L.; Moritz, R. L.; Aebersold, R.; Nesvizhskii, A. I. IProphet: Multi-Level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates. *Mol. Cell. proteomics* **2011**, *10* (12).

(88) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Anal. Chem.* **2003**, *75* (17), 4646–4658.

(89) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold - Making Protein Folding Accessible to All. *bioRxiv* **2021**, 2021.08.15.456425.

(90) Li, S. C.; Ng, Y. K. Calibur: A Tool for Clustering Large Numbers of Protein Decoys. *BMC Bioinformatics* **2010**, *11* (1), 1–12.

(91) Perez-Riverol, Y.; Bai, J.; Bandla, C.; García-Seisdedos, D.; Hewapathirana, S.; Kamatchinathan, S.; Kundu, D. J.; Prakash, A.; Frericks-Zipper, A.; Eisenacher, M.; et al. The PRIDE Database Resources in 2022: A Hub for Mass Spectrometry-Based Proteomics Evidences. *Nucleic Acids Res.* **2022**, *50* (D1), D543–D552.

(92) Yang, L.; Tang, X.; Weisbrod, C. R.; Munske, G. R.; Eng, J. K.; von Haller, P. D.; Kaiser, N. K.; Bruce, J. E. A Photocleavable and Mass Spectrometry Identifiable Cross-Linker for Protein Interaction Studies. *Anal. Chem.* **2010**, *82* (9), 3556–3566.

(93) Schweppe, D. K.; Eng, J. K.; Yu, Q.; Bailey, D.; Rad, R.; Navarrete-Perea, J.; Huttlin, E. L.; Erickson, B. K.; Paulo, J. A.; Gygi, S. P. Full-Featured, Real-Time Database Searching Platform Enables Fast and Accurate Multiplexed Quantitative Proteomics. *J. Proteome Res.* **2020**, *19* (5), 2026–2034.

(94) Lu, L.; Scalf, M.; Shortreed, M. R.; Smith, L. M. Mesh Fragmentation Improves Dissociation Efficiency in Top-down Proteomics. *J. Am. Soc. Mass Spectrom.* **2021**, *32* (6), 1319–1325.

(95) Mohr, J. P.; Perumalla, P.; Chavez, J. D.; Eng, J. K.; Bruce, J. E. *Mango: A General Tool for CID-Cleavable Cross-Linked Peptide Identification*; 2017.

(96) Iacobucci, C.; Sinz, A. To Be or Not to Be? Five Guidelines to Avoid Misassignments in Cross-Linking/Mass Spectrometry. *Anal. Chem.* **2017**, *89* (15), 7832–7835.

(97) Leitner, A.; Reischl, R.; Walzthoeni, T.; Herzog, F.; Bohn, S.; Förster, F.; Aebersold, R. Expanding the Chemical Cross-Linking Toolbox by the Use of Multiple Proteases and Enrichment by Size Exclusion Chromatography. *Mol. Cell. proteomics* **2012**, *11* (3), M111-014126.

(98) Molodenskiy, D.; Shirshin, E.; Tikhonova, T.; Gruzinov, A.; Peters, G.; Spinozzi, F. Thermally Induced Conformational Changes and Protein-Protein Interactions of Bovine Serum Albumin in Aqueous Solution under Different PH and Ionic Strengths as Revealed by SAXS Measurements. *Phys. Chem. Chem. Phys.* **2017**, *19* (26), 17143–17155.

(99) Matzinger, M.; Kandioller, W.; Doppler, P.; Heiss, E. H.; Mechtler, K. Fast and Highly Efficient Affinity Enrichment of Azide-A-DSBSO Cross-Linked Peptides. *J. Proteome Res.* **2020**, *19* (5), 2071–2079.

(100) Han, J.-D. J.; Bertin, N.; Hao, T.; Goldberg, D. S.; Berriz, G. F.; Zhang, L. V; Dupuy, D.; Walhout, A. J. M.; Cusick, M. E.; Roth, F. P.; et al. Evidence for Dynamically Organized Modularity in the Yeast Protein–Protein Interaction Network. *Nature* **2004**, *430* (6995), 88–93.

(101) M., K. P.; J., L. L.; Yu, X.; B., G. M. Relating Three-Dimensional Structures to Protein Networks Provides Evolutionary Insights. *Science (80-.).* **2006**, *314* (5807), 1938–1941.

(102) Park, S.-G.; Anderson, G. A.; Navare, A. T.; Bruce, J. E. Parallel Spectral Acquisition with an Ion Cyclotron Resonance Cell Array. *Anal. Chem.* **2016**, *88* (2), 1162–1168.

(103) Park, S.-G.; Anderson, G. A.; Bruce, J. E. Parallel Spectral Acquisition with Orthogonal ICR Cells. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (3), 515–524.

(104) Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J. Accurate and Sensitive Peptide Identification with Mascot Percolator. *J. Proteome Res.* **2009**, *8* (6), 3176–3181.

(105) Kooijman, P. C.; Nagornov, K. O.; Kozhinov, A. N.; Kilgour, D. P. A.; Tsybin, Y. O.; Heeren, R. M. A.; Ellis, S. R. Increased Throughput and Ultra-High Mass Resolution in DESI FT-ICR MS Imaging through New-Generation External Data Acquisition System and Advanced Data Processing Approaches. *Sci. Rep.* **2019**, *9* (1), 8.

(106) Chavez, J. D.; Park, S.-G.; Mohr, J. P.; Bruce, J. E. Applications and Advancements of FT-ICR-MS for Interactome Studies. *Mass Spectrom. Rev.* **2022**, *41* (2), 248–261.

(107) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. Fourier Transform Ion Cyclotron Resonance Mass Spectrometry: A Primer. *Mass Spectrom. Rev.* **1998**, *17* (1), 1–35.

(108) Anderson, L. C.; Dehart, C. J.; Kaiser, N. K.; Fellers, R. T.; Smith, D. F.; Greer, J. B.; Leduc, R. D.; Blakney, G. T.; Thomas, P. M.; Kelleher, N. L.; et al. Identification and Characterization of Human Proteoforms by Top-Down LC-21 Tesla FT-ICR Mass Spectrometry. *J. Proteome Res.* **2017**, *16* (2), 1087–1096.

(109) Grosshans, P. B.; Marshall, A. G. Can Fourier Transform Mass Spectral Resolution Be Improved by Detection at Harmonic Multiples of the Fundamental Ion Cyclotron Orbital Frequency? *Int. J. Mass Spectrom. Ion Process.* **1991**, *107* (1), 49–81.

(110) Nikolaev, E. N.; Gorshkov, M. V; Mordehai, A. V; Talrose, V. L. Ion Cyclotron Resonance Signal-Detection at Multiples of the Cyclotron Frequency. *Rapid Commun. Mass Spectrom.* **1990**, *4* (5), 144–146.

(111) Cho, E.; Witt, M.; Hur, M.; Jung, M.-J.; Kim, S. Application of FT-ICR MS Equipped with Quadrupole Detection for Analysis of Crude Oil. *Anal. Chem.* **2017**, *89* (22), 12101–12107.

(112) Park, S.-G.; Anderson, G. A.; Bruce, J. E. Parallel Detection of Fundamental and Sixth Harmonic Signals Using an ICR Cell with Dipole and Sixth Harmonic Detectors. *J. Am. Soc. Mass Spectrom.* **2020**, *31* (3), 719–726.

(113) Anderson, L. C.; DeHart, C. J.; Kaiser, N. K.; Fellers, R. T.; Smith, D. F.; Greer, J. B.; LeDuc, R. D.; Blakney, G. T.; Thomas, P. M.; Kelleher, N. L.; et al. Identification and Characterization of Human Proteoforms by Top-Down LC-21 Tesla FT-ICR Mass Spectrometry. *J. Proteome Res.* **2017**, *16* (2), 1087–1096.

(114) Zhang, Z.; Marshall, A. G. A Universal Algorithm for Fast and Automated Charge State Deconvolution of Electrospray Mass-to-Charge Ratio Spectra. *J. Am. Soc. Mass Spectrom.* **1998**, *9* (3), 225–233.

(115) Yugandhar, K.; Zhao, Q.; Gupta, S.; Xiong, D.; Yu, H. Progress in Methodologies and Quality-Control Strategies in Protein Cross-Linking Mass Spectrometry. *Proteomics* **2021**, *21* (23–24), 2100145.

(116) Schmidt, C.; Urlaub, H. Combining Cryo-Electron Microscopy (Cryo-EM) and Cross-Linking Mass Spectrometry (CX-MS) for Structural Elucidation of Large Protein Assemblies. *Curr. Opin. Struct. Biol.* **2017**, *46*, 157–168.

(117) Kelley, L. A.; Mezulis, S.; Yates, C. M.; Wass, M. N.; Sternberg, M. J. E. The Phyre2 Web Portal for Protein Modeling, Prediction and Analysis. *Nat. Protoc.* **2015**, *10* (6), 845–858.

(118) Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. PatchDock and SymmDock: Servers for Rigid and Symmetric Docking. *Nucleic Acids Res.* **2005**, *33* (suppl_2), W363–W367.

(119) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science (80-.).* **2021**, *373* (6557), 871–876.

(120) Jacquemard, C.; Kellenberger, E. A Bright Future for Fragment-Based Drug Discovery: What Does It Hold? *Expert Opin. Drug Discov.* **2019**, *14* (5), 413–416.

(121) Rinner, O.; Seebacher, J.; Walzthoeni, T.; Mueller, L. N.; Beck, M.; Schmidt, A.; Mueller, M.; Aebersold, R. Identification of Cross-Linked Peptides from Large Sequence Databases. *Nat. Methods* **2008**, *5* (4), 315–318.

(122) Götze, M.; Iacobucci, C.; Ihling, C. H.; Sinz, A. A Simple Cross-Linking/Mass Spectrometry Workflow for Studying System-Wide Protein Interactions. *Anal. Chem.* **2019**, *91* (15), 10236–10244.

(123) Jiang, P.-L.; Wang, C.; Diehl, A.; Viner, R.; Etienne, C.; Nandhikonda, P.; Foster, L.; Bomgarden, R. D.; Liu, F. A Membrane-Permeable and Immobilized Metal Affinity Chromatography (IMAC) Enrichable Cross-Linking Reagent to Advance In Vivo Cross-Linking Mass Spectrometry. *Angew. Chemie Int. Ed.* **2022**, *61* (12), e202113937.

(124) Jiao, F.; Yu, C.; Wheat, A.; Wang, X.; Rychnovsky, S. D.; Huang, L. Two-Dimensional Fractionation Method for Proteome-Wide Cross-Linking Mass Spectrometry Analysis. *Anal. Chem.* **2022**, *94* (10), 4236–4242.

(125) Steigenberger, B.; van den Toorn, H. W. P.; Bijl, E.; Greisch, J.-F.; Räther, O.; Lubeck, M.; Pieters, R. J.; Heck, A. J. R.; Scheltema, R. A. Benefits of Collisional Cross Section Assisted Precursor Selection (Caps-PASEF) for Cross-Linking Mass Spectrometry. *Mol. Cell. Proteomics* **2020**, *19* (10), 1677–1687.

(126) Ihling, C. H.; Piersimoni, L.; Kipping, M.; Sinz, A. Cross-Linking/Mass Spectrometry Combined with Ion Mobility on a TimsTOF Pro Instrument for Structural Proteomics. *Anal. Chem.* **2021**, *93* (33), 11442–11450.

(127) Lauber, M. A.; Reilly, J. P. Novel Amidinating Cross-Linker for Facilitating Analyses of Protein Structures and Interactions. *Anal. Chem.* **2010**, *82* (18), 7736–7743.

(128) Rey, M.; Dhenin, J.; Kong, Y.; Nouchikian, L.; Filella, I.; Duchateau, M.; Dupré, M.; Pellarin, R.; Duménil, G.; Chamot-Rooke, J. Advanced In Vivo Cross-Linking Mass Spectrometry Platform to Characterize Proteome-Wide Protein Interactions. *Anal. Chem.* **2021**, *93* (9), 4166–4174.

APPENDIX A: SUPPLEMENT FOR CHAPTER 2

Supplemental Methods

Searching Mango output with Msfragger

Mango was run on the F8 SCX fractions with following settings following settings:

mass_tolerance_relationship = 10.00 ppm

mass_tolerance_peptide = 20.00

reporter_neutral_mass = 751.406080 Da

export_mgf=1.

The resulting mgf file was searched in msfragger (build 20170103_v2) using a narrow window search with the following settings:

num_threads = 0

precursor_mass_tolerance = 100

precursor_mass_units = 1

precursor_true_tolerance = 20

precursor_true_units = 1

fragment_mass_tolerance = 20

fragment_mass_units = 1

isotope_error = 0/1/2

search_enzyme_name = Trypsin

search_enzyme_cutafter = KR

search_enzyme_butnotafter = P

num_enzyme_termini = 2

allowed_missed_cleavage = 1

clip_nTerm_M = 1

variable_mod_01 = 15.99490 M

# variable_mod_02 = 42.01060 [*
# variable_mod_03 = 79.96633 STY
# variable_mod_04 = -17.02650 nQnC
# variable_mod_05 = -18.01060 nE
variable_mod_06 = 197.03240 K[
allow_multiple_variable_mods_on_residue = 0
max_variable_mods_per_mod = 3
max_variable_mods_combinations = 5000
output_file_extension = pepXML
output_format = pepXML
output_report_topN = 1
output_max_expect = 50.0
precursor_charge = 0 0
override_charge = 0
digest_min_length = 7
digest_max_length = 50
digest_mass_range = 500.0 5000.0
max_fragment_charge = 2
track_zero_topN = 0
zero_bin_accept_expect = 0
zero_bin_mult_expect = 1
add_topN_complementary = 0
minimum_peaks = 10
use_topN_peaks = 100
min_fragments_modelling = 1
min_matched_fragments = 1
minimum_ratio = 0.01
clear_mz_range = 0.0 0.0
add_Cterm_peptide = 0.000000

add_Nterm_peptide = 0.000000
add_Cterm_protein = 0.000000
add_Nterm_protein = 0.000000
add_G_glycine = 0.000000
add_A_alanine = 0.000000
add_S_serine = 0.000000
add_P_proline = 0.000000
add_V_valine = 0.000000
add_T_threonine = 0.000000
add_C_cysteine = 57.021464
add_L_leucine = 0.000000
add_I_isoleucine = 0.000000
add_N_asparagine = 0.000000
add_D_aspartic_acid = 0.000000
add_Q_glutamine = 0.000000
add_K_lysine = 0.000000
add_E_glutamic_acid = 0.000000
add_M_methionine = 0.000000
add_H_histidine = 0.000000
add_F_phenylalanine = 0.000000
add_R_arginine = 0.000000
add_Y_tyrosine = 0.000000
add_W_tryptophan = 0.000000
add_B_user_amino_acid = 0.000000
add_J_user_amino_acid = 0.000000
add_O_user_amino_acid = 0.000000
add_U_user_amino_acid = 0.000000
add_X_user_amino_acid = 0.000000
add_Z_user_amino_acid = 0.000000

The resulting pep.xml was processed identically to the *E.coli* data as described in the experimental.

Searching DSSO cross-linked data

The raw file (OR8_20140721_FL_HeLa_lysate_DSSO_XL_SCX_f25_rep) for the DSSO data from the original XlinkX manuscript[34] was obtained from Chorus data repository (https://chorusproject.org) with project I.D. number 890. Due to the asymmetric fragmentation of DSSO yielding its characteristic doublets, it can be efficiently analyzed by running Mango twice to target different pairs of corresponding doublets. The Mango parameters used were as follows, using constants derived from the XlinkX software[34]:

Short arm parameters:

mass_tolerance_relationship = 10.00 ppm
mass_tolerance_peptide = 20.00
reporter_neutral_mass = 49.98264 Da
Long arm parameters:
mass_tolerance_relationship = 10.00 ppm
mass_tolerance_peptide = 20.00
reporter_neutral_mass = -13.96152 Da

The resulting ms2 files were then merged together, with precursor lists concatenated if the same scan was appeared in both Mango runs. The concatenated file was then searched twice with Comet against the uniprot human database (09/08/17) with the default settings and the following changes:

Requires short arm modified:
mango_search = 1;
variable modifications: 15.9949 M

required modifications: 54.01056 at an internal K.

Requires long arm modified:

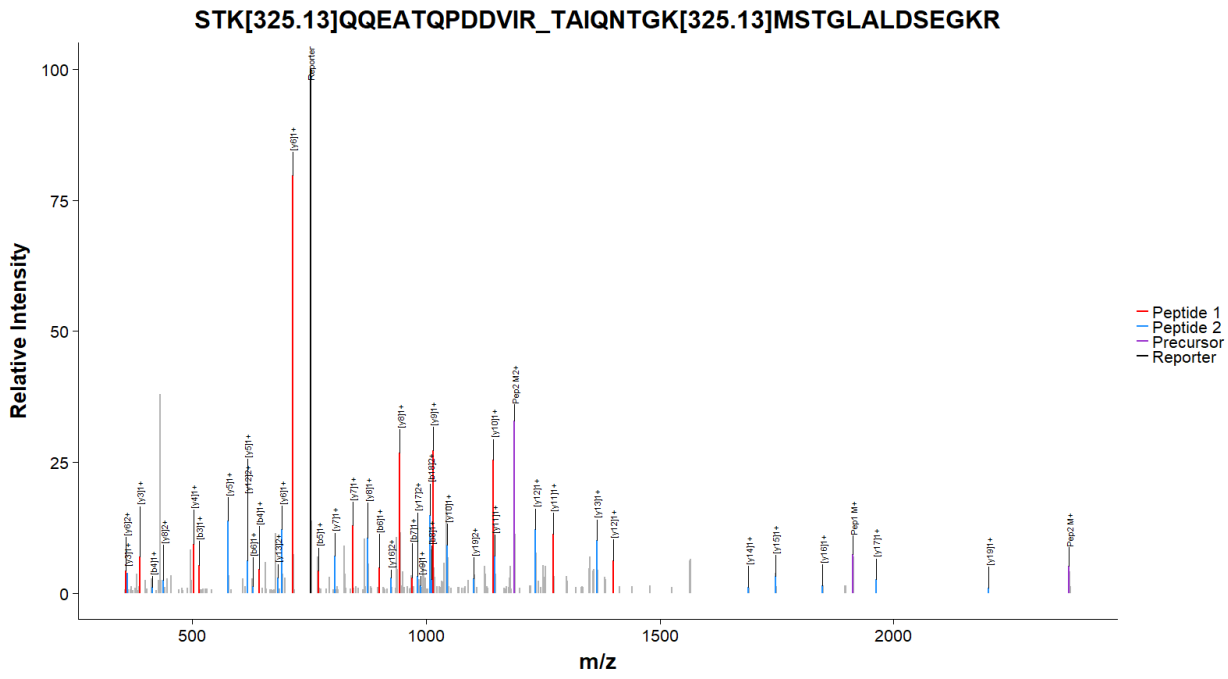
mango_search = 1;

variable modifications: 15.9949 M

required modifications: 85.94264 at an internal K.

The outputs from these two searches were concatenated and processed identically to the *E.coli* data as described in the experimental.

Supplemental Figures



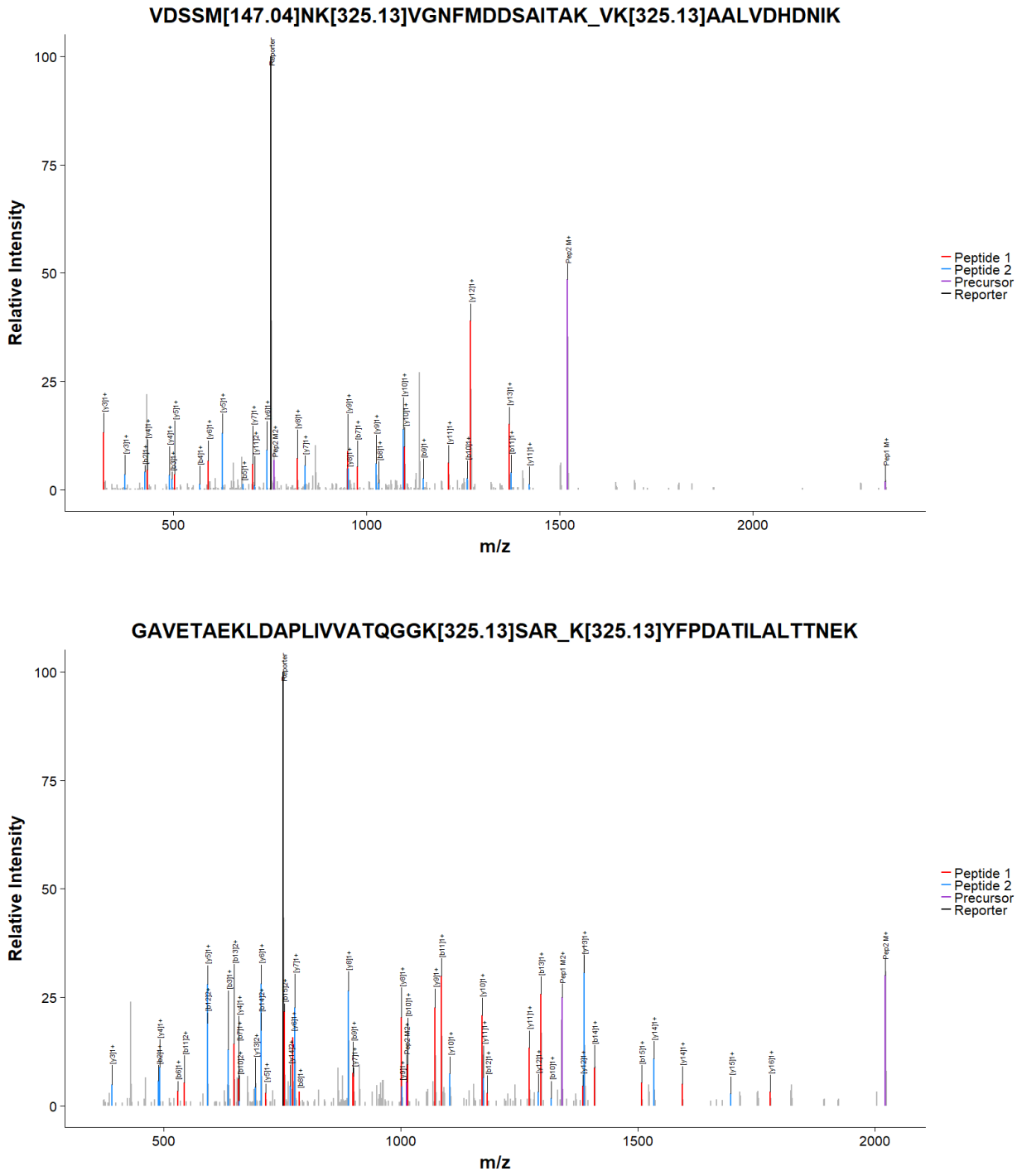


Figure A.5.1. Several examples of MS2 spectra annotated with their assigned cross-link, as well as the ions that could be used in scoring by Comet. Note that only linear peptide ions are scored,

no ions containing fragments of both peptides are used in scoring during the Comet search.

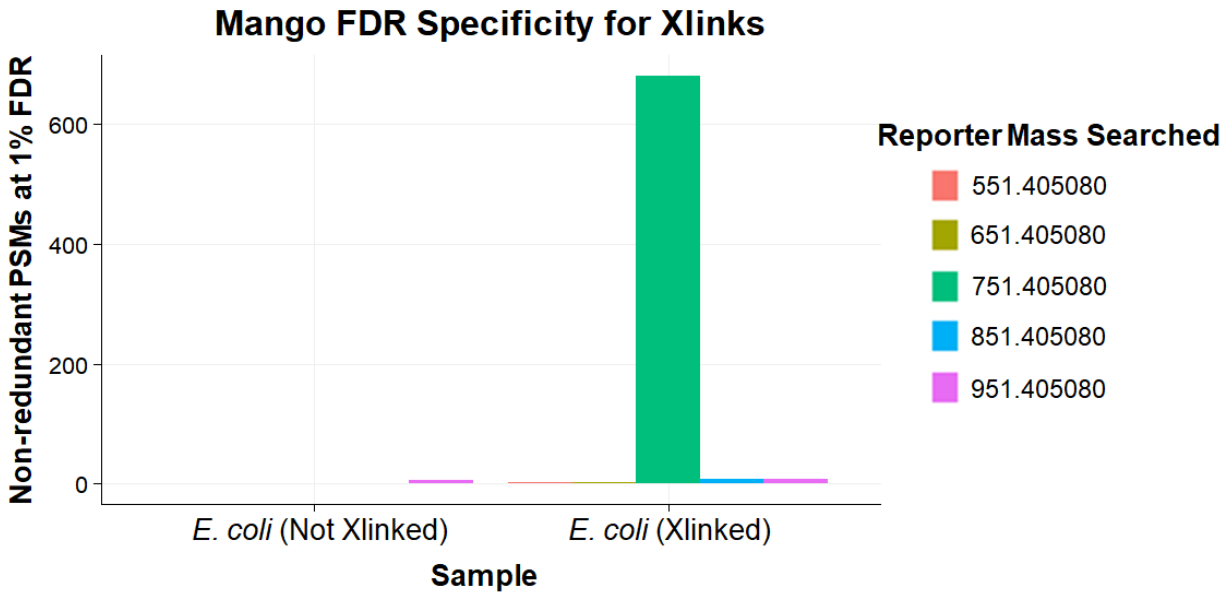


Figure A.5.2. The Mango-Comet pipeline finds only a significant number of cross-links when searching a cross-linked sample with a reporter mass that matches the cross-linker used.

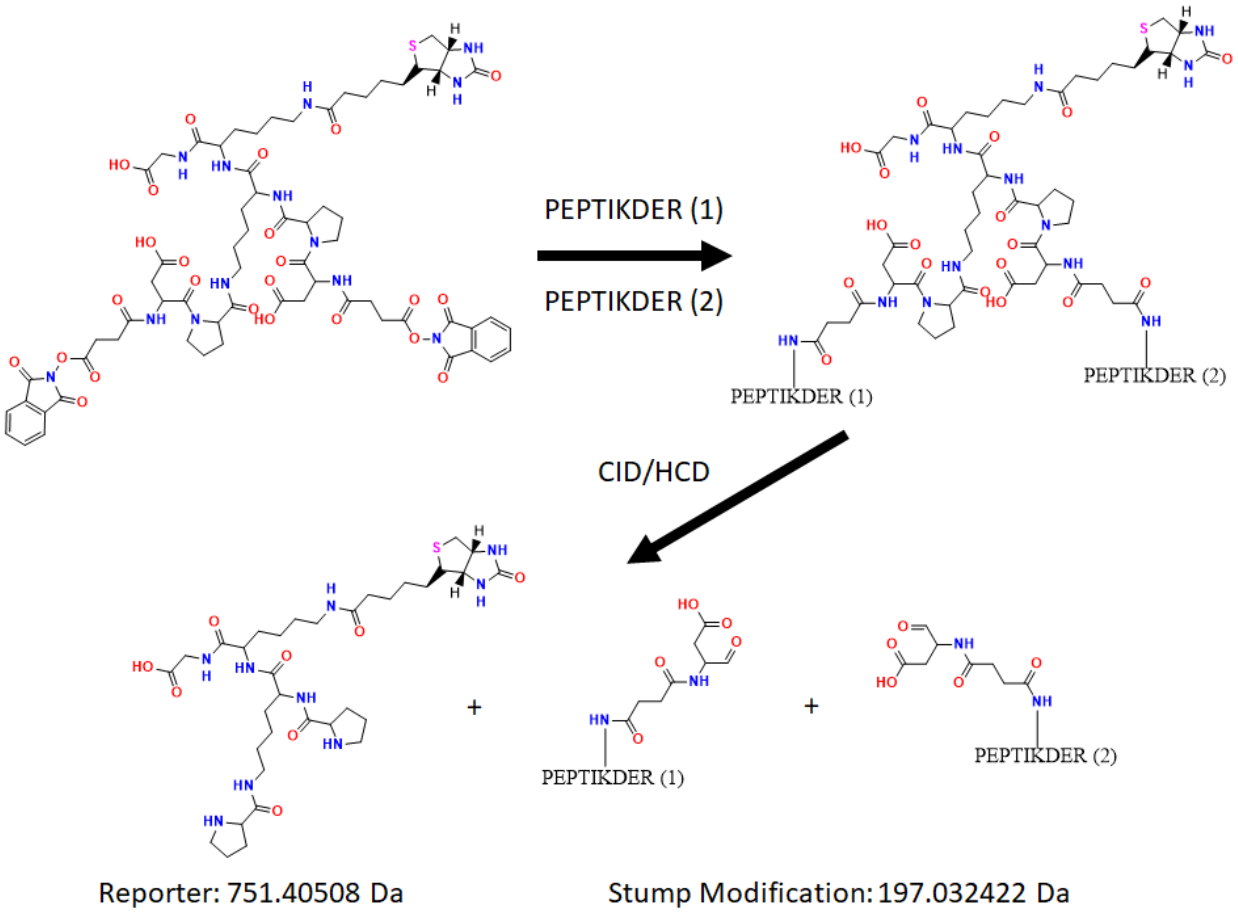
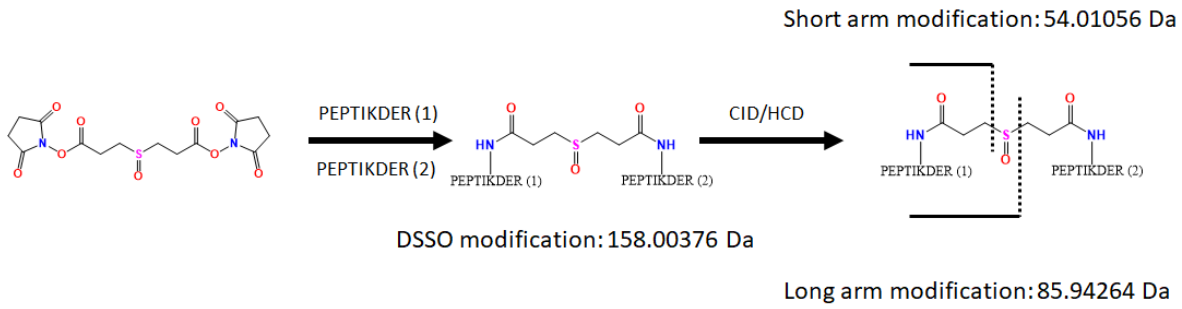


Figure A.5.3. Structure and fragmentation of BDP-NHP. The reporter species and its mass shown here are used as a parameter in Mango to direct the extraction of pairs of precursor masses from the spectra.

A



B

$$Short\ arm\ reporter = DSSO\ Modification\ - 2\ * (Short\ arm\ modification)$$
$$Short\ arm\ reporter = 158.00376\ Da\ - 2 * 54.01056\ Da$$
$$Short\ arm\ reporter = 49.98264\ Da$$

$$Long\ arm\ reporter = DSSO\ Modification\ - 2\ * (Long\ arm\ modification)$$
$$Long\ arm\ reporter = 158.00376\ Da\ - 2 * 85.94264\ Da$$
$$Long\ arm\ reporter = -13.96152\ Da$$

Figure A.5.4. Structure of DSSO. The reporter mass of DSSO is extracted directly from the known stump modifications and overall mass modification. Long arm and short arm reporters enable searches with fixed modifications on lysine, rather than combinations of variable mods that generate a larger search space. Masses for stump modifications and DSSO cross-link modification mass are taken from the XlinkX software.

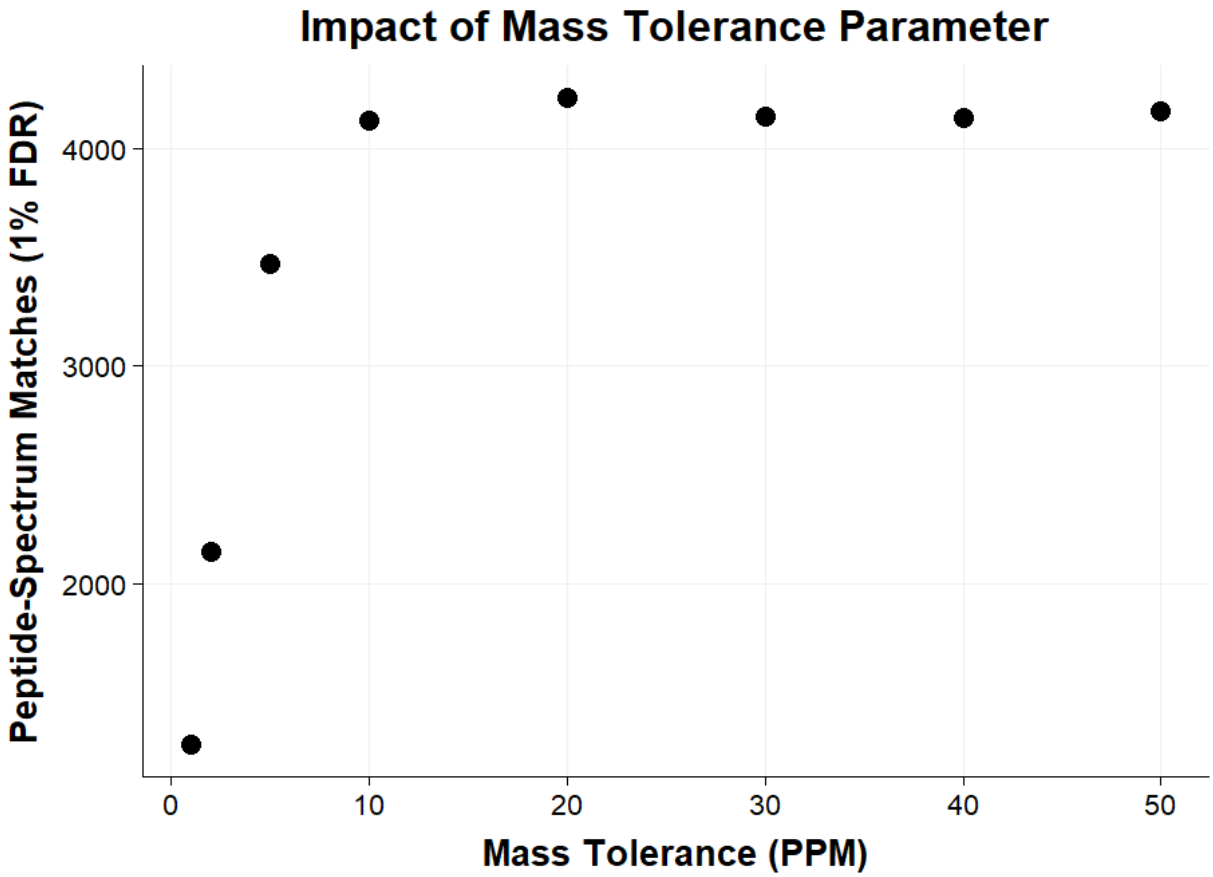


Figure A.5.5. The mass tolerance parameter in Mango determines how close a pair of peaks plus the reporter mass have to be to the isolated precursor mass to be reported. At extremely tight tolerances, a very small number of positive results are obtained due to losing results to the non-zero mass accuracy of the instrument. The number of positive results obtained plateaus around 10ppm, and increasing the tolerance further tends to lead to increased search times without a large change in the number of results.

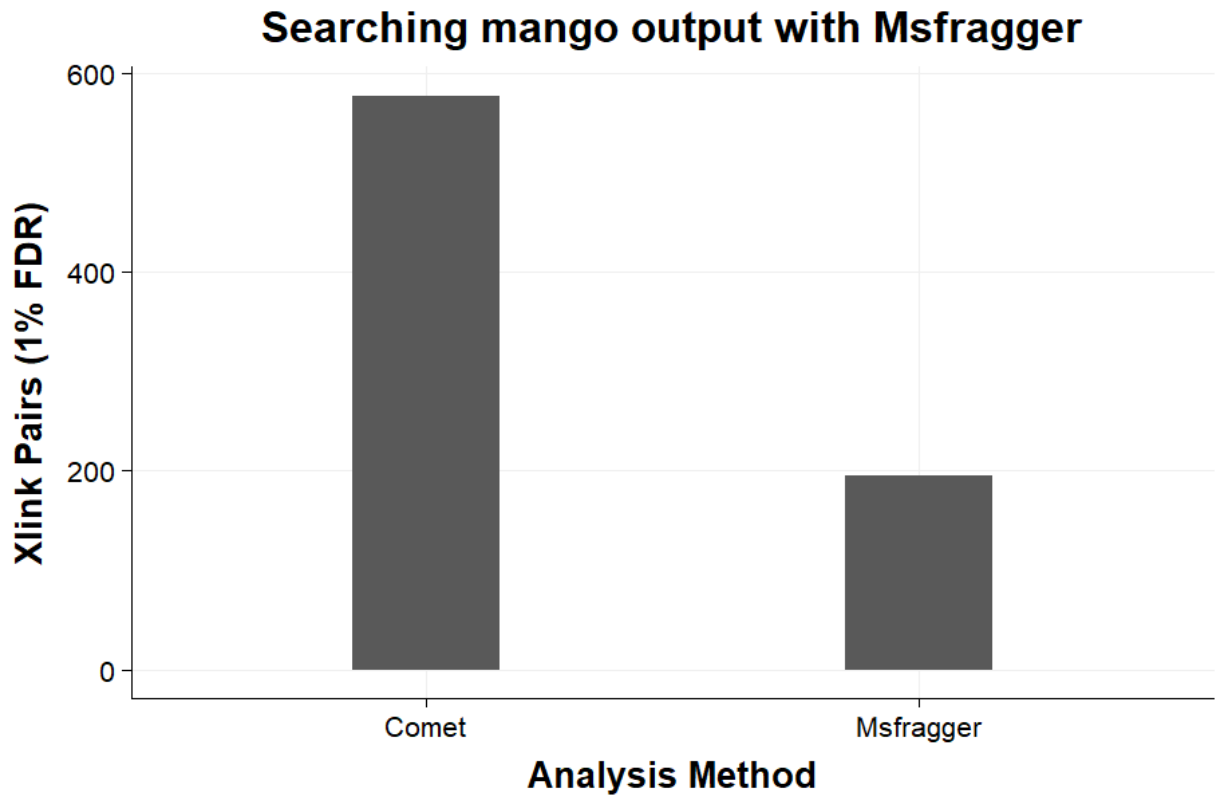


Figure A.5.6. Msfragger can be used to search the mgf output of Mango with no modifications required, although the pepXML output of Msfragger must be used to preserve spectra titles that encode the relationship. While fewer results were found with Msfragger, this is likely due to differences in the handling of variable modifications, and could be compensated for with additional optimization and post-processing. Additionally, the search is completed in a few seconds using Msfragger.

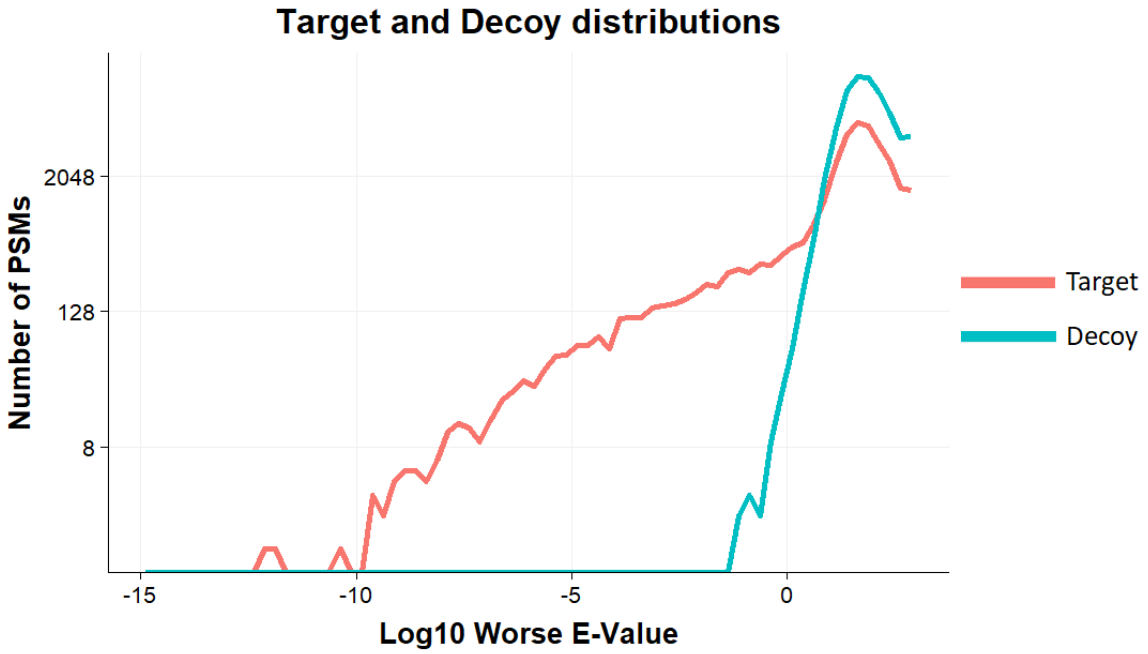


Figure A.5.7. False discovery analysis in Mango is performed by merging the results of all fractions, sorting by E-value, and then selecting an E-value cut-off such that 1% of the results above that cut-off are decoys (dashed line). A clear separation in the E-values of targets and decoys is observed in the tail of the distributions.

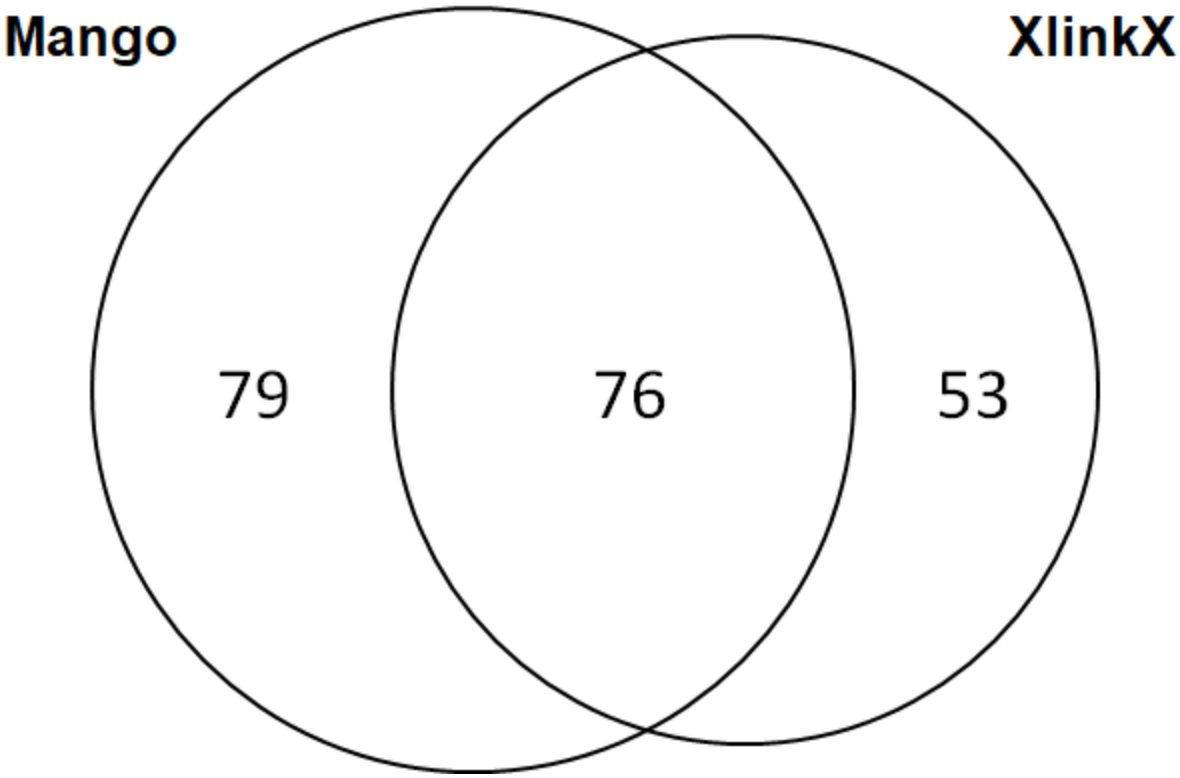


Figure A.5.8. Mango was used to search published data generated using the cross-linker, DSSO. While DSSO lacks a physical reporter ion, its predictable fragmentation facilitates the identification of released precursors using Mango. Using Mango and comet to search the CID-MS/MS spectra yields a similar number of results compared to XlinkX.

APPENDIX B: SUPPLEMENT FOR CHAPTER 3

Supplemental Methods

Cross-link level FDR estimation

FDR estimation was initially applied to the set of MS3 spectra using PeptideProphet, and results that satisfy the 1% FDR threshold were then grouped into cross-links. By obligation, this will underestimate the FDR at the group level due to accumulation of decoys. After grouping, the FDR was estimated at the unique cross-link level by performing an entrapment type search. The Mitocarta 2.0 database used in the original search (n=1042 forward sequences) was supplemented with protein sequences from Bacillus subtillis (n=4185 forward sequences), as well as reverse protein sequences for each forward sequence to perform target-decoy competition. All raw files from mitochondrial samples were searched again using this merged database and FDR filtering was performed using PeptideProphet as in the initial search. Results were then grouped into cross-links based off their scan groups, and the unique cross-link level FDR was estimated by counting the number of unique cross-links containing at least one peptide from a Bacillus subtillis sequence divided by the total number of unique cross-links. Empirically, this yields a sub 5% FDR at the unique cross-link level, with the majority of entrapment sequences appearing in binary links (**Supplemental figure 1**).

Observed distance distribution of different cross-linkers

The spacer arm length of a cross-linker determines what cross-links can be formed in a given system. XlinkDB automatically maps cross-link datasets to known structures from the Protein Data Bank and calculates the C_α- C_α distance for each mappable link. We compared the mapped distance distributions from three different mitochondrial data sets generated from three different cross-linkers: DSSO, BDP-NHP, and Bisby. DSSO data was obtained from the **Liu2017MCPmousemito_Heck** table and BDP-NHP data was obtained from the **SchweppemousemitoPNAS2017_BruceLab** table**.** From these data sets, we find that the observed distance distribution correlates to the cross-linker spacer-arm length (**Supplemental Figure 3**).

Supplemental Figures

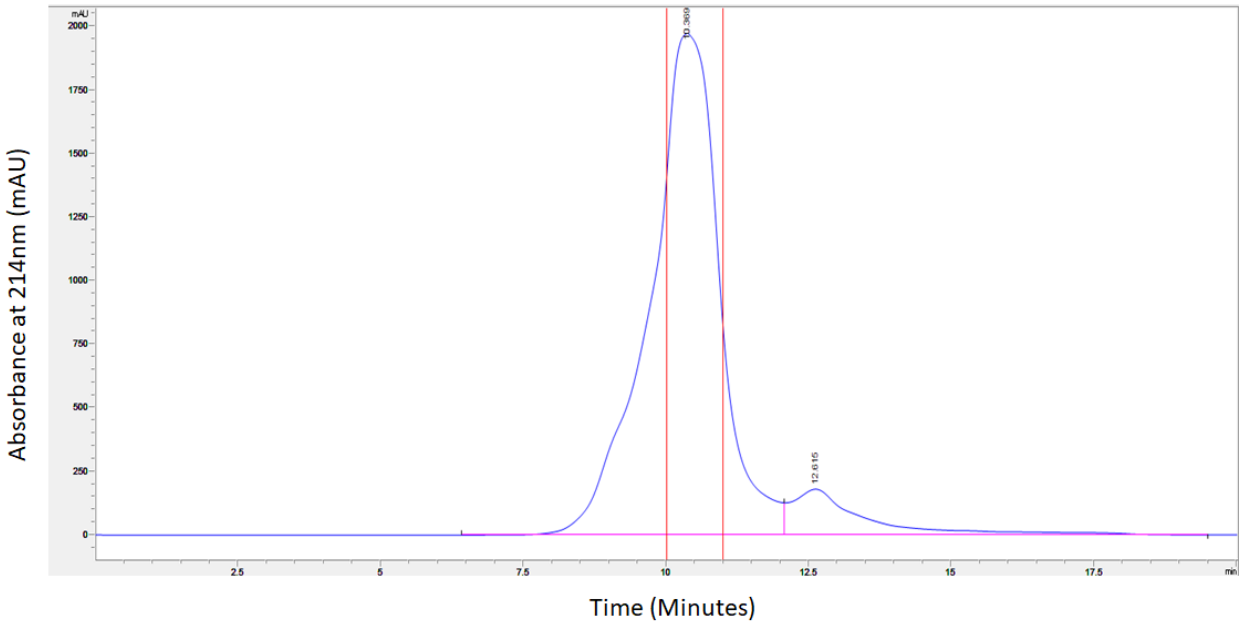


Figure B.5.9. Absorbance trace from low-PH reverse phase purification of crude Bisby product mixture after ether washes. The fraction retained for use in cross-linking experiments is between the two red lines. The left shoulder peak not retained contains 3-armed products that have a terminal proline on one of the arms due to failed aspartic acid coupling.

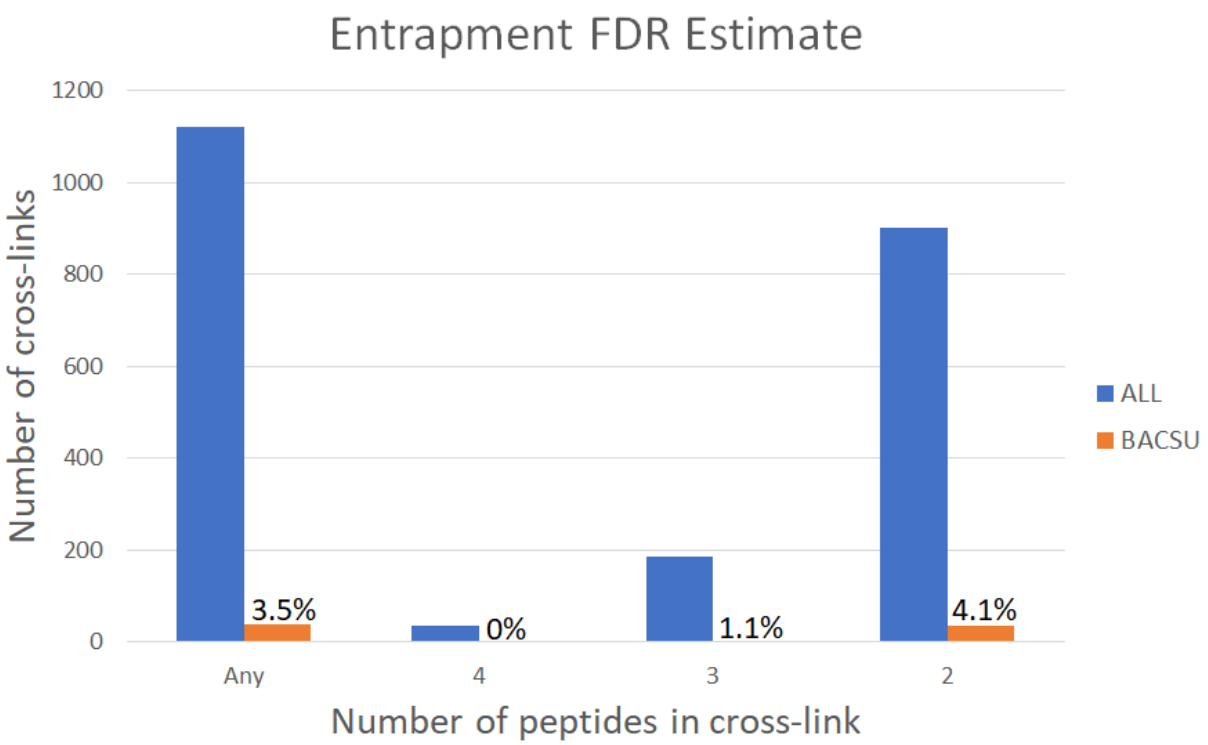


Figure B.5.10. Estimation of unique cross-link level FDR by entrapment strategy. PSMs originating from target BACSU sequences that survive target-decoy competition are necessarily incorrect, and act as a proxy to estimate the FDR at the cross-link level.

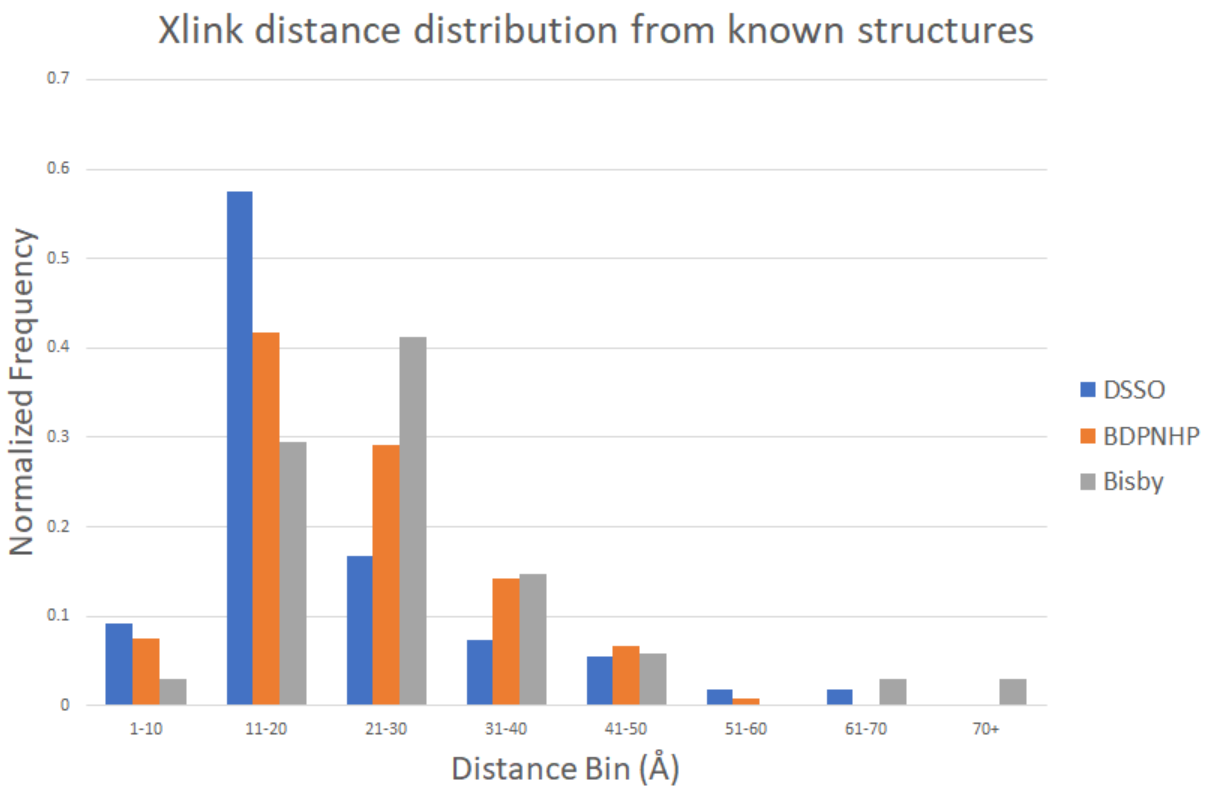


Figure B.5.11. Distance distributions ($C_\alpha - C_\alpha$) from known PDB structures obtained from three different mitochondria cross-linking experiments employing three different cross-linkers from datasets on XlinkDB. There is a correlation between the length of the spacer arm of the cross-linker and the observed distance distribution. The shortest linker, DSSO, produces the shortest cross-links on average while the longest cross-linker, Bisby, produces longer cross-links on average.

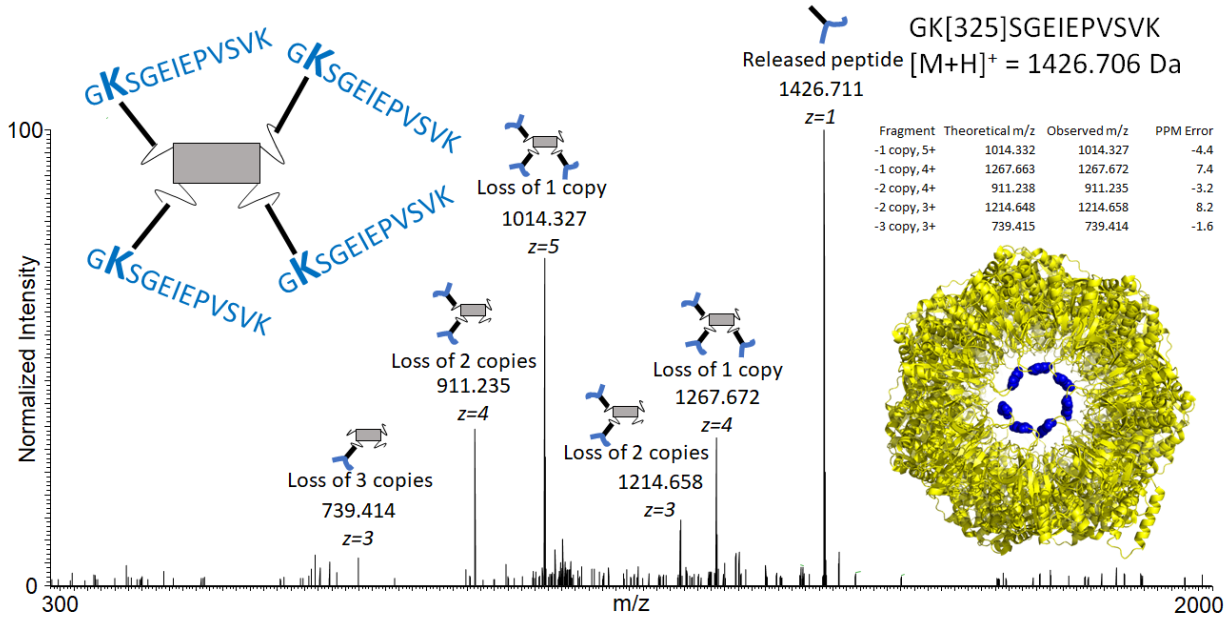


Figure B.5.12. An identified unambiguous homotetramer cross-link assigned to the CH10 heat shock protein. All possible complement ions corresponding to the loss of one to three copies of this peptide are observed in the MS2 spectrum, confirming its identity as an unambiguous homotetramer. CH10 exists as part of a homoheptameric complex (bottom right, PDB: 4PJ1), with the cross-linked lysine site highlighted in blue at the pore opening of the chaperone assembly. The observed tetra-link is consistent with any selection of four lysines from the pore.