The Development and Application of Mass Spectrometry-based Structural Proteomic Approaches to Study Protein Structure and Interactions

by

Karl Andrew Thomas Makepeace

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Biochemistry and Microbiology

We acknowledge and respect the lək̓ʷəŋən peoples on whose traditional territory the university stands and the Songhees, Esquimalt and W̱SÁNEĆ peoples whose historical relationships with the land continue to this day.

The Development and Application of Mass Spectrometry-based Structural Proteomic Approaches to Study Protein Structure and Interactions

by

Karl Andrew Thomas Makepeace

Supervisory Committee

Dr. Christoph H. Borchers, **Co-supervisor**
(Department of Biochemistry & Microbiology)

Dr. Perry L. Howard, **Co-supervisor**
(Department of Biochemistry & Microbiology)

Dr. Caroline E. Cameron, **Departmental Member**
(Department of Biochemistry & Microbiology)

Dr. John E. Burke, **Departmental Member**
(Department of Biochemistry & Microbiology)

Dr. Stephanie M. Willerth, **Outside Member**
(Department of Mechanical Engineering)

ABSTRACT

Proteins and their intricate network of interactions are fundamental to many molecular processes that govern life. Mass spectrometry-based structural proteomics represents a powerful set of techniques for characterizing protein structures and interactions. The last decade has witnessed a large-scale adoption in the application of these techniques toward solving a variety of biological questions. Addressing these questions has often been coincident with the further development of these techniques.

Insight into the structures of individual proteins and their interactions with other proteins in a proteome-wide context has been made possible by recent developments in the relatively new field of chemical crosslinking combined with mass spectrometry. In these experiments crosslinking reagents are used to capture protein-protein interactions by forming covalent linkages between proximal amino acid residues. The crosslinked proteins are then enzymatically digested into peptides, and the covalently-coupled crosslinked peptides are identified by mass spectrometry. These identified crosslinked peptides thus provide evidence of interacting regions within or between proteins.

In this dissertation the development of tools and methods that facilitate this powerful technique are described. The primary arc of this work follows the development and application of mass spectrometry-based approaches for the identification of protein crosslinks ranging from those which exist endogenously to those which are introduced synthetically. Firstly, the development of a novel strategy for comprehensive determination of naturally occurring protein crosslinks in the form of disulfide bonds is described. Secondly, the application of crosslinking reagents to create synthetic crosslinks in proteins coupled with molecular dynamics simulations is explored in order to structurally characterize the intrinsically disordered tau protein. Thirdly, improvements to a crosslinking-mass spectrometry method for defining a protein-protein interactome in a complex sample is developed. Altogether, these described approaches represent a toolset to allow researchers to access information about protein structure and interactions.

CONTENTS

Supervisory Committee **ii**

Abstract **iii**

Table of Contents **iv**

List of Tables **vii**

List of Figures **viii**

Abbreviations **x**

Acknowledgements **xv**

Dedication **xvi**

1 Introduction **1**

1.1 Proteomics 1

1.1.1 Proteins 2

1.1.1.1 Protein structure 2

1.1.2 Mass spectrometry 6

1.1.2.1 Instrumentation 11

1.1.3 Bottom-up proteomics 13

1.2 Structural Proteomics 14

1.2.1 Chemical crosslinking-combined with MS 15

1.2.1.1 Crosslinking-MS experimental workflow 16

1.2.1.2 Crosslinking-MS data analysis 17

1.3 Motivations, research questions, and objectives 19

2 Comprehensive identification of disulfide bonds using non-specific proteinase K digestion and CID-cleavable crosslinking analysis methodology for Orbitrap LC/ESI-MS/MS data 21
2.1 Introduction . . . 22
2.2 Materials and Methods . . . 25
2.2.1 Proteolytic digestion . . . 25
2.2.2 Mass spectrometry analysis . . . 25
2.2.3 Mass spectrometry data analysis . . . 26
2.3 Results & Discussion . . . 27

3 Insight into the Structure of the "Unstructured" Tau Protein 33
3.1 Introduction . . . 34
3.2 Materials and Methods . . . 37
3.2.1 Expression and Purification of the Tau Protein . . . 37
3.2.2 Crosslinking . . . 38
3.2.3 LC-MS/MS Analysis . . . 39
3.2.4 Surface Modification . . . 40
3.2.5 Discrete Molecular Dynamics Modeling . . . 40
3.2.6 Unconstrained Molecular Dynamics Modeling . . . 41
3.2.7 Data and Code Availability . . . 42
3.3 Results and Discussion . . . 42
3.4 Conclusion . . . 58

4 Improving identification of *in-organello* protein-protein interactions using an affinity-enrichable, isotopically-coded, and mass spectrometry-cleavable chemical crosslinker 59
4.1 Introduction . . . 60
4.2 Materials and Methods . . . 62
4.2.1 Mitochondria preparation and *in-organello* crosslinking . . . 62
4.2.2 Sample Lysis, Pre-fractionation and Digestion . . . 63
4.2.3 Enrichment of crosslinked peptides . . . 64
4.2.4 LC-MS/MS analysis . . . 64
4.2.5 Data-dependent Acquisition methods . . . 65
4.2.6 MS1 Feature Analysis . . . 67
4.2.7 Bioinformatics Analysis . . . 67
4.2.8 Structural Validation of Crosslink Identifications . . . 70
4.2.9 Experimental Design and Statistical Rationale . . . 70
4.3 Results . . . 71

4.3.1 Developing an integrated experimental and computational CL-MS workflow. 71
4.3.2 Affinity enrichment for improved detection of crosslinker-modified peptides 72
4.3.3 Isotopic-coding for the specific acquisition of crosslinker-modified peptides 72
4.3.4 Integrating crosslinker-specific mass-spectral feature information for improved performance in PSM validation 76
4.3.5 Overview of the identifications with respect to fractionation . 85
4.3.6 The yeast mitochondria interactome 86
4.3.7 Structural validation of crosslinks on existing structural models 87
4.4 Discussion 90

5 General Discussion **108**
5.1 Summary 108
5.1.1 Chapter 2: Disulfide bonds, non-specific digestion, and higher-order CL-MS 109
5.1.2 Chapter 3: Tau, short-distance CL-MS, and computational structural modeling 112
5.1.3 Chapter 4: Improving the detection, acquisition, and identification of crosslinks in large-scale CL-MS analyses 115
5.2 CL-MS evolution and horizons 118

Bibliography **119**

LIST OF TABLES

2.1 Disulfide crosslink determination in proteins with known and unknown disulfide connectivities 30

3.1 Short-distance crosslinks used for CL-DMD simulations 43
3.2 Photo-reactive SDA dead-end crosslinked tau residues used for surface modification analysis 49
3.3 Long-distance DSS crosslinks used for long-distance crosslinking analysis 50

4.1 Hardklör parameters 66
4.2 Krönik parameters 66
4.3 Kojak parameters 69
4.4 Percolator parameters 70
4.5 Description of all features used to represent PSMs 78
4.6 A comparison of recent mitochondria mass spectrometry-based crosslinking studies 88
4.7 Protein-protein interactions with highest number of PSMs . . . 89

LIST OF FIGURES

2.1 Diagram of the CID-cleavage of the Cys–Cys crosslink 24
2.2 Disulfide bond determination data analysis workflow 28
2.3 Disulfide determination in bovine serum albumin 31

3.1 Conformational ensemble of native tau in solution as determined by CL-DMD 51
3.2 Original and relaxed structures of the lowest energy centroid . . 53
3.3 Short-distance crosslinks used for CL-DMD of tau protein in solution 54
3.4 Circular dichroism analysis of the native tau protein in solution 55
3.5 Experimental validation of the tau structure with surface modification and long-distance crosslinking 56

4.1 Crosslinking reagent and experimental workflow 73
4.2 Affinity enrichment improves detection of crosslinker-modified peptides 93
4.3 MS^1 Δ 8.0502 Da doublet features identified in Krönik output . 94
4.4 Duty cycle utilization for each acquisition method (TopS, MTag, or Incl) 95
4.5 Targeted acquisition improves the coverage of crosslinker-modified peptides 96
4.6 Schematic of method and approximate time required for acquisition of MTag and Incl datasets for a sample fraction 97
4.7 Crosslinker-specific mass spectrum features improve crosslinker-modified peptide identification 98
4.8 Overview of the crosslink identifications 100
4.9 Circle diagram of the protein-protein interaction network determined from identified crosslinks 101

4.10 Protein-protein interaction network analysis and sub-compartment localization of the identified crosslinks 103
4.11 Identified crosslinks from high centrifugation fraction mapped to PDB structures of yeast mitochondrial ETC complexes and super-complexes 104
4.12 Identified crosslinks mapped to PDB structures of yeast mitochondrial ETC complexes and super-complexes for all sample pre-fractions 106

ABBREVIATIONS

3D	3-dimensional
ABC	ammonium bicarbonate
ACN	acetonitrile
AE	affinity enrichment
AGC	automatic gain control
atm	atmosphere
ATP	adenosine triphosphate
CBDPS	cyanurbiotin-dimercaptopropionyl-succinimide
CD	circular dichroism
CHARMM	Chemistry at HARvard Macromolecular Mechanics
CID	collision-induced dissociation
CL	crosslink
CL-PSM	crosslink-peptide-spectrum match
CL-DMD	crosslinking constraint-guided discrete molecular dynamics
CLMP	crosslinker-modified peptide
CL-MS	crosslinking mass spectrometry
DDA	data-dependent acquisition
DMD	discrete molecular dynamics
DNA	deoxyribonucleic acid
DSA	disuccinimidyl adipate
DSG	disuccinimidyl glutarate

DSS	disuccinimidyl suberate
DTT	dithiothreitol
E. coli	*Escherichia coli*
EBI	European Bioinformatics Institute
EDTA	ethylenediaminetetraacetic acid
EMBL	European Molecular Biology Laboratory
EPR	electron paramagnetic resonance
ESI	electrospray ionization
ETC	electron transport chain
ETD	electron transfer dissociation
FA	formic acid
FASP	filter-aided sample preparation
FDR	false-discovery rate
FRET	Förster resonance energy transfer
FT	Fourier transform
FTMS	Fourier transform mass spectrometry
GROMACS	GROningen MAchine for Chemical Simulations
H	heavy
HCD	higher-energy collisional dissociation
HCl	hydrogen chloride
HDX	hydrogen-deuterium exchange
HPLC	high performance liquid chromatography
Hz	hertz
IAA	iodoacetamide
ICC-CLASS	Isotopically-Coded Cleavable Crosslinking Analysis Software Suite
ID	inner diameter
IDP	intrinsically disordered protein
Incl	inclusion list
IPTG	isopropyl β-D-1-thiogalactopyranoside
KCl	potassium chloride

KOH	potassium hydroxide
L	light
LB	lysogeny broth
LC	liquid chromatography
LC-MS/MS	liquid chromatography-tandem mass spectrometry
LC-MS	liquid chromatography-mass spectrometry
LTQ	linear trap quadropole
m/z	mass-to-charge ratio
MALDI	matrix-assisted laser desorption/ionization
MD	molecular dynamics
MOPS	3-(N-morpholino)propane sulfonic acid
mRNA	messenger RNA
MS	mass spectrometry
MS^1	full scan mass spectrum
MS^2	2-stage fragment ion scan mass spectrum
MS^3	3-stage fragment ion scan mass spectrum
MS^n	multi-stage/sequential tandem mass spectrometry
MS/MS	tandem mass spectrometry
MTag	MassTag
NaCl	sodium chloride
NCE	normalized collision energy
nESI	nano-electrospray ionization
NHS	N-hydroxysuccinimide
Ni-NTA	Nickel-nitrilotriacetic acid
NMR	nuclear magnetic resonance
NMWL	nominal molecular weight limit
NOE	nuclear Overhauser effect
NPT	isothermal-isobaric Nosé–Hoover thermostat
ns	nanosecond
NVT	Nosé–Hoover thermostat
OD	outer diameter

PBS	phosphate buffered saline
PDB	Protein Data Bank
PET	positron emission tomography
PK	proteinase K
PME	particle mesh Ewald
PMSF	phenylmethylsulfonyl fluoride
PPI	protein-protein interaction
ppm	parts per million
PRE	paramagnetic relaxation enhancement
PRIDE	PRoteomics IDEntifications
PSM	peptide-spectrum match
qCL	quantitative crosslinking
RCF	relative centrifugal force
REX	replica exchange
RF	radio frequency
RMSD	root-mean-square deviation
RNA	ribonucleic acid
RPM	rotations per minute
RTime	retention time
SAXS	small-angle X-ray scattering
SCX	strong cation exchange
SDA	succinimidyl 4,4′-azipentanoate
SDS-PAGE	sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SEC	size-exclusion chromatography
SGD	Saccharomyces Genome Database
SPE	solid-phase extraction
TCEP	tris(2-carboxyethyl)phosphine
TFA	trifluoroacetic acid
Th	Thomson
TIP3P	transferable intermolecular potential with 3 points
TopN	TopN
TopS	TopSpeed

UV	ultraviolet
UVPD	ultraviolet photodissociation

ACKNOWLEDGEMENTS

I would like to thank:

Dr. Christoph Borchers, for providing an opportunity to receive an education that went above and beyond anything I had expected when undertaking my studies. I thank him especially for encouragement and backing to engage with the international mass spectrometry and proteomics research community through many conferences and symposia over the years. It has been a privilege to work in such an exciting and dynamic field and see it develop and grow with a front-row seat.

Dr. Perry Howard, for his advising, support, leadership, advocacy, genuine encouragement and optimism without which this dissertation could not have been completed.

Drs. John Burke, Stephanie Willerth and Caroline Cameron, for their valuable feedback, expertise, and support in their roles as committee members.

Drs. Chris Nelson and Caroline Cameron, for their guidance, support, and advocacy in their role as Graduate Advisor during my program.

Members of the UVic Genome BC Proteomics Centre, past and present, for helping me over the course of my time as student, then research technician, and later graduate student. A special thank you to Darryl Hardie for his always generous technical support and Drs. Nicole Sessler and Yassene Mohammed for their support in both personal and professional capacities over the years.

Drs. Jason Serpa and Nicholas Brodie, for your comradery as fellow bench mates and compadres in curiosity and crosslinking.

The Biotechnical Support Centre, for all of their help in facilitating use of departmental equipment for my experiments. Also, a special thanks to Scott Scholz for ensuring I wouldn't be stuck inspecting MS spectra at 480p on XP and assembling a workstation that was a joy to work with.

Collaborators and co-authors, for all of your efforts and contributions to the work that is represented in each and every chapter of this dissertation.

Sherri, Rei, Mom, and Dad, for your unwavering and unconditional support and love. Truly, I could not have completed this without you.

DEDICATION

To, more than anyone, Sherri.

You have always been there for me and I cannot thank you enough.

CHAPTER

ONE

INTRODUCTION

1.1 Proteomics

Following the structural determination of deoxyribonucleic acid (DNA) (1953) [1] a hierarchical framework for understanding the biochemistry of life was established. The central dogma of molecular biology (1958) [2, 3] introduces a framework that describes the flow of genetic information from DNA to ribonucleic acid (RNA), and from RNA to protein. Protein-coding gene sequences in DNA are transcribed and processed into messenger RNA (mRNA) which is in turn translated into proteins. This process links genotype to phenotype and is key to understanding the relationship between the genetic constitution of an individual and the observable characteristics or traits of that individual.

The genome represents the whole of an organism's genetic information while the proteome represents its protein complement [4]. Proteins are the ultimate products of the gene expression process and are the driving biological entities in determining phenotypes. Unlike the genome, which remains essentially constant and

unchanged throughout the life of a cell, proteins are constantly being synthesized and degraded depending upon the cellular environment. Because of the highly dynamic variability and adaptability of the proteome in response to environment it is an apt and information-rich measure of the state of an organism. The global study of this protein complement to the genome is termed proteomics.

1.1.1 Proteins

In keeping with the etymology of the word (coined by Dutch chemist Gerhard Johannes Mulder [5]), "proteins" are biological entities of primary importance to the orchestration of life in and amongst cells. They conduct their molecular processes in concert and, in so doing, define each cell's specific biological functions and characteristics.

1.1.1.1 Protein structure

The role of proteins as key effectors in cellular function is governed by their specific biomolecular structures and interactions. A paradigm to understand these structures is represented by four distinct levels of protein structure.

Primary structure The most elementary level of protein structure, primary structure, is determined by the specific linear sequence of amino acid residues that comprise a single continuous polypeptide chain. There is a set of 20 standard proteinogenic amino acids which comprise the monomeric units that can be covalently linked together into polymeric chains called peptides, polypeptides, or proteins. Each amino acid is linked to the next through a peptide bond yielding covalently linked amino acid residues. These peptide bonds are formed during protein synthesis through a condensation reaction between the carboxyl group of the amino acid

residue at the growing end of the polypeptide chain and the amino group of the next amino acid. Following synthesis of the polypeptide, amino acid residues may also be modified in various ways. These modifications can occur on the side-chains of residues and the N' or C' terminal ends of the polypeptide chain. They are important in modulating protein function, and can lead to changes in overall biochemical signalling within and between cells [6]. A modification of particular interest in this dissertation, and the subject of Chapter 2, are cystine disulfide bonds. These are modifications in which the thiol groups found on the side-chains of cysteine residues are covalently coupled resulting in a linkage between two sites within or between polypeptide chains. These linkages are particularly important in stabilizing higher order protein structure.

Secondary structure The next level of protein structure, secondary structure, is determined by locally folded regions of the protein chain where non-covalent interactions between atoms of the backbone occur. The most common of these regions are represented by secondary structures called α-helices and β-pleated sheets [7, 8]. These secondary structures establish and maintain their form through hydrogen bonds which occur between the carbonyl group of one amino acid and the amino group of another. Algorithmic computational classifications of protein secondary structure have been developed [9, 10] and are used to define and represent these structures as ribbons [11] in 3D molecular viewing software such as PyMol [12]. These classifications of secondary structure are used in Chapter 3 to highlight the transient existence of such features in an intrinsically disordered protein (IDP). Protein helices may differ in handedness or the number of residues required to complete a turn. The most common helix obsvered is a right-handed helix with each turn containing 3.6 amino acid residues. Here, the carbonyl of an amino acid residue hydrogen bonds with the amino hydrogen atom of an amino acid residue that is four residues ahead

in the chain. Sidechains of the amino acid residues point outward where they can intereact with the surroundings. A β-pleated sheet involves two or more polypeptide chain segments lining up alongside each other resulting in a sheet-like structure stabilized by backbone carbonyl to amino group hydrogen bonds. The sidechains of the amino acids point both above and below the plane of the sheet in alternating fashion. The arrangement of the adjacent strands can be either parallel, in the case that N- and C-termini are oriented in the same direction for both strands, or anti-parallel if not. Segments that are not α-helix or β-strand are called loop or coil segments. These segments have no obvious regular patterns in their structure. In addition to serving to link together helical and sheet segments or as flexible N- and C- termini they often contribute to determining functional specificity of binding sites within a protein [13]. Typically, around half of the residues in a protein are in loop segments [14].

Tertiary structure The next level, tertiary structure, is the overall three-dimensional fold of the polypeptide chain. This is primarily brought about by the non-covalent interactions between amino acid residue sidechains and hydrophobic interactions wherein amino acid residues with non-polar sidechains preferably cluster together inside of the protein fold and away from the aqueous environment. Disulfide linkages are also an important type of covalent linkage that act like molecular "staples" holding regions of a folded or folding polypeptide in place. These linkages can occur between spatially proximate but sequentially distant amino acid residues.

Quaternary structure Finally, quaternary structure occurs when two or more polypeptide chains come together to form a protein complex. The polypeptide chains may be identical or distinct and are held together by the same interactions that mediate tertiary structure. The complexation of polypeptides together to

form protein complexes represent a set of protein-protein interactions (PPIs). A diverse and dynamic network of PPIs are responsible for myriad biological activities. PPIs may differ in composition, affinity and whether the interaction is stable or transient. Stable interactions are long-lived interactions between protein subunits which form a protein complex. Transient interactions are short-lived interactions that are important in signalling or regulating changes in biological function or activity.

The extent to which a PPI is stable or transient exists on a continuum that depends on the physiological context [15, 16]. For example PPIs that are transient under normal physiological conditions may be replaced by PPIs that are stable or permanent under pathological conditions, as is the case in some proteopathies [17, 18] and discussed in more detail in Chapter 3. The large-scale identification of PPIs in a complex biological context is the subject of Chapter 4.

All of these conceptual levels of protein structure, from primary to quaternary, operate in concert to allow a protein to adopt a biologically functional form and engage in PPIs. Thus, insight into these aspects of protein structure and interactions are of great importance in understanding the basis of cellular functions at a molecular level. An improved understanding in these areas could lead to more informed drug development and improved molecular models of disease. Each of the following chapters in this dissertation utilizes and endeavors to develop methods capable of experimentally accessing protein structural information at each of these levels.

Denaturation Each protein or protein complex has a distinct native three-dimensional structure or set conformations however destabilization of this structure can be accomplished by environmental changes that disrupt the interactions leading to the higher-order structure described above. A protein is denatured when it loses its structure. Denaturation can be accomplished by various methods such as temperature change, use of chemical detergents, or pH change. Denaturation is

also a key part of the mass spectrometry-based bottom-up proteomics workflow. It allows proteases to more efficiently digest the polypeptide chains that constitute the proteins of the proteome. Covalent linkages such as disulfide bonds or synthetically introduced crosslinks between amino acid residues are preserved following denaturation. Structural information about proximal amino acid residues in the native structure are encoded in these linkages. Exploring and developing methods to exploit this fact and capture this information is the primary focus of this dissertation.

1.1.2 Mass spectrometry

Mass spectrometry (MS) is an analytical technique capable of accurately measuring the mass-to-charge ratio (m/z) of molecules in a sample. This information can then be used in determining the precise molecular weights of the molecules that constitute the sample. With this capability, the technique is useful for identifying known and unknown compounds qualitatively and quantitatively. The earliest demonstrations of the technique were reported in the early 1910's where J.J. Thomson produced the first mass spectrographs providing experimental evidence for the existence of isotopes for a stable element [19, 20]. In 1991, the Thomson (Th) was proposed [21] as a unit representing the measure of mass-to-charge ratio (m/z) in mass spectrometry (MS) in Thomson's honour. The term is occasionally used; however, one encounters units of m/z more regularly in current literature, so it is the term used throughout this dissertation.

There are three fundamental components common to all mass spectrometers: (1) ionization source; (2) mass analyzer; (3) ion detection system. These three components map to three necessary processes respectively: (1) creating ions; (2) mass selection/ion manipulation; (3) ion detection with digital recording. A mass spectrometer may also include other functional modules such as fragmentation cells

and specialized ion focusing/routing. These additional components enable deeper structural characterization of sample analytes and improved sensitivity.

Creating ions All mass analyzers function on the basis of electromagnetism with measurement dimensions of mass and charge; therefore, ionization of the molecules in the sample to be analyzed is required. Both positive and negative charge ions are amenable to mass spectrometric analysis; however, analysis of each polarity is carried out separately. Typically positively charged ions are analyzed in MS-based proteomics. Samples are introduced to the mass spectrometer at acidic pH, and, under such conditions, positively charged protein and peptide ions are produced. The ionized molecules must also be transferred from the liquid or solid phase that a sample may initially be in and into the gas phase, allowing manipulation by external electric and magnetic fields. This ionization process and introduction to the gas phase happens at what is called the ion source of the mass spectrometer.

All of the MS experiments presented in this dissertation were conducted using a molecular ionization technique called nano-electrospray ionization (nESI) [22]. This technique is performed by applying a high voltage to a mobile liquid phase sample dissolved in a polar and volatile solvent at flow rates of around 200 nanolitres per minute. An aerosol spray of charged droplets is then formed by passing the sample liquid through the tip of a needle positioned at the intake orifice of a mass spectrometer at atmospheric pressure. These droplets enter the mass spectrometer where they are exposed to a high voltage potential difference and high temperatures leading to desolvation. After the solvent molecules in the droplets evaporate, analytes, which are now in the gas phase, inherit the charges that the droplets carried. This method is considered a "soft-ionization" technique in which the ionized molecules undergo minimal fragmentation and are passed into the instrument as intact molecular ions.

Multiply-charged ions are also typically formed with this technique leading to peptide ions of charge state 2+ or greater. For any particular molecular species, a range of charge states may also be generated and observed (e.g., a peptide species observed with 2+ and 3+ charge states). Although this was initially thought of as a drawback of the ionization technique that added complexity to the resulting spectra, it is now considered a strength for the additional information it provides [23]. Also, the observed mass range is effectively extended due to producing ions with higher charge states which are consequently observed at lower m/z. ESI integrates well with upstream and online liquid chromatography (LC) separations, further enhancing the analytical depth of the experiment.

Mass selection Once ionized, the molecules are now able to be accelerated through the mass spectrometer under low pressure. On the basis of m/z ratios, the electric or magnetic fields of a mass analyzer can deflect or trap individual ions. A variety of mass analyzers exist, which include time-of-flight, Orbitraps, quadrupoles, and ion traps. Each type has distinct qualities; however, all function to manipulate ions and either separate analytes in the sample to record a mass spectrum and/or to selectively accumulate analytes of a particular m/z range and shuttle them to other functional modules of the mass spectrometer.

The first measurement (mass detector readout) taken of the intact molecular ions generated at the ion source, absent any fragmentation step, results in a mass spectrum often referred to as the "full scan". In the context of a multi-stage/sequential tandem mass spectrometry (MS^n) experiment where those ions observed in the full scan are subsequently selectively isolated, accumulated, and fragmented, the full scan may also be referred to as an MS^1 scan.

The subsequent spectra of fragment ions generated from fragmentation of a subset of the ions observed in the MS^1 scan are called MS^2 scans. Alternative names for

these scans may include product ion scan, fragment ion scan, and daughter ion scan. The intact ions which were fragmented may also be referred to as the precursor or parent ions. Further cycles of ion isolation, fragmentation, and measurement are named according to MS^n, where "n" is the number of measurement cycles. These cycles are carried out on a sub-second timescale compatible with the online coupling of high-resolution LC. Throughout the dissertation, spectra are referred to in MS^n terms.

Ion detection In order to record a mass spectrum, a mass analyzer works as a component in an ion detection system. In this system, the mass analyzer functions to separate ions based on m/z ratios. A mass detector then functions to detect and convert those ions to a digital output representing distinct m/z signals with associated signal intensities.

Electron multipliers are commonly used as mass detectors. They function by first colliding the ions into a metal dynode which releases electrons which are then amplified through a series of dynode collisions and additional electron emissions. These mass detectors have very high gain and low noise where a signal from a single ion is amplified several orders of magnitude. This amplification is captured as an electrical signal and recorded by a computer as mass spectrum data.

A non-destructive alternative mass detector is based on image current detection. Here, instead of ions colliding with metal, axial oscillating movement of ions around a central spindle-like electrode induces a current on the sections of a halved outer barrel-like electrode kept at ground potential. This image current produces a waveform that is then Fourier-transformed by a computer into frequencies representing the m/z ratios and intensities of the analytes. Detection systems using Fourier transform (FT) allow for very high resolving power, mass accuracy, and dynamic range [24].

The product of an ion detection system is ultimately a mass spectrum: a 2-

dimensional plot of m/z ratio and intensity. The signal intensity is related to the relative abundance of the various analytes in the sample. However, it is not directly interpretable as the precise abundance of an analyte as many factors related to the characteristics of individual analytes (e.g., ionization efficiency) and instrumentation can influence this parameter.

Ion fragmentation Structural information about ions can be apprehended through MS^n. In order to access this structural information, fragmentation of the gas-phase ions is carried out between cycles of mass analysis and detection. This information can include primary sequence and amino acid modification location(s) in proteins or peptides. Indeed, the process of fragmenting and collecting mass spectra for a peptide is often referred to as "sequencing" a peptide.

Typically precursor ions within a narrow m/z ratio range are selectively isolated and accumulated in an ion trap mass analyzer. Once a predetermined number of ions or length of time passes, the accumulated ions are either activated and fragmented in the trap or shuttled to another ion trapping module where fragmentation will take place. Several fragmentation methods exist. Different methods may employ fundamentally different fragmentation mechanisms, which can result in distinct patterns of fragment ions, and therefore distinct information about the structure of the precursor ion.

The work presented in this dissertation uses fragmentation methods called collision-induced dissociation (CID) and higher-energy collisional dissociation (HCD) [25]. Both methods accomplish fragmentation with the principle mechanism involving the collision of precursor ions into atoms or molecules of an inert collision gas (e.g., argon, helium, or nitrogen). Collisional kinetic energy is transferred to internal energy in the ion, which can lead to bond breakage [26]. A mass analyzer applies the collision energy. For all data presented in this dissertation, CID occurs in a

linear trap quadropole (LTQ), and HCD occurs in a specialized HCD module consisting of an ion trapping multipole coupled to a C-trap wherein fragment ions are retained using higher radio frequency (RF) voltages prior to mass detection [25]. For peptide parent ions, both CID and HCD fragment ion mass spectra typically feature fragment ion series from breakage at the peptide bonds (-CO-NH-) between amino acid residues. These fragments are called b- and y-fragment ions [27]. A characteristic series-specific relationship between these peptide backbone fragment ions and their observed intensities also exists [28]. HCD is better suited for resolving lower mass fragment ions (e.g., immonium ions) than CID and can help determine precise locations of protein modifications (e.g., phosphorylation).

1.1.2.1 Instrumentation

The research presented in this dissertation was carried out using various Thermo Scientific Orbitrap mass spectrometers. These hybrid instruments combine mass analyzers of different types to benefit from the performance attributes of each [29]. In each of the instruments a combination of linear ion trap mass analyzer together with a downstream C-trap, Orbitrap mass analyzer, and HCD fragmentation module enable MS^n analyses allowing for information relevant to structural characteristics of the analyte to be captured. These instruments deliver very high analytical performance.

Analytical performance The analytical performance of the instrument can be assessed on a variety of dimensions with some of the most broadly important being: (1) mass accuracy, (2) resolving power, (3) dynamic range, and (4) speed of a data acquisition.

Mass accuracy Mass accuracy represents how closely an observed m/z measurement matches the theoretical m/z for the measured ion. The difference between

these two values is the measurement error which is usually expressed in parts per million (ppm). Orbitrap mass analyzers generally give measurements with mass accuracy of less than 10 ppm, while other ion trap mass analyzers (e.g., quadrupoles) may be in the 100 ppm to 1000 ppm range [30]. Higher mass accuracy allows one to limit the possible elemental compositions that may explain an observed signal in the mass spectrum and thus serves as a powerful "filter" useful in confirming the identity of a compound or identifying an unknown. As sample complexity increases this parameter becomes a critical consideration when conducting a mass spectrometry experiment.

Resolving power Resolving power refers to the ability of the mass spectrometer to resolve different ion species as distinct mass peaks in a mass spectrum. This parameter is especially important as sample complexity increases and the extent to which ion species may produce narrowly-separated signals or overlapping isotopic patterns in the mass spectrum becomes more likely.

Dynamic range Dynamic range describes the range in analyte concentrations across which the instrument is able to capture information. More specifically, it is the ratio between the maximum and minimum intensity for distinct spectral features that an instrument can produce. A higher dynamic range is important for capturing information from analytes that exist across a range of concentrations in the sample. This can be an especially critical parameter to consider when the sample to be analyzed has both very high and low abundance compounds.

Scan rate The speed of a data acquisition refers to the rate at which the mass spectrometer can record mass spectra. An instrument that can acquire mass spectra rapidly while maintaining high mass accuracy and resolution is desirable because

it enables one to observe more unique analytes (i.e., achieve greater coverage) and acquire multiple observations of a particular analyte (i.e., greater evidence for identification) in a given time frame. This parameter may also be referred to as the "scan rate" of the instrument and may be expressed in hertz (Hz) (i.e., cycles per second).

Liquid chromatography-MS Each instrument used in this dissertation was also coupled to an LC system with a nESI source allowing for automated and continuous sample analysis. Liquid chromatography-mass spectrometry (LC-MS) allows for separation of analytes prior to ionization and mass analysis. Sample molecules are delivered via a liquid mobile phase to a column (i.e., stationary phase) on which they adsorb. The composition of the mobile phase is then, typically, altered along a concentration gradient of a solvent or solute component that will compete or disrupt the degree to which the adsorbed sample molecules are retained on the column. The elution time of each molecular species depends upon its interaction with the stationary phase and current solvent composition. The time at which mass analysis occurs is recorded in the mass spectrometry data as the retention time (RTime).

With a high level of performance in each of these analytical aspects, Orbitrap mass spectrometers have become a widely used instrument of choice for bottom-up proteomic experiments and are used in all such analyses presented in this dissertation.

1.1.3 Bottom-up proteomics

The LC-MS analyses of proteins and proteomes are most commonly performed using a preparatory and analytical approach known as "bottom-up proteomics", which is also referred to as "shotgun proteomics" when the approach is performed on a mixture of proteins [31]. In this approach proteolytic digestion of the protein-containing

sample is performed prior to LC-MS separation and analysis. Most often digestion is accomplished using the enzyme trypsin which specifically cleaves at the C-terminal end of lysine and arginine amino acid residues. The basic sidechains of these residues and the N-terminal residue amine-group of the resulting peptides are protonated at acidic pH at which the reverse-phase liquid-chromatography separations take place. Although, trypsin has become the most popular protease for which digestion is carried out, various other proteases can be used to generate an alternative set of peptides that may serve a particular proteomic analysis better, this is further explored in Chapter 2. In recent years this technique has become capable of identifying whole proteomes with analysis times on the order of minutes to hours [32].

1.2 Structural Proteomics

Early use of the term "structural proteomics" referred to the project of characterizing the three-dimensional structures of all proteins at a genome-wide scale [33]. More recently the term is commonly used interchangeably with the term "structural mass spectrometry", which itself refers to mass spectrometry experiments used to derive structural information about the molecules under study [34]. This usage co-evolved with mass spectrometers becoming the dominant technology with which proteomics experiments are now carried out. Throughout the dissertation the term "structural proteomics" is used in this more recent sense.

Structural proteomics comprises a suite of techniques that use mass spectrometry to yield insight into various structural aspects (e.g., protein folds, conformational dynamics, and interactions) of the proteins under analysis. Protein structural information can be captured independently of analysis in an experimentally desirable buffer environment or sample context and then later read-out via LC-MS. Due to this

flexibility afforded in preparation and analysis structural proteomic approaches are applicable to a wide-variety of sample conditions and protein types. Some of these techniques include: hydrogen-deuterium exchange (HDX) which measures uptake of deuterium and can yield information related to solvent accessibility and hydrogen bonding of a protein; surface modification mass spectrometry and limited proteolysis which can probe solvent exposed regions of protein tertiary or quaternary structure; photoaffinity labeling which can be used to obtain structural information about ligand–protein interaction sites; and ion mobility which can yield information about the overall size and shape of a protein or protein complex. This dissertation will focus primarily on use and development of the technique of CL-MS which yields information on inter-residue distances within and between proteins.

1.2.1 Chemical crosslinking-combined with mass spectrometry

Chemical crosslinking has been used since at least the early 1970's together with gel electrophoresis to elucidate interactions between subunits of the ribosome [35, 36], translocases [37], and the proteasome [38]. CL-MS in its most general definition is defined as follows: "*non-covalent interactions or proximities within or between biomolecules are covalently fixed for their detection in an otherwise dissociative analytical process involving a mass spectrometer*" [39]. In the context of this dissertation, the term is used in a more exclusive sense which specifically involves the mass spectrometric analysis of crosslinked proteins and peptides. Here, crosslinks are covalent linkages between residues and as such contain information useful for gaining structural insights into a protein's three-dimensional structure if the linked residues can be experimentally identified. These crosslinks may already exist endogenously in the form of disulfide bonds or other less common covalent linkages (e.g., sulfilimine

bond [40, 41]). More often in these experiments they are introduced synthetically to capture interactions between proximal residues that would not otherwise form covalent linkages. Experiments utilizing CL-MS were initially demonstrated in the early 1990's and began gaining traction in the early 2000's [42, 43, 44, 45, 46, 47, 48]

1.2.1.1 Crosslinking-MS experimental workflow

The general workflow for a CL-MS experiment is to first obtain a protein sample in conditions that are amenable to the crosslinking chemistry to be used. This sample may be isolated (i.e., "purified") protein(s) (as in Chapters 2 and 3), whole cells and sub-cellular compartments (as in Chapter 4), or tissues [49]. A crosslinking reaction is then carried out using chemical, photo-reactive, or enzymatic reagents. Depending on the reagent used, a quench reagent may be introduced to stop the crosslinking reaction. Following this, excess crosslinker is removed via dialysis, gel electrophoresis, or gel-filtration. The proteins are then enzymatically digested with the enzyme trypsin which yields a complex mixture of peptides and crosslinked peptides. At this point, crosslinked peptides maybe enriched via a variety of methods such as affinity enrichment (AE) [50, 51], size-exclusion chromatography (SEC) [52], or strong cation exchange (SCX) [53, 54]. A bottom-up proteomics approach can then be used to analyze the crosslink-containing peptide mixture via LC-MS. Data generated from this experiment can then be analyzed using algorithms specifically designed to identify crosslinked peptides. Aspects of these algorithms are developed in Chapter 2 and Chapter 4.

There are a wide variety of crosslinking reagents available. These include so-called "zero-length" crosslinkers which covalently couple amino acid residues without incorporating any additional atoms. However, more commonly used in CL-MS analyses are crosslinking reagents which consist of at least two reactive groups

capable of covalently coupling residues and separated by a "spacer-group". This intermediating section can be as simple as an alkyl chain or more complex with multi-functional designs. Additional functionalities designed into a crosslinking molecule may include MS-cleavable groups, enrichable groups, and isotopically-labelled groups. A method for CL-MS analysis of a complex sample, yeast mitochondria, that utilizes one of these multi-functional crosslinking reagents is presented in Chapter 4.

1.2.1.2 Crosslinking-MS data analysis

A bottom-up LC-MS/MS analysis of a protein digest sample will, typically, yield data that consists of a set of MS^1 and MS^2 spectra. These spectra are then evaluated with respect to a reference protein sequence database in order to identify peptides. This evaluation is accomplished by comparison of the experimental MS^2 spectrum data (e.g., precursor mass, fragment ion m/zs) to theoretical MS^2 spectra generated for peptides that can be expected to exist in the sample (given a specified proteolytic digest) and match (within tolerance) the observed precursor ion mass. The same analysis of a crosslinked protein digest sample will require a specialized algorithm to evaluate the data to identify crosslinked peptides. The main challenge that arises when trying to identify crosslinked peptides from this data is that there exists the possibility that any given peptide may be coupled to any other given peptide in the protein sequence database under consideration. The implication of this fact is that the number (n) of potential crosslinked peptide species one must consider grows exponentially (e.g., $(n^2 + n)/2$ for crosslinked peptide pairs) as the protein sequence database to be considered increases in size. It has been calculated that it requires just 50 proteins to reach a theoretical number of unique crosslinked peptide species that would exceed the number of peptides in the entire human proteome [55].

Several algorithmic approaches to dealing with this issue exist ranging from

brute-force enumeration [56, 57] to heuristic-based search [58, 59, 60]. On current consumer-grade computational hardware brute-force approaches are feasible when the sample data and sequence database to be analyzed are relatively simple (e.g., less than 50 proteins). However, as database size increases and whole proteomes are considered or protease specificity is relaxed then alternative approaches should, or must, be used. Each of these approaches is used to meet different purposes throughout this dissertation.

In addition to algorithmic approaches, experimental approaches exist that can mitigate or bypass entirely the issue of an exponentially expanded search-space by releasing the individual peptides constituting the crosslinked peptide species and acquiring MS^3 spectra for each [61, 62, 63, 64]. Both algorithmic and experimental approaches toward overcoming this challenge are applied and discussed in Chapter 4.

One of the major challenges in bottom-up proteomics studies is in the correct identification of peptides from the vast amount of collected MS^2 data. Although peptide-spectrum match (PSM) scoring functions aim to maximize precision in peptide-spectral matching there still exists some possibility of returning incorrect PSM identifications. Several strategies to estimate the confidence a particular set of identifications from a database search have been developed. One of the most common strategies for this is referred to as "Target-Decoy". This strategy involves searching the MS^2 data with both expected protein sequences ("targets") and reversed or scrambled protein sequences which are not to be expected ("decoys"). By observing the proportion of identifications returned from a database search that are matched to decoy sequences it is possible to estimate the false-discovery rate (FDR) (a.k.a., false positive rate) for a particular set of identifications. Chapter 4 uses this strategy together with a machine learning algorithm [65] as a foundation upon which a method for improving the FDR estimates in CL-MS experiments is explored and developed.

1.3 Motivations, research questions, and objectives

CL-MS has developed significantly since the first demonstrations of the technique in the early 2000's [43, 44, 45, 47, 66]. Despite this, at the outset of the work that is now presented in this dissertation there still existed, in a variety of contexts, several challenges to the successful application of the technique and significant room for further methodological and analytical development. My overarching objective is to explore, develop, and apply MS-based structural proteomic techniques, with a primary focus on CL-MS, toward understanding protein structure and PPI networks.

I begin by developing and describing a method for identifying naturally occurring crosslinks, in the form of endogenous disulfide bonds, using MS in Chapter 2. The method aims to improve upon the coverage achieved in typical CL-MS analyses by employing an algorithm capable of higher-order crosslink identification and a non-specific proteolytic digest approach. In Chapter 3, I describe an application of CL-MS in which crosslinks are synthetically introduced to a protein and, following MS analysis, information about the covalently linked residue-residue pairs is utilized in molecular dynamics simulations to create a *de novo* structural model of the intrinsically disordered protein tau. The study is amongst the first demonstrations of the crosslinking constraint-guided discrete molecular dynamics (CL-DMD) approach [67, 68, 69, 70] and aims to deliver experimentally constrained all-atom structural models of the monomeric full-length (2N4R) human tau protein, of which few other examples currently exist [71, 72]. Finally, in Chapter 4, I develop and describe improvements to the CL-MS technique in its application toward elucidation of large interactomes. Here I aim to specifically address challenges with respect to the detection (i.e., MS^1 precursor ion signal observation), acquisition (i.e., MS^2

collection of fragment ion information), and identification (i.e., confident assignment of PSMs) of crosslinks.

The main contributions detailed in this dissertation are the development of MS-based structural proteomic methodologies for the characterization of protein structure and interactions. The methodologies developed and applied in each of the following chapters can be referenced as a foundation to be built upon in future structural proteomic studies that aim to access structural information on proteins from monomers to larger assemblies and entire proteome PPI networks.

CHAPTER

TWO

COMPREHENSIVE IDENTIFICATION OF DISULFIDE BONDS USING NON-SPECIFIC PROTEINASE K DIGESTION AND CID-CLEAVABLE CROSSLINKING ANALYSIS METHODOLOGY FOR ORBITRAP LC/ESI-MS/MS DATA

This chapter was adapted from the publication:

Makepeace, Karl A.T., Jason J. Serpa, Evgeniy V. Petrotchenko, and Christoph H. Borchers. "Comprehensive identification of disulfide bonds using non-specific proteinase K digestion and CID-cleavable crosslinking analysis methodology for Orbitrap LC/ESI-MS/MS data." *Methods* 89 (2015): 74-78.

DOI: 10.1016/j.ymeth.2015.02.021

Contribution disclosure: Work presented in this chapter was carried out in the laboratory of Christoph Borchers. Karl Makepeace and Jason Serpa performed all experiments. The software that was created and used for the identification of disulfide bonds in mass spectrometry data was developed by Evgeniy Petrotchenko. All analyses of the data were completed by Karl Makepeace. All authors contributed to development of the methodology described in the chapter. Karl Makepeace and Evgeniy Petrotchenko wrote the first draft of the manuscript. Christoph Borchers oversaw the project and all authors contributed to the final version of the published manuscript.

2.1 Introduction

Disulfide bonds are naturally occurring post-translational modifications in proteins, which are important for stabilization of the tertiary protein structure. This modification involves the formation of a S–S bond between the thiol groups of two cysteine residues. Disulfide bonds carry valuable structural information about a protein. Because the Cys–Cys bond is a zero-length crosslink, forming a covalent linkage between two amino acid side-chains directly while incorparting no intermediate (i.e., "spacer-arm") atoms, disulfide bonds can serve as valuable short distance constraints in protein structure determination and molecular modeling.

Disulfide bonds have traditionally been determined by amino acid sequencing and using mass spectrometry [73, 74]. One of the approaches to the identification of disulfide bonds by mass spectrometry is the LC-MS analysis of the enzymatic protein digest absent the disulfide-breaking reduction step which is typically performed. An attractive feature of the MS analysis of the Cys–Cys disulfides is that the R1-C-S–S-C-R2 linkage undergoes CID-fragmentation, producing a specific pattern of cleavage

products resulting from the fragmentation of the C–S and S–S bonds. This cleavage pattern has been exploited for the identification of disulfide-bonded peptides by MALDI- and ESI-MS/MS [75, 76]. The DBond program, which takes into account such cleavage products, was developed for the analysis of the MS^2 spectra of the disulfide-linked peptides, obtained with high-specificity proteolytic enzymes [77]. The Borchers research group has previously reported the application of the non-specific proteolytic enzyme proteinase K for the comprehensive identification of the inter-peptide crosslinks [78]. Digestion of crosslinked proteins with proteinase K produces a series of inter-peptide crosslinks consisting of short peptides that are related by sequence overlap, which increase confidence in the discovery and identification of crosslinks. Digestion with proteinase K can be crucial for the determination of the disulfide bonds in proteins resistant to digest with a standard protease (e.g., trypsin) [79, 80, 81] and proteins with complex disulfide connectivities.

In general, the current methodology for the MS-based crosslink determination consists of digestion of the crosslinked proteins with proteolytic enzymes and subsequent analysis of the obtained peptides with LC-ESI-MS/MS using high performance mass spectrometers, such as Orbitrap-based hybrid MS systems. For easier detection and identification of the crosslinked peptides by mass spectrometry, the Borchers group had previously developed the isotopically-coded affinity-enrichable CID-cleavable crosslinker cyanurbiotin-dimercaptopropionyl-succinimide (CBDPS)-H_8/D_8 [82] which can be used in combination with non-specific digestion of the crosslinked proteins with proteinase K [78]. The Borchers group also developed algorithms and software tools to facilitate analyses using such a crosslinker [83, 57]. The continued development and application of the CBDPS-H_8/D_8 crosslinker is the topic of Chapter 4. The CID-cleavable sites in CBDPS are C–S bonds, which are the same bonds that are preferentially cleaved by CID in disulfide-containing peptides

(Figure 2.1).

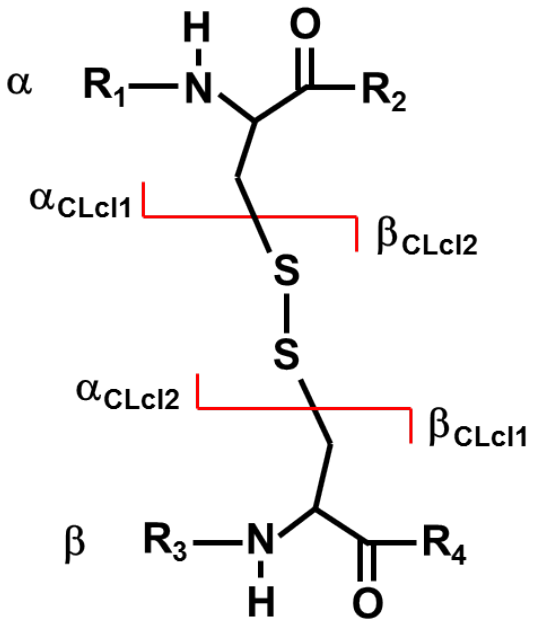


Figure 2.1: Diagram of the CID-cleavage of the Cys–Cys crosslink.
Disulfide bond fragments, which were used in the analysis, are labeled as CLcl1 and CLcl2 according to the DXMSMS Match program nomenclature [57].

Due to this similarity, the Borchers lab crosslinking analysis "toolkit" could be applied to the determination of disulfide crosslinks in proteins. Unfortunately, disulfide-crosslinked peptides are not isotopically-coded nor can they be specifically enriched. However, the natural occurrence of these disulfides in a digest is approximately 100 %, which is much higher than the occurrence of crosslinks introduced chemically with crosslinking reagents. Thus, this seemed appropriate to apply CID-cleavable crosslinking analysis methodology for the analysis of the disulfide crosslinked peptides. A similar approach—specifically designed for the analysis of the disulfide-bonded tryptic peptides—was also developed independently from chemical crosslinking applications [77].

Here, I present the further development of an approach designed for use with non-specific proteinase K digestion, along with software for the analysis of second-order

(i.e., three inter-crosslinked peptides) disulfide crosslinks. The software has options for the use of digestion with non-specific enzymes and second-order disulfide bond analysis. Finally, I describe the application of these tools to the analysis of disulfide bonds, by treating them as zero-length crosslinks. The software suite has also now been supplemented with an additional program for the analysis of second-order crosslinks, enabling robust and comprehensive disulfide bond determination.

2.2 Materials and Methods

All chemicals were from Sigma–Aldrich unless noted otherwise. Human monoclonal IgG1 antibody was from Genscript, Trypanosoma CISSA (Q26806), PSSA-2 (G0UXP9) and CESP (A8QZK1) were provided by Dr. Martin Boulanger (University of Victoria).

2.2.1 Proteolytic digestion

Proteins were digested in phosphate buffered saline (PBS) pH 7.2 with trypsin (Promega) for 18 h or proteinase K (Worthington) for 2 h at 37°C, using a 1:10 enzyme:substrate ratio.

2.2.2 Mass spectrometry analysis

Mass spectrometric analysis was performed as previously reported for protein crosslinking analysis [84], except that the sample was not reduced prior to HPLC separation. A nano-HPLC system (Easy-nLC II, Thermo Fisher Scientific, Bremen, Germany) was coupled to the ESI source of an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific). Samples were injected onto a 100 μm outer diameter (OD), 360 μm OD trap column packed with Magic C18AQ (Bruker-Michrom,

Auburn, CA), 100 Å, 5 μm pore size (prepared in-house), and desalted by washing for 15 min with 0.1 % formic acid (FA). The column was equilibrated with 95 % solvent A (2 % acetonitrile (ACN) and 98 % water, both containing 0.1 % FA) before the peptides were separated with a 70 min ACN:water gradient. The gradient used was 0–60 min: 4–40 % B, 60–62 min: 40–80 % B, 62–70 min: 80 % B, with solvent B containing 90 % ACN and 10 % water, both 0.1 % FA. Separations were done on a 75 μm OD, 360 μm OD analytical column packed (in-house) with Magic C18AQ, 100 Å particle size, 5 μm pore size, and with an IntegraFrit™ (New Objective Inc., Woburn, MA).

2.2.3 Mass spectrometry data analysis

MS data were collected with a data-dependent acquisition in which the six most intense peaks in each full scan mass spectrum (MS^1) scan were selected for fragmentation. Dynamic exclusion was set to 60 seconds with a repeat count of 2. MS^1 scans (m/z 400–2000 range) and MS^2 scans were acquired at 60,000 and 30,000 resolution, respectively. MS^2 fragmentation was performed by CID at normalized collision energy of 35 %. Fourier transform mass spectrometry (FTMS) full scan automatic gain control (AGC) target was 1,000,000 and FTMS MS^n AGC target was 100,000.

Proteome Discoverer (ver. 1.4.0.288) was used to generate ".MGF" files from ".RAW" files. DXMSMS Match [57] of the Isotopically-Coded Cleavable Crosslinking Analysis Software Suite (ICC-CLASS) [56] was used to identify crosslinked peptides. DXMSMS Match 2^{nd} Level was used for the search of the second-order crosslinks (three peptides with two crosslinking bridges).

2.3 Results & Discussion

There are several connectivities possible for the disulfide bond connected peptides (cysteine crosslinks) which are obtained by enzymatic digestion of disulfide-containing proteins. Cases which involve one or two peptides, can be handled with the existing DXMSMS Match software, and cases with three inter-linked peptides can now be analyzed using the DXMSMS Match 2^{nd} Level program introduced here.

Thus, the overall analytical strategy would be to assign crosslinks that contain disulfide bonds in one peptide (zero-order), and two peptides (first-order) with DXMSMS Match, and then to use DXMSMS Match 2^{nd} Level to assign crosslinks where three peptides are connected by disulfide bonds (second-order) (Figure 2.2). When using the DXMSMS Match program, the Mip parameter (i.e., the crosslinker mass value for the inter-peptide crosslink formation), which is used to assign crosslinks containing a single disulfide bond, is set to -3.02293 because two hydrogen atoms with mass of 1.00783 Da are lost due to disulfide bond formation and one proton with mass of 1.00728 Da is lost due to the crosslinking of two peptides with charge state 1+ into one inter-peptide crosslink with charge state 1+. For cases where two disulfide bonds are present in a dipeptide, a Mip value of -5.03723 is used. Zero-order crosslinks are reported as the matched peptide sequence together with "(-)i(-)" as the designation for intra-peptide crosslinks in the program output. First-order crosslinks are reported as both matched peptides in the program output. When using the DXMSMS Match 2^{nd} Level program, the Mip parameter that is used to assign crosslinks containing single disulfide bonds between each outer peptide and the middle peptide (i.e., 2 disulfide bonds) is -3.02293. To assign cases where an additional disulfide bond is present (i.e., 3 disulfide bonds), a Mip of -5.03723 is used.

There is a high computational load for the analysis of digests resulting from the

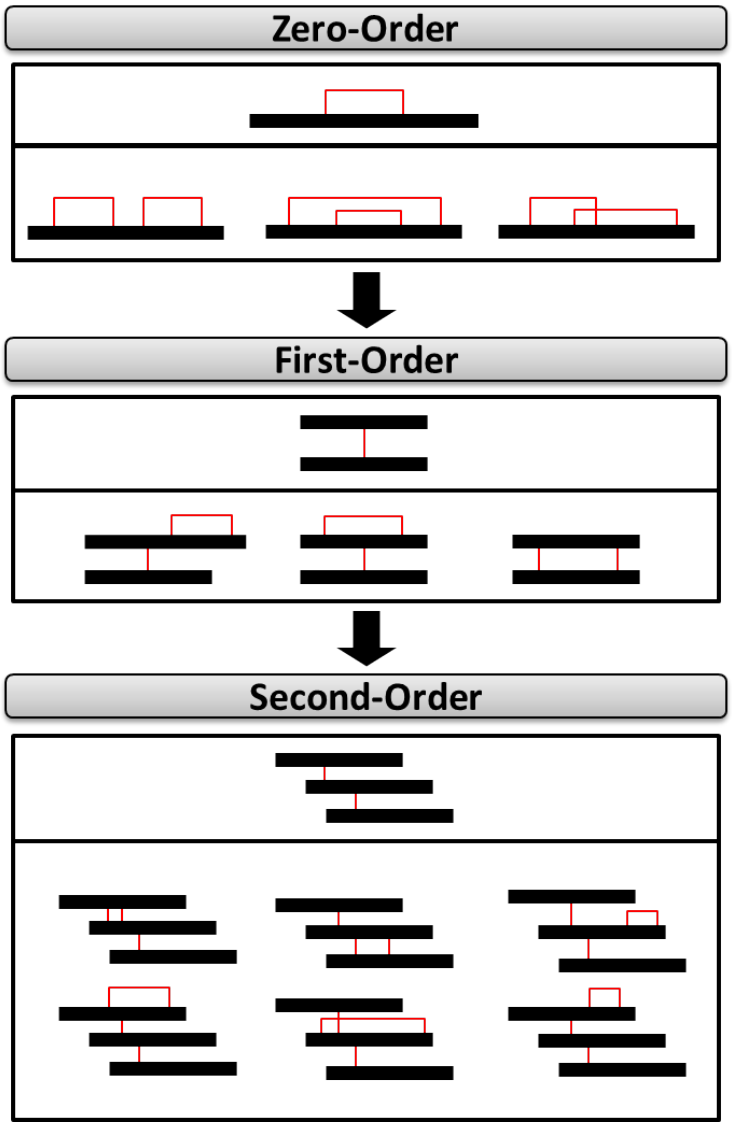


Figure 2.2: Disulfide bond determination data analysis workflow.
Zero-order (intra-peptide) and first-order (two bonded peptides) disulfide analysis with DXMSMS Match are followed by second-order (three bonded peptides) disulfide analysis with DXMSMS Match 2^{nd} Level.

use of non-specific enzymes in the DXMSMS Match 2^{nd} Level program, because the number of possible second-order crosslinked peptides formed from n peptides is proportional to n^3. Therefore, it is practical to use this program after the regular DXMSMS Match program has assigned the zero- and first-order disulfides (one and two peptide crosslinks), and to restrict the search space for potential

second-order disulfides (three peptide crosslinks) to those regions of the sequence immediately adjacent to cysteine residues that are still unassigned to disulfide bonds (approximately 20 residues on each side).

It is important to perform the MS^2 acquisition with high mass-accuracy (i.e., by using an instrument such as the Orbitrap mass analyzer) in order to keep the fragment mass match tolerances low and minimize the number of incorrect fragment ion matches. This becomes especially important when analyzing data from proteinase K digests as the number of crosslink predictions for each precursor is generally much greater than with site-specific proteases.

I have validated this strategy on a number of model disulfide-containing proteins and have applied it to the determination of the disulfide bonding pattern in not-yet-characterized proteins (Table 2.1). For those proteins that have disulfide connectivities that have been reported in the literature or noted in the UniProt database (i.e., BSA, fibrinogen, IgG1, insulin, lysozyme, and RNase S), I have identified and reconfirmed these connectivities with this approach and software. For those proteins without known disulfide connectivities (i.e., CESP, CISSA, and PSSA-2), I have identified novel disulfide connectivities. Overall, I was able to identify most of the disulfide bonds in the proteins tested. In the case of fibrinogen, the masses of four disulfide-containing peptides were outside the analysis mass range due to being higher than second-order (i.e., more than three inter-crosslinked peptides). Digestion with proteinase K provided more disulfide identifications, probably due to the fact that the digestion sites were closer to the cysteine residues. As noted before in previous work from the Borchers lab, proteinase K is a robust non-specific proteolytic enzyme producing short peptides, which, when crosslinked, fall into the ideal mass range for the mass spectrometric analysis [78]. This is even more apparent for second-order crosslinks. Digestion results in "families" of related peptides, which additionally

benefits the assignment process by producing multiple distinct confirmations of the crosslinked sites (Figure 2.3).

Table 2.1: Examples of disulfide crosslink determination in proteins with known and unknown disulfide connectivities.

Values represent the number of disulfides identified in each step of the analysis (0^{th}, 1^{st}, 2^{nd} order); T- and PK-denote digestion with trypsin or proteinase K, respectively. "n/a" indicates that the analysis was not performed (due to only proteinase K digests, not trypsin, being applied for analysis of a protein or to all disulfides having been identified in lower-order crosslink analyses rendering higher-order analysis unnecessary for that protein).

Protein	UniProt	0^{th} Order		1^{st} Order		2^{nd} Order		Total found			Total known or
	Accession No.	T	PK	T	PK	T	PK	T	PK	Total	possible [ref]
BSA	P02769	0	1	14	18	2	n/a	16	18	18	18* [85]
CESP	A8QZK1	n/a	0	n/a	3	n/a	0	n/a	3	3	3 [n/a]
CISSA	Q26806	3	1	1	3	n/a	n/a	4	4	4	4 [n/a]
Fibrinogen	P02671, P02675, P02679	n/a	2	n/a	10	n/a	0	n/a	12	12	16 [86]
IgG1	A0A087WTX5, A0A087WV47	0	0	5	5	0	3	5	8	9	9 [87]
Insulin	P01308	n/a	0	n/a	3	n/a	n/a	n/a	3	3	3 [88]
Lysozyme	P00698	n/a	0	n/a	4	n/a	2	n/a	4	4	4 [89]
PSSA-2	G0UXP9	3	1	1	3	n/a	n/a	4	4	4	4 [n/a]
RNase S	P61823	n/a	0	n/a	3	n/a	0	n/a	3	3	4 [90]

* There is one free cysteine (BSA Cys58) in the protein monomer.

The determination of peptides crosslinked by Cys–Cys bonds after proteolytic digestion of the proteins, is inherently susceptible to the recombination of the disulfide containing peptides ("scrambling"), especially when digestion is performed at neutral to high pH, as, for example, is required for optimal digestion with trypsin. Lowering the pH to pH 6.5–6.8 minimizes this effect when tryptic digestion is used ([91], presented by A. Heck at 4^{th} Symposium on Structural Proteomics, Antwerp, Belgium, 2014). Digestion at pH 6.5–6.8 range can also be used for proteinase K, as this protease is fully active in this pH range. Alternatively, the described approach has also successfully been used with pepsin digestion at low pH, which also minimizes

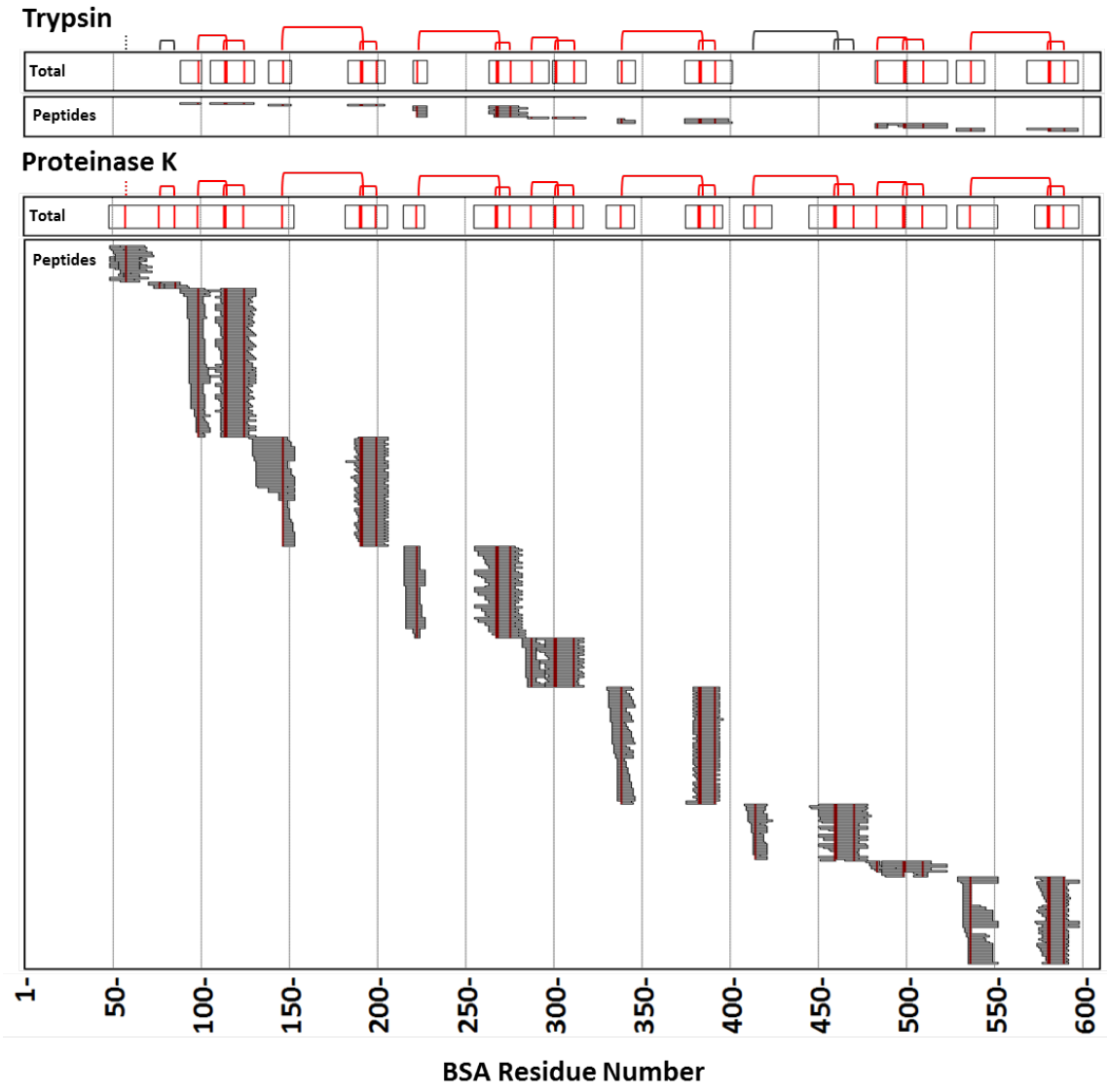


Figure 2.3: An example of disulfide determination in BSA. The tryptic and proteinase K peptides which form the identified disulfide crosslinks are shown as horizontal bars.

Cysteine residue positions are highlighted in red. Disulfide connectivity is shown as brackets, identified and non-identified disulfide bonds are shown in red and black, respectively. Panels immediately below the brackets represent total coverage of the cysteine-containing regions of the protein sequence, which were identified by the analysis of the respective digests.

disulfide bond recombination.

In my analysis I observed disulfides which are not present in the known X-ray

crystallography or NMR-based protein structures. I believe this is due to a small amount of pre-processing scrambling that occurred in the samples. In most of these cases, the intensities of the corresponding signals and their score values were low, and choices for the alternative assignments were made based on signal intensities and score values. As a first filtering step, I used an empirical scoring function that reports the product of % matched b- and y-ions, crosslinker-cleavage ions, % matched spectrum intensity, and number of specific crosslinker-cleavage ions matched. There was no automatic FDR evaluation, but "true positive" assignments were selected in a similar manner to the $< 1\,\%$ FDR evaluation, which was done for analogous CID-cleavable isotopically-coded crosslinkers [84], i.e., $> 30\,\%$ of fragment ions matched and $> 5\,\%$ of total intensity matched.

In summary, a CID-cleavable crosslinking analysis methodology employing robust non-specific proteolytic digestion and bottom-up mass spectrometric analysis with Orbitrap LC-ESI-MS/MS can be successfully applied for the comprehensive identification of the disulfide bonds in proteins. The approach presented here can also be satisfactorily applied for the characterization of the proper disulfide connectivities in protein therapeutics, "biosimilars", as illustrated by the insulin and IgG1 examples. The combination of this approach with the quantitative characterization of the disulfide bonding red/ox status in these molecules is a potential avenue of further development.

CHAPTER

THREE

INSIGHT INTO THE STRUCTURE OF THE "UNSTRUCTURED" TAU PROTEIN

This chapter was adapted from the publication:

Popov, Konstantin I., Karl A.T. Makepeace, Evgeniy V. Petrotchenko, Nikolay V. Dokholyan, and Christoph H. Borchers. "Insight into the structure of the "unstructured" tau protein." *Structure* 27, no. 11 (2019): 1710-1715.

DOI: 10.1016/j.str.2019.09.003

Contribution disclosure: Work presented in this chapter was carried out in the laboratories of Christoph Borchers, and Nikolay Dokholyan. The project was initially conceived by Evgeniy Petrotchenko, Christoph Borchers, and Nikolay Dokholyan. All authors contributed to methodological design. All structural proteomic mass spectrometry experiments and related data analyses presented in this chapter were conducted by Karl Makepeace. Molecular dynamics simulations were performed by Konstantin Popov. Karl Makepeace, Evgeniy Petrotchenko, and Konstantin Popov wrote the first draft of the manuscript. Christoph Borchers and Nikolay Dokholyan oversaw the project and all authors contributed to the final version of the published manuscript.

3.1 Introduction

The tau protein is involved in the pathogenesis of protein misfolding-related neurodegenerative diseases, including Alzheimer's disease [92, 93]. A misfolding event leads to the formation of oligomers and eventually neurofibrillary tangles. The underlying molecular mechanisms leading to neurotoxicity and disease progression are not well understood, however, oligomeric species are suspected to be toxic and eventually lead to the death of neuronal cells [94]. Moreover, a prion-like spread of this aggregation via the conversion of native protein molecules by toxic oligomers has been suggested to be involved in the pathogenesis of neurodegenerative diseases [95]. Recent longitudinal investigations using tau-specific positron emission tomography (PET) radiotracers suggest tau pathology is a major driver of local neurodegeneration [96]. Native tau is considered to be an intrinsically disordered protein (IDP), although evidence of some globular structure can be found in the literature [97]. How this native form converts to give rise to the recently described panoply of tau

aggregate morphologies characteristic of specific tauopathic diseases [98, 99, 100, 101] is not fully understood. Thus, the native structure of tau in solution may serve as a basis for understanding the various misfolding and oligomerization pathways which may give rise to the panoply of tau aggregate morphologies that have recently been described.

Traditional all-atom molecular dynamics simulations of the tau structure are hampered by the large size of the full-length tau protein (441 residues) [102]. A number of biophysical methods, such as nuclear magnetic resonance (NMR), electron paramagnetic resonance (EPR), Förster resonance energy transfer (FRET), and small-angle X-ray scattering (SAXS)—in combination with computational methods—has been applied to the study of IDPs, including the structure of tau in solution [103, 104, 102, 105]. Tau conformational ensembles have been proposed based on long-distance paramagnetic relaxation enhancement (PRE)-NMR, single-molecule FRET data, or time-resolved ESI-MS with HDX and molecular modeling [71, 72]. Nuclear Overhauser effect (NOE)-derived short-distance restrained molecular dynamics simulation are traditionally used for the determination and refinement of protein structures by NMR [106, 107]. Incorporation of the crosslinking data as experimental distance constraints into molecular dynamics protein structure modeling is an alternative promising approach for protein structure determination [108]. Previously, together with collaborators, the Borchers research group developed a method for determining protein structures, called short-distance crosslinking constraint-guided discrete molecular dynamics (CL-DMD) simulations, in which the folding process is guided by short-distance experimental constraints that are incorporated into the DMD force-field energy function [67]. Adding short-distance crosslinking constraints to DMD simulations results in a reduction of the possible conformational space and allows simulations to converge more rapidly to folded protein conformations [109].

This approach was tested on proteins with significant amounts of secondary structure in their native conformations and, more recently, for the prediction of the conformational ensemble of the 140-residue-long IDP, α-synuclein [68]. The conformational flexibility of IDPs, such as α-synuclein and tau, brings additional challenges to the computational process [110], because, in these cases, proteins exist as a collection of inter-converting conformational states, and the crosslinking data represent multiple conformations of a protein rather than a single structure.

The Medusa force field [111, 112, 113], which is utilized in DMD simulations, is discretized to mimic continuous inter-atom potentials, and can readily integrate any additional potentials, such as pairwise distance constraints [109, 114] and solvent accessibility information [115, 116]. Using CL-DMD simulations, conformational ensembles can be generated that satisfy the optimal number of constraints, thereby naturally helping to resolve possibly conflicting experimentally derived constraints. Previous work has shown that CL-DMD simulations are a viable computational platform for the structural analysis of IDPs, and α-synuclein in particular [68, 117].

Here, the CL-DMD approach is extended to determine the conformational ensembles of the full-length human tau protein in solution. During this process, tau was crosslinked with a panel of short-range crosslinkers (spacer length < 7 Å), the crosslinked proteins were enzymatically digested, the crosslinked residues were determined by LC-MS/MS analysis, and the resulting information on inter-residue distances was introduced into the DMD force field as external constraints. To experimentally validate the predicted structures, tau was analyzed using surface modification and long-distance crosslinking.

3.2 Materials and Methods

All reagents were from Sigma-Aldrich unless otherwise noted and except the following: DSS, DSA, DSG crosslinking reagents (Creative Molecules Inc.), SDA (Thermo Fisher Scientific).

3.2.1 Expression and Purification of the Tau Protein

The pET28a plasmid vector was received as a gift from Prof. Dr. David Vocadlo (Simon Fraser University), and containing a synthetic gene corresponding to the human 2N4R tau isoform (441-residue isoform) with an additional 21-residue N-terminal fusion tag containing 6x-His and a thrombin cleavage site (MGSSHHHHHHSSGLVPRGSHM). The plasmid was transfected into *E. coli* BL21(DE3) bacteria for protein expression. Cell cultures were grown in lysogeny broth (LB) media at 37°C, with 150 rpm shaking, to an optical density of 0.8–1.0 at 600 nm. Protein expression was induced by the addition of isopropyl β-D-1-thiogalactopyranoside (IPTG) to give a final concentration of 0.5 or 1.0 mM. The tau protein was then overexpressed overnight ($\sim$ 16 hours) or for 4 hours at ambient room temperature (16–20°C), with 150 rpm shaking.

Cells were harvested by centrifugation at 4000 RCF for 10 min at 4°C. The supernatant was discarded and the cells were resuspended in 30 mL of ice-cold phosphate buffered saline (PBS) at pH 7.4. The centrifugation step was repeated, the supernatant was discarded, and the pelleted cells were then frozen at −80°C until needed. Frozen cell pellets were thawed and resuspended in 30 mL of lysis buffer (20 mM sodium phosphate, 500 mM NaCl, 5 mM imidazole, pH 7.4), and two Roche cOmplete™ protease inhibitor cocktail tablets were added, along with phenylmethylsulfonyl fluoride (PMSF) to give a final concentration of 1 mM. The

resuspended cells were then sonicated using a Misonix S-3000 with six cycles of 20 seconds pulsing (power setting = 5), followed by a 40 second rest period. Cellular debris was removed by centrifugation at 13,000 rpm using a JA-20 rotor in a Beckman L8-M Ultracentrifuge for 30 min at 4°C. The supernatant was loaded onto immobilized Nickel-nitrilotriacetic acid (Ni-NTA) beads (1 mL bed volume). The column was washed with 40 mL wash buffer (20 mM sodium phosphate, 500 mM NaCl, 10 mM imidazole, pH 7.4) and the protein was subsequently eluted with 5 separate additions of 2 mL elution buffer (20 mM sodium phosphate, 500 mM NaCl, 250 mM imidazole, pH 7.4). Elutions containing tau protein were pooled and subsequently concentrated, and buffer exchanged into PBS pH 7.4 using Amicon Ultra-15 centrifugal filters (10,000 NMWL).

In preparations where significant tau proteolytic breakdown fragments were observed in the Ni-NTA elution fractions, the concentrated and buffer-exchanged sample was subsequently loaded onto a Superdex™ 200 10/300 GL (GE Healthcare) column using a 500 μL capillary injection loop. Protein was eluted from the column with PBS (pH 7.4) buffer at a flow rate of 0.25 mL/min with fraction collection every 3 min (0.75 mL/fraction). Fractions containing intact tau protein with few or no proteolytic fragments were pooled and concentrated as described above.

3.2.2 Crosslinking

Samples were crosslinked by the addition of DSS (Creative Molecules Inc.), DSA-$^{12}C_6/^{13}C_6$ (Creative Molecules Inc.), DSG-H_6/D_6 (Creative Molecules Inc.), or SDA (Thermo Fisher Scientific) to final concentrations ranging from 0.1–1 mM (DSS, DSA, and DSG) or 1–5 mM (SDA), at room temperature for 15 min. Reactions were quenched by adding ammonium bicarbonate (ABC) to a final concentration of 10 mM.

SDA crosslinking reaction mixtures were incubated for 10 min in the dark to allow the NHS-ester reaction to take place, followed by 10 min of UV irradiation under a 25 W UV lamp (Model UVGL-58 Mineralight lamp, UVG) with a 366 nm wavelength filter. SDA reaction mixtures were quenched with 10 mM ammonium bicarbonate. A portion of each crosslinking reaction mixture was checked by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) gel to see the extent of potential intermolecular crosslinked products. Aliquots were subsequently run on SDS-PAGE and monomeric tau bands were excised and in-gel digested with trypsin [118]. The recovered protein digests were lyophilized to dryness and stored at −80°C until analysis by LC-MS, at which time they were resuspended in 0.1 % FA containing 10 mM tris(2-carboxyethyl)phosphine (TCEP) prior to autosampler loading.

3.2.3 LC-MS/MS Analysis

Mass spectrometric analysis was then performed using a nano-HPLC system (Easy-nLC II, Thermo Fisher Scientific), coupled to the ESI-source of an LTQ Orbitrap Velos or Fusion (Thermo Fisher Scientific), using conditions described previously [67]. Briefly, samples were injected onto a 100 μm ID, 360 μm OD trap column packed in-house with Magic C18AQ 100 Å, 5-μm pore size (Bruker-Michrom, Auburn, CA), and desalted by washing with solvent A (2 % ACN:98 % water, both containing 0.1 % FA). Peptides were separated with a 60-min gradient as follows: (0–60 min: 4–40 % solvent B (90 % ACN, 10 % water, 0.1 % FA, 60–62 min: 40–80 % B, 62–70 min: 80 % B), on a 75-μm ID, 360-μm OD analytical column packed with Magic C18AQ 100 Å, 5-μm pore size (prepared in-house), with IntegraFrit™ (New Objective Inc., Woburn, MA) and equilibrated with solvent A. MS data were acquired using a data-dependent acquisition (DDA) method. The DDA also utilized dynamic exclusion, with an exclusion window of 10 ppm and exclusion duration of 60 seconds. MS^1 and

MS^2 events used 60,000- and 30,000-resolution FTMS scans, respectively, with a scan range of m/z 400–2000 in the MS scan. For MS/MS, the CID collision energy was set to 35 %. Crosslinking data were analyzed using Kojak [58] and Percolator [65] using filtering criteria of Percolator q-value $\leq$ 0.01, score $\geq$ 1.675, population type = "intra", ppm error $\pm$ 2.5.

3.2.4 Surface Modification

Chemical surface-modification data in form of photoreactive dead-ends were obtained from SDA crosslinking data sets. Amino acid residues modified with the photoreactive moiety of the crosslinker, and the hydrolyzed amino-reactive moiety of the crosslinker (i.e., a mass addition of 100.05 Da), were considered.

3.2.5 Discrete Molecular Dynamics Modeling

CL-DMD simulations were performed according to the protocol described in previous work [67]. Briefly, in order to incorporate the experimental data for inter-residue distances between specific atoms into the DMD simulations [114], a series of well-shape potentials is introduced that energetically penalize atoms whose interatomic distance do not satisfy the experimentally-determined inter-atom proximity constraints. The widths of these potentials are determined by the crosslinker spacer length and the side-chain flexibility [67]. Starting from a completely unfolded structure of the tau molecule, an all-atom replica exchange (REX) [119] simulation of the protein is performed where 24 replicas with temperatures equally distributed in the range from 0.375 to 0.605 kcal/(mol k_B), are run for 6×10^6 DMD time-steps. The first 2×10^6 time-steps of system equilibration are discarded during the data analysis. In addition, all of the structures among all of the simulation trajectories are ranked, and the ones with the lowest 10 % of the energies, as determined by the DMD

Medusa force field function [120] are selected. These structures were then clustered using the GROMACS distance-based algorithm [121]. The centroids of the most populated clusters were selected as the representative model of the tau protein structure. Because there may be no distinct lowest-energy state for the protein, picking just one of these states would potentially introduce a bias to the structure predictions. Thus, all of these low energy states are presented here as predicted models of the tau conformational ensemble.

3.2.6 Unconstrained Molecular Dynamics Modeling

Unconstrained MD simulations were performed in the GROMACS 2018 package [121, 122] using CHARMM36 force field [123]. Each structure was solvated using a transferable intermolecular potential with 3 points (TIP3P) water model and neutralized with sodium and chloride ions. Particle mesh Ewald (PME) was used for long-range electrostatic interactions with a 10 Å cutoff for non-bonded interactions. Systems were initially equilibrated using a Nosé–Hoover thermostat (NVT) and were later subjected to 4 heating-cooling cycles during a 4-ns simulated annealing procedure. During these cycles the NVT equilibrated system (both the protein and the solvent) were linearly heated from 300 K to 325 K for 0.5 ns and then cooled back to 300 K during next 0.5 ns. This procedure was repeated 4 times. After that, the systems were further equilibrated using isothermal-isobaric Nosé–Hoover thermostats (NPTs). Next, regular MD simulations were performed at a constant pressure and temperature of 1 atm and 298 K for additional 50 ns per replicate. After the simulations were completed, trajectories were clustered and a representative structure as a centroid of the most populated cluster was selected.

3.2.7 Data and Code Availability

Structures of the centroids of the main clusters were deposited to PDB-Dev [124, 125] and the accession code of the entry for these models is PDBDEV_00000033. The CL-DMD software is available upon request from Prof. Dr. Nikolay Dokholyan.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium [126] via the PRIDE [127, 128] partner repository with the dataset identifier PXD015044.

3.3 Results and Discussion

The CL-DMD approach [67] was used to structurally model the monomeric tau protein. The approach uses experimental inter-residue short-distance crosslinking constraints for guiding DMD simulations and enabling modeling of the protein folding in reasonable time.

In brief, tau was crosslinked with a panel of short-range reagents succinimidyl 4,4′-azipentanoate (SDA) [129], DSG, and DSA. SDA is a hetero-bifunctional amino group-reactive and photoreactive reagent, spacer length approximately 5 Å, and DSG and DSA are homobifunctional amino group-reactive crosslinkers with spacer lengths of approximately 6 and approximately 7 Å, respectively. Crosslinked proteins were separated by SDS-PAGE, the monomer band was in-gel digested with proteinase K or trypsin proteolytic enzymes, and the digest was analyzed by LC-MS/MS to identify crosslinked peptides (Table 3.1). The monomer band of the crosslinked tau was used to exclude potential inter-protein crosslinks from the analysis. The distances between crosslinked residues are based on the length of the crosslinker reagents (Table 3.1), and were introduced as constraints into the DMD potentials (see Brodie et al., 2017 [67] for additional details). In addition, tau was characterized by surface

modification and long-distance crosslinking. Surface modification experimental data were extracted from crosslinking experiments using non-selective photoreactive diazirine-based reagent SDA to determine the characteristics of the residues as exposed or buried (Table 3.2). Long-distance crosslinking was used to estimate the overall protein topology (Table 3.3).

Table 3.1: Short-distance crosslinks used for CL-DMD simulations.

Constraint	Crosslinker	Spacer Arm (Å)	Residue Number 1	Residue 1	Atom	Amino Acid Length 1	Residue Number 2	Residue 2	Atom	Amino Acid Length 2	Total Length (Å)
Q6-K353	SDA	5	6	Q	CD	4	353	K	NZ	6.4	15.4
K24-Q6	SDA	5	24	K	NZ	6.4	6	Q	CD	4	15.4
K24-E9	SDA	5	24	K	NZ	6.4	9	E	CD	4	15.4
K24-D13	SDA	5	24	K	NZ	6.4	13	D	CG	2.6	14
K24-G16	SDA	5	24	K	NZ	6.4	16	G	CA	0	11.4
K24-G21	SDA	5	24	K	NZ	6.4	21	G	CA	0	11.4
K24-D22	SDA	5	24	K	NZ	6.4	22	D	CG	2.6	14
K24-K225	DSA	6	24	K	NZ	6.4	225	K	NZ	6.4	18.8
K24-K240	DSG	4	24	K	NZ	6.4	240	K	NZ	6.4	16.8
G37-K281	SDA	5	37	G	CA	0	281	K	NZ	6.4	11.4
E45-K24	SDA	5	45	E	CD	4	24	K	NZ	6.4	15.4
T101-K150	SDA	5	101	T	CG2	2.6	150	K	NZ	6.4	14
T102-K150	SDA	5	102	T	CG2	2.6	150	K	NZ	6.4	14
D110-K150	SDA	5	110	D	CG	2.6	150	K	NZ	6.4	14

Continued on next page

Continuation of Table 3.1

Constraint	Crosslinker	Spacer Arm (Å)	Residue Number 1	Residue 1	Atom	Amino Acid Length 1	Residue Number 2	Residue 2	Atom	Amino Acid Length 2	Total Length (Å)
P112-K150	SDA	5	112	P	CG	2.4	150	K	NZ	6.4	13.8
S113-K150	SDA	5	113	S	CB	1.5	150	K	NZ	6.4	12.9
K130-K150	DSG	4	130	K	NZ	6.4	150	K	NZ	6.4	16.8
K132-K150	DSA	6	132	K	NZ	6.4	150	K	NZ	6.4	18.8
K140-K130	DSA	6	140	K	NZ	6.4	130	K	NZ	6.4	18.8
K140-K150	DSA	6	140	K	NZ	6.4	150	K	NZ	6.4	18.8
K141-K130	DSA	6	141	K	NZ	6.4	130	K	NZ	6.4	18.8
K141-K148	DSA	6	141	K	NZ	6.4	148	K	NZ	6.4	18.8
K141-K150	DSA	6	141	K	NZ	6.4	150	K	NZ	6.4	18.8
K143-K132	DSA	6	143	K	NZ	6.4	132	K	NZ	6.4	18.8
K143-K148	DSA	6	143	K	NZ	6.4	148	K	NZ	6.4	18.8
K143-K150	DSA	6	143	K	NZ	6.4	150	K	NZ	6.4	18.8
K143-K163	DSA	6	143	K	NZ	6.4	163	K	NZ	6.4	18.8
K148-K148	DSA	6	148	K	NZ	6.4	148	K	NZ	6.4	18.8
K148-K150	DSG	4	148	K	NZ	6.4	150	K	NZ	6.4	16.8
K148-K174	DSA	6	148	K	NZ	6.4	174	K	NZ	6.4	18.8
K150-K130	DSG	4	150	K	NZ	6.4	130	K	NZ	6.4	16.8
K150-K132	DSA	6	150	K	NZ	6.4	132	K	NZ	6.4	18.8
K150-K148	DSA	6	150	K	NZ	6.4	148	K	NZ	6.4	18.8

Continued on next page

Continuation of Table 3.1

Constraint	Crosslinker	Spacer Arm (Å)	Residue Number 1	Residue 1	Atom	Amino Acid Length 1	Residue Number 2	Residue 2	Atom	Amino Acid Length 2	Total Length (Å)
K150-K180	DSA	6	150	K	NZ	6.4	180	K	NZ	6.4	18.8
K150-K190	DSA	6	150	K	NZ	6.4	190	K	NZ	6.4	18.8
K163-K24	DSG	4	163	K	NZ	6.4	24	K	NZ	6.4	16.8
K163-A89	SDA	5	163	K	NZ	6.4	89	A	CB	1.5	12.9
K163-A91	SDA	5	163	K	NZ	6.4	91	A	CB	1.5	12.9
K163-Q92	SDA	5	163	K	NZ	6.4	92	Q	CD	4	15.4
K163-P98	SDA	5	163	K	NZ	6.4	98	P	CG	2.4	13.8
K163-G100	SDA	5	163	K	NZ	6.4	100	G	CA	0	11.4
K163-K130	DSA	6	163	K	NZ	6.4	130	K	NZ	6.4	18.8
K163-K132	DSA	6	163	K	NZ	6.4	132	K	NZ	6.4	18.8
K163-K148	DSA	6	163	K	NZ	6.4	148	K	NZ	6.4	18.8
K163-K150	DSA	6	163	K	NZ	6.4	150	K	NZ	6.4	18.8
K163-K174	DSA	6	163	K	NZ	6.4	174	K	NZ	6.4	18.8
K163-K180	DSA	6	163	K	NZ	6.4	180	K	NZ	6.4	18.8
K163-K190	DSG	4	163	K	NZ	6.4	190	K	NZ	6.4	16.8
K163-K225	DSA	6	163	K	NZ	6.4	225	K	NZ	6.4	18.8
K163-K240	DSG	4	163	K	NZ	6.4	240	K	NZ	6.4	16.8
K163-S396	SDA	5	163	K	NZ	6.4	396	S	CB	1.5	12.9
K174-K130	DSA	6	174	K	NZ	6.4	130	K	NZ	6.4	18.8

Continued on next page

Continuation of Table 3.1

Constraint	Crosslinker	Spacer Arm (Å)	Residue Number 1	Residue 1	Atom	Amino Acid Length 1	Residue Number 2	Residue 2	Atom	Amino Acid Length 2	Total Length (Å)
K174-K143	DSG	4	174	K	NZ	6.4	143	K	NZ	6.4	16.8
K174-K150	DSA	6	174	K	NZ	6.4	150	K	NZ	6.4	18.8
K174-K190	DSA	6	174	K	NZ	6.4	190	K	NZ	6.4	18.8
K174-K225	DSA	6	174	K	NZ	6.4	225	K	NZ	6.4	18.8
K174-K240	DSA	6	174	K	NZ	6.4	240	K	NZ	6.4	18.8
K180-K150	DSG	4	180	K	NZ	6.4	150	K	NZ	6.4	16.8
K190-K224	DSA	6	190	K	NZ	6.4	224	K	NZ	6.4	18.8
K224-P219	SDA	5	224	K	NZ	6.4	219	P	CG	2.4	13.8
K224-T220	SDA	5	224	K	NZ	6.4	220	T	CG2	2.6	14
K225-K150	DSA	6	225	K	NZ	6.4	150	K	NZ	6.4	18.8
K225-K180	DSA	6	225	K	NZ	6.4	180	K	NZ	6.4	18.8
K225-K190	DSA	6	225	K	NZ	6.4	190	K	NZ	6.4	18.8
K225-K234	DSA	6	225	K	NZ	6.4	234	K	NZ	6.4	18.8
K225-K240	DSA	6	225	K	NZ	6.4	240	K	NZ	6.4	18.8
K225-K257	DSA	6	225	K	NZ	6.4	257	K	NZ	6.4	18.8
K225-K259	DSA	6	225	K	NZ	6.4	259	K	NZ	6.4	18.8
K225-K343	DSG	4	225	K	NZ	6.4	343	K	NZ	6.4	16.8
K234-K224	DSG	4	234	K	NZ	6.4	224	K	NZ	6.4	16.8
K240-K280	DSA	6	240	K	NZ	6.4	280	K	NZ	6.4	18.8

Continued on next page

Continuation of Table 3.1

Constraint	Crosslinker	Spacer Arm (Å)	Residue Number 1	Residue 1	Atom	Amino Acid Length 1	Residue Number 2	Residue 2	Atom	Amino Acid Length 2	Total Length (Å)
K254-K240	DSA	6	254	K	NZ	6.4	240	K	NZ	6.4	18.8
K254-K259	DSA	6	254	K	NZ	6.4	259	K	NZ	6.4	18.8
K257-K240	DSA	6	257	K	NZ	6.4	240	K	NZ	6.4	18.8
K259-K240	DSA	6	259	K	NZ	6.4	240	K	NZ	6.4	18.8
K267-K225	DSA	6	267	K	NZ	6.4	225	K	NZ	6.4	18.8
K267-K240	DSA	6	267	K	NZ	6.4	240	K	NZ	6.4	18.8
K267-K257	DSA	6	267	K	NZ	6.4	257	K	NZ	6.4	18.8
K267-K280	DSA	6	267	K	NZ	6.4	280	K	NZ	6.4	18.8
K267-K281	DSA	6	267	K	NZ	6.4	281	K	NZ	6.4	18.8
K274-K240	DSG	4	274	K	NZ	6.4	240	K	NZ	6.4	16.8
K274-K257	DSA	6	274	K	NZ	6.4	257	K	NZ	6.4	18.8
K274-K259	DSA	6	274	K	NZ	6.4	259	K	NZ	6.4	18.8
K274-K281	DSA	6	274	K	NZ	6.4	281	K	NZ	6.4	18.8
K281-S241	SDA	5	281	K	NZ	6.4	241	S	CB	1.5	12.9
K281-K257	DSA	6	281	K	NZ	6.4	257	K	NZ	6.4	18.8
K343-K353	DSG	4	343	K	NZ	6.4	353	K	NZ	6.4	16.8
K347-K353	DSA	6	347	K	NZ	6.4	353	K	NZ	6.4	18.8
K369-K141	DSA	6	369	K	NZ	6.4	141	K	NZ	6.4	18.8
K369-T231	SDA	5	369	K	NZ	6.4	231	T	CG2	2.6	14

Continued on next page

Continuation of Table 3.1

Constraint	Crosslinker	Spacer Arm (Å)	Residue Number 1	Residue 1	Atom	Amino Acid Length 1	Residue Number 2	Residue 2	Atom	Amino Acid Length 2	Total Length (Å)
K369-K347	DSA	6	369	K	NZ	6.4	347	K	NZ	6.4	18.8
K369-K370	DSG	4	369	K	NZ	6.4	370	K	NZ	6.4	16.8
K370-K347	DSG	4	370	K	NZ	6.4	347	K	NZ	6.4	16.8
K370-K353	DSA	6	370	K	NZ	6.4	353	K	NZ	6.4	18.8
K375-K369	DSA	6	375	K	NZ	6.4	369	K	NZ	6.4	18.8
K375-K395	DSG	4	375	K	NZ	6.4	395	K	NZ	6.4	16.8
N381-K353	SDA	5	381	N	CG	2.6	353	K	NZ	6.4	14
A382-K353	SDA	5	382	A	CB	1.5	353	K	NZ	6.4	12.9
K383-K369	DSA	6	383	K	NZ	6.4	369	K	NZ	6.4	18.8
K383-K375	DSA	6	383	K	NZ	6.4	375	K	NZ	6.4	18.8
K383-K395	DSA	6	383	K	NZ	6.4	395	K	NZ	6.4	18.8
K385-K369	DSG	4	385	K	NZ	6.4	369	K	NZ	6.4	16.8
K385-K370	DSA	6	385	K	NZ	6.4	370	K	NZ	6.4	18.8
K385-K375	DSA	6	385	K	NZ	6.4	375	K	NZ	6.4	18.8
K385-E380	SDA	5	385	K	NZ	6.4	380	E	CD	4	15.4
K385-N381	SDA	5	385	K	NZ	6.4	381	N	CG	2.6	14
K385-A382	SDA	5	385	K	NZ	6.4	382	A	CB	1.5	12.9
K385-S396	SDA	5	385	K	NZ	6.4	396	S	CB	1.5	12.9

Table 3.2: Photo-reactive SDA dead-end crosslinked tau residues used for surface modification analysis.

G2	D32	E64	P131	P198	P252	K278	K300	H348	K389	E410
S3	Y37	S65	L133	K199	S254	I279	L301	K350	T392	I411
S4	K43	T69	D135	T200	S256	G280	D302	K362	H393	Y413
H5	D44	P78	A137	G205	K259	S281	L303	D364	K394	K414
H6	Q45	T82	Q143	P207	S260	T282	S304	K366	L395	S415
H7	G46	A85	A144	K209	R261	E283	N305	Q370	E399	T422
H8	Y48	P89	K149	D212	L262	N284	V306	K372	N400	S423
H9	H51	A91	G163	S217	T264	L285	K309	I373	A401	P424
H10	Q52	A109	A171	S218	P266	K286	K317	S375	K402	S432
S11	D53	G119	T172	G223	V267	H287	K330	P383	A403	D440
Q25	Q54	T120	Q181	R228	M269	G290	K336	G384	K404	P442
E26	E55	T121	K182	R230	P270	G291	V337	G385	T405	T446
E28	D57	E123	K193	P232	D271	G292	K340	G386	D406	L447
V29	T58	G126	T194	L234	L272	K293	C341	N387	H407	E450
E31	K63	D129	A197	P238	S277	K299	G342	K388	G408	L455

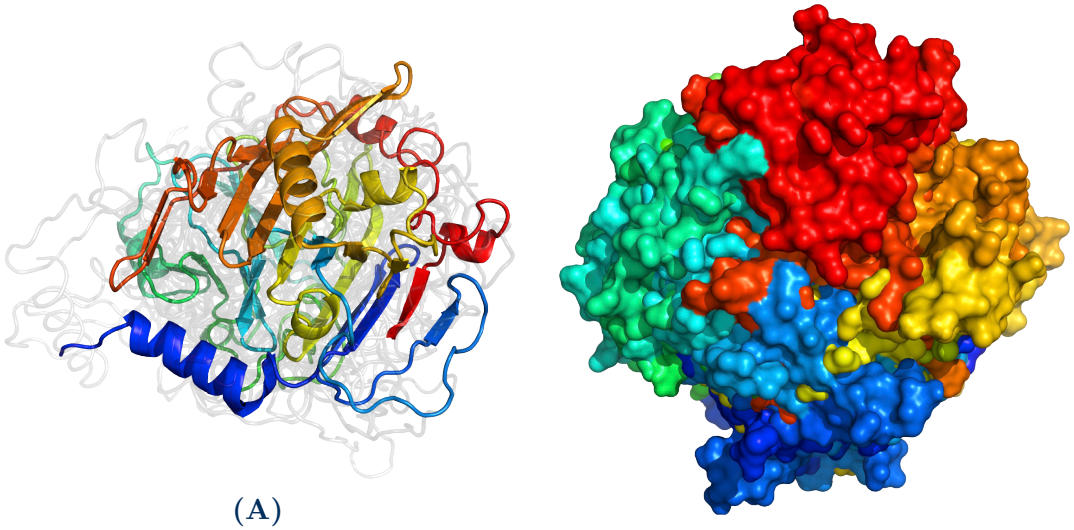
Table 3.3: Long-distance DSS crosslinks used for long-distance crosslinking analysis.

*Crosslinks labeled with an asterisk were also identified in the Mirbaha et al., 2018 study [130], **a double asterisk indicate crosslinks identified only in the Mirbaha et al., 2018 study, and not in our data. The K163-K438 crosslink exceeds 30 Å in length. Crosslinks without asterisks were only found in our data.

K24-K163**	K140-K148	K148-K180	K163-K224**	K224-K240	K240-K259	K259-K281
K130-K140*	K140-K150	K150-K163*	K163-K395**	K225-K234	K240-K274	K267-K280
K130-K150	K140-K163	K150-K174*	K163-K438**	K225-K240	K240-K281	K267-K281
K132-K143	K141-K150	K150-K180	K174-K190	K225-K254	K254-K259	K274-K281
K132-K148	K143-K150*	K150-K190	K174-K225	K225-K259	K257-K267	K353-K370
K132-K150	K143-K180	K150-K225	K180-K225	K225-K281	K257-K281	K370-K385
K132-K163	K148-K163	K163-K174*	K180-K240	K234-K259	K259-K274	K375-K383
K140-K143	K148-K174	K163-K180*	K224-K234	K240-K257	K259-K280	K383-K395

To obtain information on the global folding of tau, clustering analysis was performed on the lowest-energy structures obtained during CL-DMD simulations. In ordered proteins, the lowest-energy structures are usually represented by only one or few states, whose conformations are close to the corresponding native structures. In contrast, disordered proteins are usually represented by a broader variety of distinct structures, reflecting the conformational freedom of the intrinsically disordered protein [131]. In this case, the lowest-energy structures forming the conformational ensemble for tau are shown as overlays in Figure 3.1A. They can be described as relatively compact globular structures with a common general topology, containing a number of secondary-structure elements. A distinct topology which is shared between all of these conformers can be observed. Four subdomains can be recognized in the structures, which arranged in a tetrahedral fashion (Figure 3.1B). The C-terminal portion of the molecule was not found to be in close proximity to the N-terminal portion with extensive contacts, as was suggested earlier [71]. In contrast, it is found here to be confined on both sides by two intermediate subdomains.

In any structural proteomics experiment, some of the experimentally derived



(A) (B)

Figure 3.1: Conformational ensemble of native tau in solution as determined by CL-DMD.

(A) Conformational ensemble of tau. Representatives of the major clusters are aligned and colored grey. Lowest energy centroid is shown in cartoon representation and colored N-to-C from blue-to-red. (B) Sub-domain topology of the tau conformational ensemble. Residues are colored as in (A).

constraints may conflict with each other because proteins exist in multiple conformations. During DMD simulations, the system is driven to satisfy the maximum possible number of constraints, so the generated ensemble of structures satisfies subsets of consistent experimental constraints. The representative models are then selected by performing geometrical clustering of this ensemble and selecting the centroids of the most populated clusters. Depending on the conformation and the number of constraints that are satisfied, these structures will have various energy scores assigned by the force field [112, 120]. This allows us to take into account both the nature of the conformational changes of a protein in solution (i.e., while the structure of the protein is "breathing" and transitioning between conformations), as well as mitigate possible errors that may have occurred during the determination of the experimental constraints [110]. Nevertheless, it appears that the lowest-energy

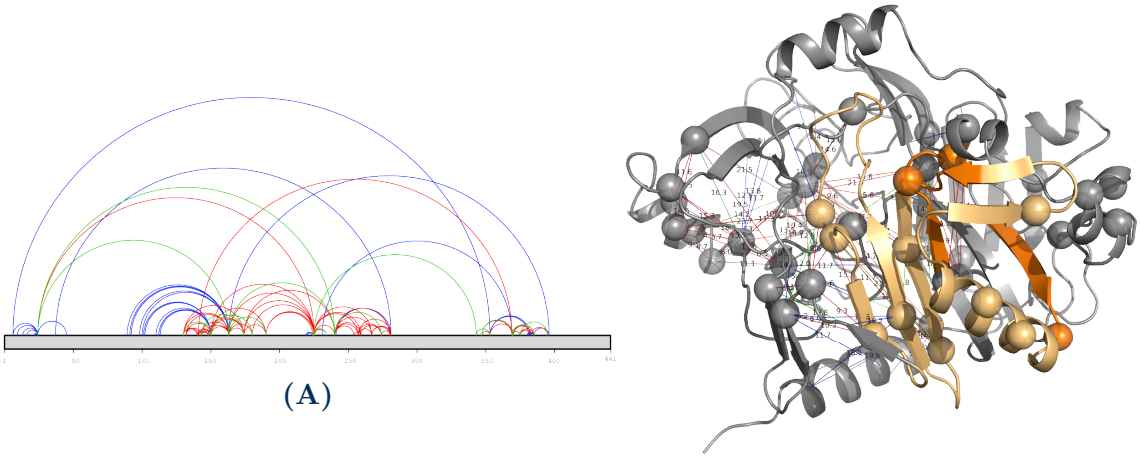
conformers satisfy most of the experimental distance constraints, while the models with higher energy satisfy a lower number of constraints. A comparison of predicted structures within an ensemble revealed that the conformational changes to the protein structure were mainly due to internal rearrangements within the N-, C-terminal, and intermediate subdomains. One could argue that the short-distance constraints used in the simulations might drive the structures to more compact states. To address this concern, all of the experimental constraints from the obtained structures are removed, and an unconstrained all-atom simulation using GROMACS 2018 and CHARMM36 force field for an additional 54 ns in 3 replicatesis run. During the first 4 ns each system was subjected to the simulated annealing procedure, where both the protein and the solvent underwent four heating cycles from a temperature of 300 K to 325 K for 0.5 ns and then cooled back to 300 K for another 0.5 ns. After that, regular molecular dynamics simulations were performed for 50 ns. The replicate trajectories were subjected to clustering, and the centroids of the most populated clusters were selected as the new relaxed models. The relaxed structures had generally the same conformation (3 Å root-mean-square deviation between the lowest constrained and unconstrained energy states), with only a slight relaxation of the more flexible regions, which suggests that the experimental constraints did not significantly bias the models toward artificially overly compacted states. For simplicity, only the relaxed structure of the first centroid is presented in Figure 3.2.

Some secondary-structure elements were observed in the conformers of the ensemble (Figure 3.3). The extent of the predicted secondary-structure motifs in the lowest-energy conformer was higher than previously proposed [71], but, as shown in the case of α-synuclein [68], this probably reflects the presence of transient secondary-structure elements. Prediction of the secondary-structure content from circular dichroism (CD) data [130] using the BeStSel server [132] (Figure 3.4) was in general

Figure 3.2: Original and relaxed structures of the lowest energy centroid. The starting conformer is shown in rainbow colors and the relaxed without crosslinking distance constraints conformer is shown in grey.

agreement with the content found in the lowest-energy centroid (Figure 3.3): 11 % versus 7 % of helix and 21 % versus 31 % of antiparallel β structure for the model and predictions, respectively.

The surface modification results were in agreement with the final tau structures (Figure 3.5A). Photoreactive SDA dead-end crosslinks (the photoreactive part of the reagent is attached to the protein residues and the NHS moiety of the reagent is hydrolyzed) were used as quasi surface modification reagents. SDA is a non-selective diazirine-based crosslinking reagent and has the potential to modify any protein residue, although some reactivity preferences for diazirine-based reagents have been reported [133]. In total there were 155 modified protein residues detected (Table 3.2). Most modifications were consistent with the lowest-energy conformer structure



(A) **(B)**

Figure 3.3: Short-distance crosslinks used for CL-DMD of tau protein in solution.

Lowest-energy conformer from the ensemble as in Figure 3.1 is shown. (A) Short-distance crosslinks used in the CL-DMD simulations are shown as lines connecting Cα atoms and as circular arcs on the diagram. Red, green and blue lines correspond to DSA, DSG, and SDA crosslinks, respectively. (B) Amiloidogenic R2, R3, and R4 peptides ^{275}VQIINK^{280}, ^{306}VQIVYK^{311}, and ^{337}VEVK^{340} are highlighted in orange. V306-F378 region which was found by cryo-EM forming cross-beta structure in fibrils [98] shown in light orange.

(only 8 of the 155 modified residues in the structure cannot be directly accessed from protein surface by the modifying reagent) (Figure 3.5A). Long-distance crosslinking cannot be used directly in CL-DMD simulations but can be employed for confirmation of the overall topology of the protein models [67]. DSS crosslinker (spacer length of approximately 11 Å) was used for long-distance crosslinking (Table 3.3). There were 52 long-distance intra-protein DSS crosslinks in agreement with the final lowest-energy conformer of the protein (Figure 3.5B). Previously reported long-distance DSS crosslinks [130] were also found to be in agreement with the obtained structure (Figure 3.5B), with only one out of ten observed crosslinks exceeding 30 Å length (Table 3.3).

Overall, the tau conformational ensemble determined here can be considered as a representation of a wide energy-funnel characteristic of the molten globular

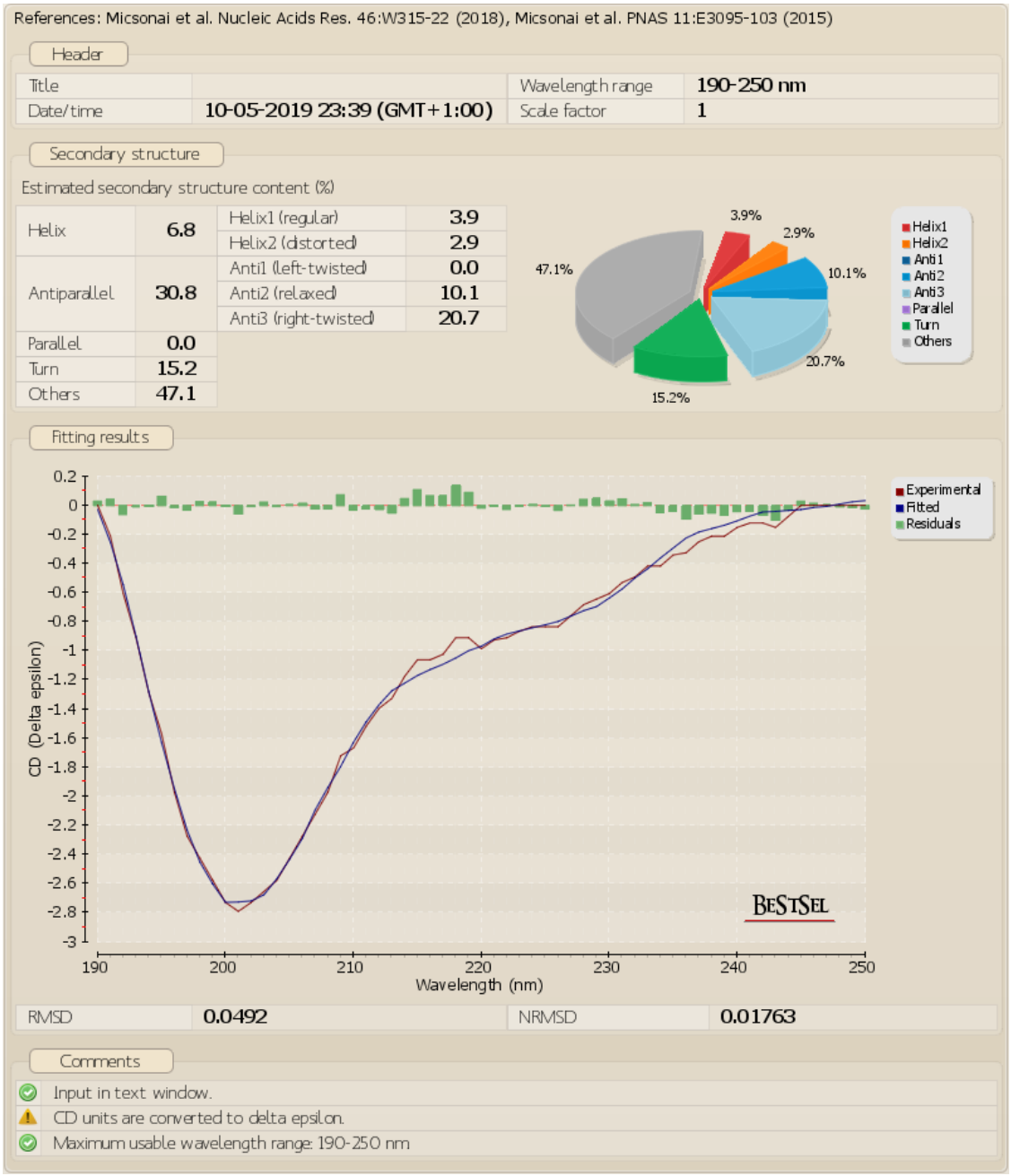


Figure 3.4: CD analysis of the native tau protein in solution.
CD data from Mirbaha et al., 2018 [130] were analyzed using BeStSel server [132].

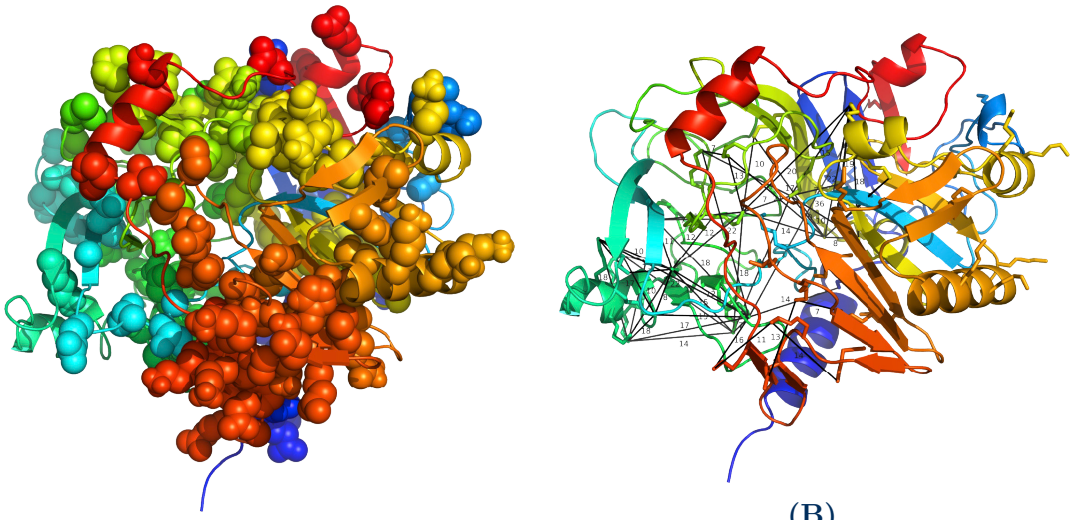


Figure 3.5: Experimental validation of the tau structure with surface modification and long-distance crosslinking.

(A) Side chains of residues that are found to be modified as dead-ends by photo-reactive moiety of hydrolyzed SDA crosslinking reagent are shown as spheres. (B) long-distance crosslinking DSS crosslinks are shown as black lines.

state of disordered proteins. For each particular representative within the ensemble, CL-DMD was able to capture the structural features present. This structure is in agreement with an earlier finding of residual native-like structural features in unfolded proteins [134]. This distribution of near-ground-state inter-converting conformations may be the basis for the structural plasticity of intrinsically disordered proteins [131], and the existing persistent structural features can determine the subsequent misfolding and oligomerization pathways. The 2N4R isoform of tau protein used in this study contains an N-terminal projection domain with N1 and N2 repeats, a proline-rich domain, a microtubule binding region containing R1, R2, R3, and R4 repeats, and a C-terminal domain [97]. The R1–R4 repeats are reported to be involved in microtubule binding and are important for aggregation, with the R3 and R4 repeats forming a cross-β sheet core of paired helical filaments *in vivo* [98]. Interestingly, in the structure presented in this chapter this region was

found to contain multiple β structure elements (Figure 3.3), which is consistent with its tendency to form β structures in the aggregates. In fact, seven out of eight β strands of the fibrils core (the exception being the β2 strand) were already found to be in a β structural form in the lowest-energy conformer (Figure 3.3). The central ^{347}KDR^{349} turn was also found as a loop connecting two β strands. These observed structural features lead to a hypothesis that the mechanism of conformational change in this region (which leads to aggregation) is a "pulling away" of the amino acid 341–357 hairpin from the core of the molecule, with further re-orientation of the hydrogen bonding from intra-protein to inter-protein cross-β bonding. Most of this region is buried in the structures presented here, which is also consistent with the idea that a dramatic rearrangement of the protein molecule needs to occur to initiate pathological aggregation.

Hyper-phosphorylation has been proposed as a factor facilitating conversion of the native monomeric form of tau to the pathological aggregated form of tau. Tau can be phosphorylated at approximately 30 sites by multiple kinases [135]. Hypothetically, phosphorylation in this region—at least at residues S293, S324, and S356—may enable the "loosening" and opening up of the structure, thereby exposing the residues to the solvent and to a potentially interacting tau molecule during aggregate formation.

In summary, a structural conformational ensemble of native tau in solution was determined by CL-DMD using short-distance constraints derived from experimentally observed inter-residue crosslinks. The predicted structure was corroborated by surface modification and long-distance crosslinking experimental data. The general folding pattern and the transient secondary-structure motifs found in the predicted structure of the tau protein may help to explain the structural transitions of the molecule that occur during the pathological conversion and oligomerization.

3.4 Conclusion

A *de novo* structural model of the native structure of tau in solution was produced using short-distance CL-DMD simulations. The model was validated against experimental data from surface modification and long-distance crosslinking experiments. The predicted structure is represented by an ensemble of relatively compact globular conformations with transient secondary-structure elements. The region containing R1–R4 repeats was found to contain multiple β structure elements, which is consistent with the tendency of this region to form a cross-β structure in the aggregates. The obtained structure can serve as a starting point for analyzing the misfolding and oligomerization of the tau protein.

CHAPTER

FOUR

IMPROVING IDENTIFICATION OF *IN-ORGANELLO* PROTEIN-PROTEIN INTERACTIONS USING AN AFFINITY-ENRICHABLE, ISOTOPICALLY-CODED, AND MASS SPECTROMETRY-CLEAVABLE CHEMICAL CROSSLINKER

This chapter was adapted from the publication:

Makepeace, Karl A.T., Yassene Mohammed, Elena L. Rudashevskaya, Evgeniy V. Petrotchenko, F-Nora Vögtle, Chris Meisinger, Albert Sickmann, and Christoph H. Borchers. "Improving Identification of *In-organello* Protein-Protein Interactions Using an Affinity-enrichable, Isotopically Coded, and Mass Spectrometry-cleavable Chemical Crosslinker." *Molecular & Cellular Proteomics* 19, no. 4 (2020): 624-639. DOI: 10.1074/mcp.RA119.001839

Contribution disclosure: Work presented in this chapter was carried out in the laboratories of Christoph Borchers, Albert Sickmann, and Chris Meisinger. The project was initially conceived by Evgeniy Petrotchenko and Christoph Borchers. Chris Meisinger and F-Nora Vögtle prepared the mitochondria to be used in the crosslinking experiments and contributed to evaluation of the CL-MS results. Elena Rudashevskaya prepared and collected all experimental data. Karl Makepeace, Yassene Mohammed, and Elena Rudashevskaya all contributed to experimental design and data analysis. Karl Makepeace and Yassene Mohammed developed the computational pipeline. Karl Makepeace, Yassene Mohammed, Elena Rudashevskaya, and Evgeniy Petrotchenko wrote the first draft of the manuscript. Albert Sickmann and Christoph Borchers oversaw the project and all authors contributed to the final version of the published manuscript.

4.1 Introduction

Proteins and their intricate networks of interactions are fundamental to many of the molecular processes that govern life [136, 137]. Insights into the structures of individual proteins and their interactions with other proteins in a proteome-wide context has been made possible by recent developments in the relatively new field of chemical crosslinking mass spectrometry (CL-MS) [59, 62]. Crosslinkers stabilize transient interactions by forming covalent chemical linkages between amino acid residues. The crosslinked proteins are then enzymatically digested into peptides, and the covalently-coupled crosslinked peptides are identified by mass spectrometry. These identified crosslinked peptides thus provide evidence of interacting regions within or between proteins [138, 139, 140, 141, 142, 143, 144, 145, 146].

Proteome-wide crosslinking analysis has the potential to provide structural char-

acterization of protein-protein interactions (PPIs) and protein complexes in their natural cellular and tissue environments. Moreover, the technique is well suited for capturing the "molecular sociology" of the cell, including the more weakly interacting and transient complexes. Such interactions may not be identified through traditional biochemical techniques using rigorous purification procedures that tend to only be compatible with robust complexes [136, 147].

Although this technique is fairly straightforward, for proteome-wide applications it is made considerably more complex by the combinatorial nature of the crosslinked peptides, which can originate from any of the proteins in the proteome. To address this issue, cleavage of the crosslinker itself—which then provides information on the masses of the individual peptides constituting a crosslink—has been recognized as a critical feature for the crosslinking analyses of complex samples [61, 82, 146, 148, 149, 150, 151, 152]. Several successful analytical strategies exploiting this feature have recently been reported for proteome-wide crosslinking studies [60, 146, 153]. The relative and absolute abundances of crosslinked peptides in typical peptide digests are much lower than those of single peptides, so specific enrichment of crosslinked peptides from the total peptide digest has also been shown to be critical for successful analyses [82]. Another advantageous feature that can be incorporated into the crosslinker is isotopic coding. It enables specific selection of the crosslink signals in MS^1 for subsequent MS^2 analysis, and adds additional characteristic features to the spectra of the crosslinks, which can then be used to further improve the confidence of the identification [82].

Here I report the application of the membrane-permeable affinity-enrichable isotopically-coded and CID-cleavable crosslinker cyanurbiotin-dimercaptopropionyl-succinimide (CBDPS) to *in-organello* crosslinking analysis [82]. I describe a CL-MS workflow that improves upon previously published workflows in terms of detection,

acquisition, and identification of crosslinked peptides [57, 84, 153, 154]. This study yielded a rich crosslinking dataset, revealing hundreds of intra- and inter-molecular PPIs within the mitochondrial organelle. Using this analytical approach, I have uncovered system-wide interaction patterns that would not be accessible through classic protein-chemistry research techniques.

4.2 Materials and Methods

All materials were from Sigma-Aldrich unless noted otherwise. The mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium [126] via the PRIDE [127, 128] partner repository with the dataset identifier PXD014055 and PXD017066.

4.2.1 Mitochondria preparation and *in-organello* crosslinking

Highly purified yeast mitochondria, strain YPH499, were prepared as described previously [155, 156]. The mitochondrial sample was thawed on ice, and then diluted gently to 5 mg/mL in isotonic buffer (250 mM sucrose, 1 mM EDTA, 10 mM MOPS-KOH, pH 7.2). Mitochondria were crosslinked with an equimolar mixture of isotopically light (L) and heavy (H) cyanurbiotin-dimercaptopropionyl-succinimide (CBDPS-H8 and CBDPS-D8, respectively) (Creative Molecules Inc.) at 2 mM as follows: samples were pre-warmed at 21°C for 5 min; after addition of the crosslinker mixture, samples were kept at 21°C for 10 min and then put on ice for 110 min. The crosslinking reaction was quenched with the addition of ammonium bicarbonate to a final concentration of 50 mM for 20 min. Crosslinked mitochondria were collected by centrifugation at 18,000 g for 20 min in the cold, and immediately proceeded with

lysis.

4.2.2 Sample Lysis, Pre-fractionation and Digestion

The pellet of crosslinked mitochondria was resuspended in a hypotonic buffer consisting of 1 mM EDTA, 10 mM MOPS-KOH, pH 7.2, left on ice for 20 min and lysed by sonication using a Vibra Cell Ultrasonic Processor for a total processing time of 1 min (70 % amplitude, 5 pulses). The lysate was centrifuged at 18,000 g for 20 min, and the resulting pellet (Pellet 1) and supernatant were collected, frozen in liquid nitrogen and stored at −80°C until the next day. Pellet 1 was used to prepare all of the samples, and is hereafter referred to as "membrane 1" or "membrane low centrifugation". The supernatant was centrifuged at 100,000 g for 45 min and the resulting pellet (Pellet 2) and supernatant used to prepare all of the samples are hereafter referred to as "membrane 2" or "membrane high centrifugation" and "soluble", respectively. Proteins were solubilized from Pellet 1 and Pellet 2 with 2 % SDS in 10 mM MOPS-KOH pH 7.2, at 37°C for 30 min and 300 RPM, with subsequent centrifugation at 18,000 g for 20 min.

Proteolysis was performed with trypsin (Promega, Sequencing Grade Modified, trypsin:protein ratio 1:20) using the FASP protocol [157] with modifications and ultrafiltration units with a nominal molecular weight cutoff of 30 kDa (Vivacon® 500, Sartorius). Samples were loaded to pre-washed filtration units ($\leq$ 400 μg of protein per unit). After pre-concentration, samples were washed with 400 μL of 8 M urea buffer, treated with 200 μL 0.1 M DTT solution, 200 μL 0.05 M IAA solution, washed 3x with 200 μL 8 M urea solution, 3x with 50 mM Tris-HCl buffer pH 8.5. Digestion was performed overnight (18 h) at 37°C. Peptides were collected by washing the filter units with 100 μL 50 mM Tris-HCl buffer pH 8.5 and then 200 μL 0.5 M NaCl.

4.2.3 Enrichment of crosslinked peptides

The resulting peptide mixture was acidified with formic acid (FA), desalted using C18 SPE columns (BondElute SPEC C18AR, Agilent Technologies), eluted with 0.4 % FA with 90 % acetonitrile (ACN), and dried completely. Samples were reconstituted with SCX buffer A (10 mM KH_2PO_4, 20 % ACN, pH 2.7) , and separated by strong cation exchange (SCX) chromatography using an Dionex UltiMate 3000 (Thermo Fisher Scientific) HPLC system and an POLYSULPHOETHYL A column (PolyLC Inc., Columbia, US, 5μm particle size, 200 Å pore size, 150 x 1.0 mM) [54]. A ternary buffer system was used: SCX buffer A (10 mM KH_2PO_4, 20 % ACN, pH 2.7), SCX buffer B (10 mM KH_2PO_4, 250 mM KCl, 20 % ACN, pH 2.7) and SCX buffer C (10 mM KH_2PO_4, 600 mM KCl, 20 % ACN, pH 2.7). From each sample, 19 SCX fractions were collected at 37.5–250 mM KCl and dried. Collected fractions were further enriched for CBDPS crosslinked peptides on monomeric avidin beads (Pierce Biotechnology) as described previously [84] and analyzed by LC-MS/MS.

4.2.4 LC-MS/MS analysis

Mass spectrometric analysis was performed using a UltiMate 3000 coupled to the ESI-source of an Orbitrap Fusion Lumos or Q Exactive HF (Thermo Fisher Scientific). Samples were loaded in 0.1 % TFA onto a trapping column (Acclaim PepMap 100 C18, 5 μm particle size, 100 μm x 2 cm, Thermo Fisher Scientific) for pre-concentration. Peptides were separated on C18 analytical column (Acclaim PepMap RSLC, 75 μm x 500 mM, 2 μm, 100 Å, Thermo Fisher Scientific) using a binary gradient (solvent A: 0.1 % formic acid (FA); solvent B: 0.1 % FA, 84 % ACN). For MS analysis on the Lumos, peptides were separated with a 120-min gradient (0–100 min: 3–35 % solvent B (84 % ACN, 0.1 % FA), 100–110 min: 35–42 % B, 110–120 min: 42–80 % B, 0.250 μL/min flowrate). On the Q Exactive HF, peptides

were separated with 180 min gradient: 0–160 min: 3–35 % solvent B, 160–170 min: 35–42 % B, 170–180 min 42–80 % B.

MS data were acquired using data-dependent methods utilizing either TopSpeed (TopS) or TopN; targeted mass difference (MTag); or inclusion list (Incl) precursor selection modes [84].

4.2.5 Data-dependent Acquisition methods

The data-dependent acquisition utilized dynamic exclusion, with an exclusion duration of 30 seconds and exclude after n times set to 1 (Lumos). MS^1 and MS^2 events used 120,000 and 60,000 resolution FTMS scans, respectively, with a scan range of 350–1800 m/z in the MS mode. For the TopN methods a loop count of 10 was used. For the TopS method, a cycle time of 3 seconds was used. For MS^2 acquisition, the HCD collision energy was set to 28 % NCE for Q Exactive HF runs and CID of 35 % for Orbitrap Fusion Lumos runs. Only precursor ions with charge states of 3+ to 7+ were selected for fragmentation. The acquisition method for MTag runs was identical to the method described for the TopS acquisitions except that a "Targeted Mass Difference" filter with the mass difference set to 8.0502 Da with a light-heavy analogue intensity range set to 50–100 was used. The acquisition method for Incl runs was identical to the method described for the TopS acquisitions except that a "Targeted Mass" filter was used. The parent mass lists used in the "Targeted Mass" filter for these analyses were calculated using Hardklör (ver. 2.3.0; see Table 4.1 for parameters [158]), Krönik (ver. 2.02; see Table 4.2 for parameters) [159], and in-house scripts. Only doublets that were identified as charge state 3 and greater were included in the parent mass list.

Table 4.1: Hardklör parameters.

Parameter	Setting
instrument	Orbitrap
resolution	120000
centroided	0
ms_level	1
scan_range_min	0
scan_range_max	0
signal_to_noise	2
sn_window	250
static_sn	0
boxcar_averaging	0
boxcar_filter	0
boxcar_filter_ppm	5
mz_min	0
mz_max	0
smooth	0
algorithm	Version2
charge_algorithm	Quick
charge_min	1
charge_max	9
correlation	0.95
averagine_mod	0
mz_window	5.25
sensitivity	2
depth	2
max_features	12
distribution_area	1
xml	0

Table 4.2: Krönik parameters.

Parameter	Setting
−c	5
−d	3
−g	1
−m	10000
−n	400
−p	10

4.2.6 MS1 Feature Analysis

The identification of doublets ($\Delta 8.0502\,\mathrm{Da}$) in MS^1 and evaluation of crosslinker-modified precursors was accomplished using a combination of Hardklör, Krönik, and in-house scripts. Criteria for classifying an MS^1 feature from the Krönik output as a doublet was that the light and heavy monoisotopic peaks for MS^1 features were separated by $8.0502\,\mathrm{Da} \pm 0.01\,\mathrm{Da}$, that the heavy-isotopic peak had a maximum intensity that occurred at a retention time that is between $-0.4\,\mathrm{min}$ and $0.05\,\mathrm{min}$ of the maximum intensity of the light-isotopic peak, that the $\log_2$ of the heavy-isotopic partner summed intensity divided by the light-isotopic partner summed intensity was 0 ± 2, and that the maximum intensities observed for both the light and the heavy isotopic peaks are each greater than or equal to $25,000$ intensity units.

4.2.7 Bioinformatics Analysis

All RAW format data files were converted to mzXML format using MSConvertGUI (release 3.0.10730) of the ProteoWizard tool suite (release 3.0.11252) [160] and the data analysis was completed using the in-house Qualis-CL software pipeline. The pipeline consists of 5 external open source software packages and 4 in-house developed modules to allow identification of crosslinked peptides, MS^1 and MS^2 PSM feature annotation, and statistical validation.

Inter-protein, intra-protein and loop-linked Lys-Lys and protein-N-term-Lys crosslinked peptides as well as single peptide PSMs were obtained using Kojak search engine (ver. 1.5.5) (see Table 4.3 for parameters) [58]. In its diagnostic mode, Kojak reports additional details for each PSM assignment. This was an essential aspect as this detailed information is utilized in the Qualis-CL pipeline. The database for data analysis included list of proteins identified earlier in highly purified mitochondrial samples [161], and proteins that have reference to mitochondria in

their description in Saccharomyces Genome Database (SGD) and/or UniProt. Thus, it contained known mitochondrial proteins, associated proteins and contaminants. The database included concatenated target and decoy protein sequences, in which the decoy entries were generated by shuffling each peptide's amino acids in each protein target entry (see Supplemental Material 1 of Makepeace et al. 2020 [162], DOI: 10.1074/mcp.RA119.001839). This generates decoy entries that have distributions of protein and peptide lengths that are similar to the target proteins. The database contains 1295 protein entries, and same number of decoy entries. All searches were performed with carboxyamidomethylated cysteine as a fixed modification, methionine oxidation as a variable modification and a maximum of 3 tryptic missed cleavages.

Hardklör [158] and Krönik [159] software tools were used to determine MS^1 spectral and chromatographic features associated with the MS^1 parent masses that had been acquired with MS^2 in the raw data. The MS^1 features detected by Hardklör and Krönik software tools were then used as input for the Qualis-CL algorithm to find those features that exist as multiplets (doublets, triplets, or quadruplets) within user-specified tolerance settings between the heavy and light pairs. The tolerances allow for variations in retention times between the labeled and unlabeled pairs, relative intensity differences (20 % in this work), and variations in the mass differences of the doublets (0.01 Da here).

The MS^2 features that result from cleavage products of the crosslinker were detected and annotated using an algorithm which was developed in-house. For each assigned MS^2 spectrum, the presence of 4 crosslinker cleavage products is determined. Following calculation of MS^1 and MS^2 features additional logic calculates meta-features that check for agreement between MS^1 and MS^2 features.

Identifications, scores, MS^1 features, MS^2 features, and meta-features (described in Table 4.5) were combined in one table per crosslink type, i.e., inter-protein, intra-

Table 4.3: Kojak parameters.

Parameter	Setting
threads	7
database	...\UP-Mitos-20161223-TDrev-K.fasta
export_percolator	1
export_pepXML	1
percolator_version	2.08
enrichment	0
instrument	0
MS1_centroid	0
MS2_centroid	0 or 1 depending on dataset
MS1_resolution	120000
MS2_resolution	60000
cross_link	nK; nK; 509.097364; CBDPS_Light
cross_link	nK; nK; 517.147578; CBDPS_Heavy
mono_link	nK; 526.1238898
mono_link	nK; 527.107929
mono_link	nK; 534.1741038
mono_link	nK; 535.158143
modification	M; 15.9949
modification_protC	0
modification_protN	0
diff_mods_on_xl	1
max_mods_per_peptide	1
mono_links_on_xl	1
enzyme	[KR] |{P}
fragment_bin_offset	0
fragment_bin_size	0.01
ion_series_A	0
ion_series_B	1
ion_series_C	0
ion_series_X	0
ion_series_Y	1
ion_series_Z	0
decoy_filter	decoy
isotope_error	1
max_miscleavages	3
max_peptide_mass	8000
min_peptide_mass	600
max_spectrum_peaks	0
ppm_tolerance_pre	3
prefer_precursor_pred	2
spectrum_processing	1
top_count	300
truncate_prot_names	0
turbo_button	0

protein, loop, or single. The Percolator algorithm (ver. 2.08) was used to perform the validation and the calculation of the q-values (see Table 4.4 for parameters). All software used was combined into a single pipeline that takes raw data in mzXML format and generates the result tables. An additional module combines these result tables with interactome databases to generate statistics, and highlights the known and novel interaction.

Table 4.4: Percolator parameters.

Parameter	Setting
-w	. . . [filename].SVMweights.txt
-U	n/a
-Y	n/a
-F	0.001

4.2.8 Structural Validation of Crosslink Identifications

XiView [163] and open source PyMOL [164] were used to map crosslinks to existing structural models for yeast electron transport chain complexes and super-complexes.

4.2.9 Experimental Design and Statistical Rationale

Crosslinked sample fractionation into one soluble and two membrane fractions was performed to allow assessment of the sub-compartment localizations of the detected protein interactions. Each sample fraction sample was digested and further separated by SCX chromatography. Each SCX chromatographic sample for the soluble fractions was analyzed with 3 different acquisition strategies allowed by the instrument software, these were: (1) data-dependent acquisition of the top 10 features or for 3 seconds—TopS/TopN, (2) triggering on the presence of mass difference of two MS^1 features equal to the difference of heavy and light crosslinker pairs—MTag, and (3) inclusion

list based on post analysis of the TopS/TopN and MTag data—Incl. Both membrane fractions were analyzed with TopS only. These multiple analyses of fractions are complementary replicates. Biological replicates would be prohibitively expensive in terms of the total number of LC-MS samples used for the experimental design of this study. The decoy entries in the search database were generated by randomizing the sequence while keeping the C-terminus amino acid unchanged for all tryptic peptides in each protein. This ensures decoy entries which are very similar to the forward ones in terms of the number, length, and composition of the tryptic peptides. The q-values were estimated by Percolator software and were used for all FDR thresholds. FDR cutoff value was put at 2 % at the identified crosslinked PSM level.

4.3 Results

4.3.1 Developing an integrated experimental and computational crosslinking-MS workflow

Previously, the Borchers lab developed a multifunctional crosslinking reagent, CBDPS, that combined several features which improve the performance of CL-MS analyses: affinity-enrichability, isotopic-coding, and MS-cleavability [82]. By taking advantage of the specific biochemical and physical features of the CBDPS crosslinking reagent (Figure 4.1A), the detection, acquisition, and identification of crosslinker-modified peptides in a complex sample were improved. These improvements affect three critical points in the analytical workflow (Figure 4.1B): (1) affinity-enrichment of crosslinker-modified peptides; (2) specific MS^2 acquisition of crosslinker-modified peptides using targeted mass difference (MTag) or inclusion list (Incl) data-dependent acquisition methods; (3) utilization of crosslinker-specific spectral features in the validation of crosslinks. This chapter describes a utilization of this reagent for proteome-

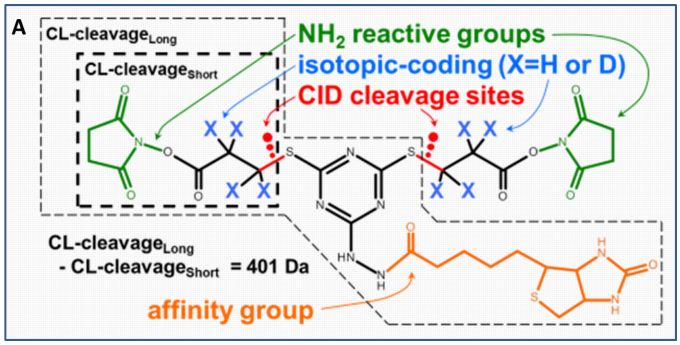
wide analyses by taking advantage of these features in both the experimental and computational aspects of the CL-MS workflow.

4.3.2 Affinity enrichment for improved detection of crosslinker-modified peptides

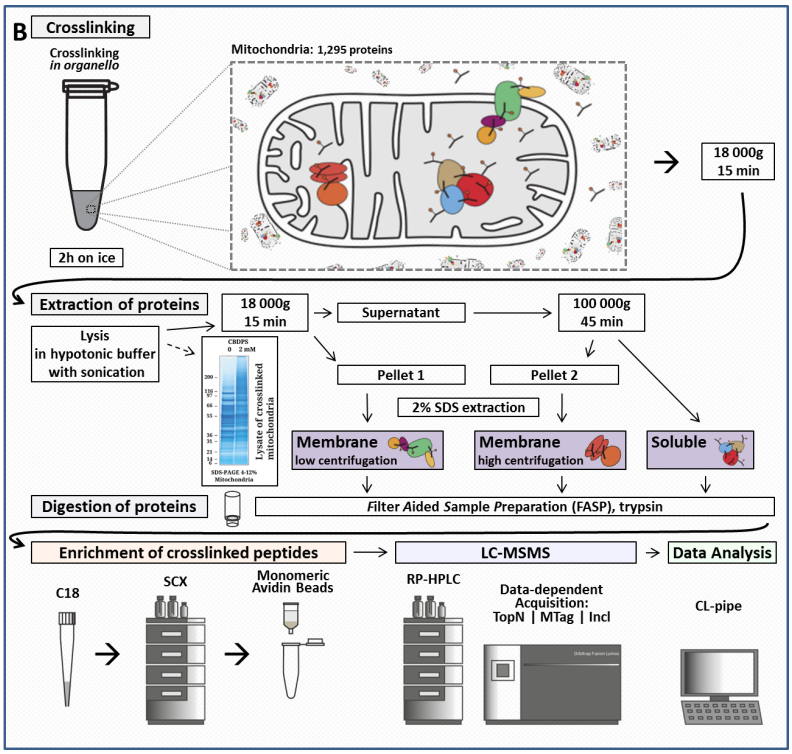
The yield of crosslinking products is typically low; therefore enrichment procedures prior to mass spectrometry analysis dramatically improve the detection and identification of these crosslinks (Figure 4.2A). Specific enrichment of CBDPS crosslinker-modified peptides is achieved through the use of the biotin tag which has been incorporated into the reagent enabling enrichment with avidin. It should be noted that inter-peptide crosslinks and single peptides containing CBDPS "dead-end" or "loop-link" modifications would both be enriched. For this reason, a chromatographic step to separate inter-peptide crosslinks from single peptides (e.g., SCX chromatography [165] or size-exclusion chromatography (SEC) [52]) is often performed prior to affinity enrichment, and can further assist in the CL-MS analysis. In order to quantitate the resulting improvement in detection of crosslinker-modified peptides, I compared the number of MS^1 doublet features (Figure 4.3) that were observed in crosslinked samples before and after enrichment. Almost twice as many (170 %) CL-modified MS^1 features were detected after the affinity-tag enrichment procedure (Figure 4.2B).

4.3.3 Isotopic-coding for the specific acquisition of crosslinker-modified peptides

In my preliminary crosslinked sample analyses, a distinct bimodal distribution in the TopN spacing (i.e., the number of MS^2 scans occurring between MS^1 scans)



(A)



(B)

Figure 4.1: Crosslinking reagent and experimental workflow.
(A) CBDPS molecular diagram showing NH_2 reactive groups, CID-cleavable bonds, isotopic-coding positions, and biotin affinity-tag. Short and long crosslinker-cleavage portions are also indicated. (B) General experimental workflow for *in-organello* crosslinking. The affinity enrichment, LC-MS analysis, and data analysis steps all take advantage of the various features of the CBDPS crosslinker shown in (A). Specifically, the biotin-tag of the crosslinker allows affinity enrichment of crosslinker-modified peptides prior to MS analysis, the isotopic-coding allows the use of targeted MS acquisition methods, and the mass spectral; features relating to both the isotopic-coding and crosslinker-cleavage result in improved confidence in peptide-spectrum match identifications.

was observed for the SCX fractions that were expected to be most abundant in crosslinker-modified peptides (Figure 4.4). This was an indicatation that the duty cycle was frequently reaching its maximum allowed time duration between consecutive MS^1 scans and may, therefore, be unable to acquire all of the unique precursors at those particular retention times. The time-dependent effects of the TopN spacing indicated that the duty cycle limit was most often met between the retention times of 25–80 min which corresponded to the most feature-rich portion of the LC-MS run. This was an indication that the MS^2 spectra of many potential crosslinked peptides were not being acquired when using the conventional TopSpeed (TopS) acquisition method (i.e., an acquisition mode where the maximum number of the most intense peaks in an MS^1 scan are acquired in a defined time period for each duty cycle), or the TopN acquisition method, where the N most-abundant peaks in an MS^1 scan are acquired for each duty cycle.

Next, I compared the number of observed MS^1 doublet features that were acquired as MS^2 spectra across three different data-dependent acquisition methods: the TopS method; the targeted mass difference (MTag) method; or inclusion lists derived from the MTag LC-MS data from a prior injection (Incl) (Figure 4.5A). In order to maximize the number of crosslinked peptides acquired in the MS/MS mode, a CL-specific acquisition strategy was developed and employed (Figure 4.6). First, "MTag" LC-MS data were collected for each SCX fraction. For this, an acquisition method was used that had a targeted mass difference filter set for the isotopic mass difference between the two forms of the crosslinker—e.g., Δ 8.0502 Da for CBDPS-H_8/D_8—along with a L-H partner intensity range set to 50–100. With this MTag method, the instrument is instructed to monitor MS^1 scans for the presence of MS^1 signals separated by the specified mass delta as they are being acquired, and to trigger MS^2 acquisition of both the light and heavy partner precursor ions when the

delta is observed (Figure 4.5B). Because MS^2 is only triggered when the mass delta is observed, the amount of duty cycle time the instrument expends acquiring MS^2 scans for precursors that do not contain crosslinker is kept to a minimum (Figure 4.4). This should result in the MS^2 acquisition of spectra from additional low-abundance crosslinker-modified precursor ions—ions which may have been otherwise not been acquired due to a low ranking in a standard TopN method parent mass list. In addition, I would expect to observe the acquisition of a larger number of unique MS^1 peptide features when using an MTag acquisition method than when using a TopS method because the instrument can spend comparatively more time collecting MS^1 scans with the MTag method.

To ensure that MS^2 spectra are acquired for as many crosslinker-modified precursors as possible from each sample/fraction in a single LC-MS run, "Incl" LC-MS data were also acquired for each fraction. Here a crosslinker-specific parent mass inclusion list was calculated using the MS^1 data from the MTag acquisition. This was accomplished by processing the MS^1 data from the MTag acquisition with a software pipeline that incorporates Hardklör [158], Krönik [159], and in-house scripts to generate ".csv" crosslinker-specific parent mass inclusion lists (Figure 4.6). Calculation of these inclusion lists and construction of the inclusion list methods can be performed immediately after the MTag LC-MS run is completed. With the Incl method, the instrument is instructed to monitor MS^1 scans as they are being acquired for the presence of the specific masses in the parent mass list (which was calculated from the prior MTag run) and to trigger MS^2 acquisition of both the light and heavy partner precursor ions when observed.

Surprisingly, I found that the targeted methods showed no improvement in the number of MS^1 doublet precursor ions acquired with MS^2 compared to untargeted methods for those samples in which enrichment was performed (Figure 4.5C). In

fact, the TopS method appeared to outperform both the MTag and Incl methods with respect to the number of MS^1 doublet features acquired with both or only light or heavy isotopic precursor ion partners being acquired. The expected advantage of using a targeted acquisition method was only realized in the analysis of sample fractions that had not previously undergone the affinity enrichment step (Figure 4.5B). In this case, a 40 % improvement in MS^1 doublet acquisition was observed for the MTag method, and a 63 % improvement was observed for the Incl method, compared to the TopS method. Presumably, the benefit realized by using targeted acquisition modes will increase together with the increasing complexity of the sample analyzed. This will be an important consideration when extending the technique to systems of increasing complexity (organelles, to cells, to tissues constituted of different cell types, etc.) or, potentially, in shortened analyses in which there is a lesser degree of sample pre-fractionation performed prior to enrichment in which greater performance improvements may be expected.

4.3.4 Integrating crosslinker-specific mass-spectral feature information for improved performance in peptide-spectrum match validation

The data-analysis pipeline integrates existing software tools and in-house-developed logic into a single tool (Figure 4.7A). Briefly, MS data (.raw file format) were converted into mzXML format. Searches were performed using Kojak [58], which was configured to output all Kojak diagnostic files for each input mzXML file. A protein database of the yeast mitochondrial proteome was assembled based on prior proteomic investigations [156, 161, 166] and was subsequently used for all of the searches and identifications.

Concurrently the mzXML files were processed using Hardklör [158] and Krönik [159] software tools to produce a list of MS^1 features. This list was then analyzed with the Qualis-CL algorithm to yield a list of crosslinking MS^1 features, (i.e., paired MS^1 features that exhibit the specific mass delta corresponding to the difference between heavy and light CBDPS crosslinker (e.g., 8.0502 Da)).

The search results from Kojak, as well as MS^1 features from Hardklör/Krönik, were combined and further annotated with additional information on these features based on peptide-spectrum matchs (PSMs) using the Qualis-CL algorithm. Specifically, now added to each PSM is corresponding information from the Kojak diagnostic output including: preliminary and final scores and ranks for both the individual peptides, the Hardklör score for the precursor mass, the score difference between the best ranking and second best ranking PSM for all tested precursor masses, the label class (light or heavy) of the crosslink moiety, the relative mass error for the precursor as determined by Kojak and which exists in the mzXML, the total ion current, base peak m/z, and intensity for the MS^2. The PSMs were also annotated with information derived from the list of paired MS^1 features (e.g., the H/L intensity ratio for the isotopic partners, the retention time deltas for the isotopic partners, whether the isotopic pattern matches the expected pattern), in addition to information on the crosslinker-cleavage fragment ions in MS^2 obtained directly from the mzXML file (e.g., the matched short and long crosslinker-cleavage fragment ions matched, the matched dead-end signature ions), and meta-PSM information (e.g., if a corroborating PSM exists for the corresponding isotopic partner).

A complete list of PSM features with descriptions is given in Table 4.5. In order to take advantage of the benefits that can be obtained by considering all of these feature dimensions simultaneously in a statistical validation of the PSMs, the popular semi-supervised machine learning algorithm Percolator [65] was used. Each PSM is

described by 41 feature dimensions and 5 Percolator-required dimensions (specid, Label, scannr, Peptide, Proteins). The new list of PSMs, now containing additional feature information, was then processed using Percolator for statistical validation.

Table 4.5: Description of all features used to represent PSMs.

#	Feature Name	Category	Description	Datatype
1	specid	pIN	Unique PSM identifier	string
2	Label	pIN	Target/Decoy PSM label (1 = Target PSM; −1 = Decoy PSM)	integer
3	scannr	pIN	MS^2 scan number for PSM	integer
4	Score	Score	The Kojak cross-correlation score	decimal
5	dScore	Score	The difference between the reported PSM score and the next best PSM score as reported in the original Kojak output	decimal
6	NormRank	Score	The sum of the Peptide #1 and Peptide #2 ranks in the first scoring pass of the Kojak algorithm	integer
7	PPScoreDiff	Score	The score contribution of only the lower scoring peptide (always reported as "pep_b") in a crosslink	decimal

Continued on next page

Continuation of Table 4.5

#	Feature Name	Category	Description	Datatype
8	Charge	Precursor	Precursor ion charge state (determined by Kojak)	integer
9	Mass	Precursor	Theoretical neutral mass of the PSM (value returned from Kojak)	decimal
10	PPM	Precursor	Difference (in parts per million) between the PSM mass (Kojak) and the Obs Mass (Kojak)	decimal
11	LenShort	Sequence	The residue count of the shortest of Peptide #1 and Peptide #2.	integer
12	LenLong	Sequence	The residue count of the longest of Peptide #1 and Peptide #2.	integer
13	LenSum	Sequence	The summed residue count of Peptide #1 and Peptide #2.	integer
14	pep_a_rank_1st _pass1	Diagnostic Info	The rank of Peptide #1 based on the first reported score in the Kojak diagnostic output	decimal
15	pep_a_pass1_score	Diagnostic Info	The first reported score for Peptide #1 in the Kojak diagnostic output	decimal

Continued on next page

Continuation of Table 4.5

#	Feature Name	Category	Description	Datatype
16	pep_a_pass2_score	Diagnostic Info	The second reported score for Peptide #1 in the Kojak diagnostic output	decimal
17	pep_b_rank_1st _pass1	Diagnostic Info	The second reported score for Peptide #1 in the Kojak diagnostic output	decimal
18	pep_b_pass1_score	Diagnostic Info	The second reported score for Peptide #1 in the Kojak diagnostic output	decimal
19	pep_b_pass2_score	Diagnostic Info	The second reported score for Peptide #1 in the Kojak diagnostic output	decimal
20	hardklor_prec_score	Diagnostic Info	The second reported score for Peptide #1 in the Kojak diagnostic output	decimal
21	dscore	Diagnostic Info	The difference between the reported PSM score and the next best PSM score as calculated from the best scoring PSM score minus the next best scoring PSM score in the Kojak diagnostic output.	decimal

Continued on next page

Continuation of Table 4.5

#	Feature Name	Category	Description	Datatype
22	l_or_h	Isotopic-coding	Label for the isotopically light (="−1") or isotopically heavy (="1") crosslinker moiety	integer
23	rel_err_mass_meas_koj	Precursor	The mass difference between the theoretical PSM Mass and the precursor mass as reported by Kojak	decimal
24	rel_err_mass_mzxml_calc	Precursor	The mass difference between the theoretical PSM Mass and the precursor mass as reported in the mzXML	decimal
25	ms2_feat_pep_a	CL-cleavage	The sum of booleans for matched crosslinker-cleavage ions for Peptides #1 for charge states of precursor $z-1$ (possible values = 0,1,2)	integer
26	ms2_feat_pep_b	CL-cleavage	The sum of booleans for matched crosslinker-cleavage ions for Peptides #2 for charge states of precursor $z-1$ (possible values = 0,1,2)	integer
27	log_ms2_totIonCurrent	Precursor	Log_{10} of the total ion current in the MS^2 scan	decimal

Continued on next page

Continuation of Table 4.5

#	Feature Name	Category	Description	Datatype
28	ms2_basePeakMZ	Precursor	The base peak intensity in the MS^2 scan	decimal
29	log_ms2 _-basePeakIntensity	Precursor	Log_{10} of the base peak intensity in the MS^2 scan	decimal
30	de_sig_expectation _match	CL-cleavage	Whether the observed matching of the MS^2 DE-signatures match the expectation for the PSM.	boolean
31	log2_int_most_l _int_most_h	Isotopic-coding	The binary logarithm of the heavy/light ratio of the maximum XIC intensity observed for the precursor MS^1 feature (± 1 min). If PSM DX Level (i.e., the number of isotopically-coded moieties that exist for the PSM) > 1 then the MS^1 signals corresponding to entirely light and entirely heavy isotopic forms are used in this calculation. If PSM DX Level $= 0$ then 0 is returned.	decimal

Continued on next page

Continuation of Table 4.5

#	Feature Name	Category	Description	Datatype
32	rt_diff_most_l _most_h	Isotopic-coding	The retention time (sec) difference from the observed light isotopic partner XIC maximum intensity to the heavy isotopic partner XIC maximum intensity. If PSM DX Level > 1 then the MS[1] signals corresponding to entirely light and entirely heavy isotopic forms are used in this calculation. If PSM DX Level = 0 then 0 is returned.	decimal
33	prec_charge_mzxml _best_kj_agree	Precursor	Whether the best scoring precursor charge determination in the Kojak diagnostic output match the precursor charge reported in the mzXML.	boolean
34	prec_charge_kj _best_kj_agree	Precursor	Whether the best scoring precursor charge determination in the Kojak diagnostic output match the best scoring PSM in the Kojak diagnostic output.	boolean

Continued on next page

Continuation of Table 4.5

#	Feature Name	Category	Description	Datatype
35	heavy_mods_pos_in _multiplits_agree	Isotopic-coding	Whether the acquired precursor match the appropriate isotopic MS^1 precursor signal for the PSM.	boolean
36	koj_kr_lh_agree	Isotopic-coding	Whether the isotopic label assignment for heavy and light by Kojak PSM and Kronik agree.	boolean
37	iso_part_psm	Isotopic-coding	The number of unique isotopic partners that match the PSM. If no isotopic partners are possible then equals -1.	boolean
38	charge2	Precursor	Precursor ion charge state = 2 (determined by Kojak)	boolean
39	charge3	Precursor	Precursor ion charge state = 3 (determined by Kojak)	boolean
40	charge4	Precursor	Precursor ion charge state = 4 (determined by Kojak)	boolean
41	charge5	Precursor	Precursor ion charge state = 5 (determined by Kojak)	boolean
42	charge6	Precursor	Precursor ion charge state = 6 (determined by Kojak)	boolean
43	charge7	Precursor	Precursor ion charge state = 7 (determined by Kojak)	boolean

Continued on next page

Continuation of Table 4.5

#	Feature Name	Category	Description	Datatype
44	charge8	Precursor	Precursor ion charge state = 8 (determined by Kojak)	boolean
45	Peptide	pIN	Peptide #1 and Peptide #2 strings combined (for pIN)	string
46	Proteins	pIN	Non-redundant tab-separated list of all proteins listed in Protein #1 and Protein #2 (for pIN)	string

A 20–30 % increase was observed in confidently-identified PSMs representing inter-protein crosslinks the additional feature information was included (Figure 4.7B. A complete list of crosslinked peptide-spectrum matches can be found online as Supplemental Material 2 for Makepeace et al. 2020 [162] (DOI: 10.1074/mcp.RA119.001839).

4.3.5 Overview of the identifications with respect to fractionation

The early SCX fractions contain predominantly single and loop peptides, while crosslinked peptides are appearing in the later SCX fractions (Figure 4.8). The overlap in identification between the three centrifugation fractions shows the benefit of the pre-fractionation steps by centrifugation. A slight enrichment of inter-protein crosslinked peptides is observed in the high centrifugation fraction, while intra-protein identifications are almost distributed between the high centrifugation and soluble fractions. Having a relatively higher number of identifications in the low

centrifugation fraction, i.e., 2–3 times higher, in the intra-protein crosslinks as well as single and loop peptide identifications suggests that extra centrifugation steps could perhaps be beneficial for even better fractionation.

4.3.6 The yeast mitochondria interactome

To demonstrate the applicability of the analytical strategy described above toward elucidation of a protein-protein interactome, highly purified yeast mitochondria was analyzed. Found in this analysis were 751 non-redundant crosslinked inter-protein inter-peptide pairs (FDR 2 %), involving 264 yeast mitochondrial proteins representing 338 unique protein-protein interactions. An abbreviated list of the most frequently identified PPIs is given in Table 4.7. The complete list of list of PPIs identified can be found online as Supplemental Material 3 for Makepeace et al. 2020 [162] (DOI: 10.1074/mcp.RA119.001839) in the form of a Microsoft Excel spreadsheet. The results reported here are compared against other previous studies involing CL-MS at a proteome-wide scale in Table 4.6. These data provide structural insight into the PPIs of 20 % of the proteins in the yeast mitochondrial proteome (Figure 4.9) and, to my knowledge, represents the most comprehensive set of yeast mitochondrial PPIs determined in a single CL-MS experiment at publication date. Furthermore, soluble, peripheral, and integral protein classes were approximately evenly represented in the interacting proteins accounting for 31 %, 29 %, and 24 % of the proteins involved in PPIs, respectively (Figure 4.10B) [161]. Of the yeast mitochondrial interactions that were identified, 71.7 % were not previously described in the EMBL-EBI IntAct Molecular Interaction Database (downloaded on October 18, 2019) (Figure 4.9) [167, 168]. In addition, 185 previously unknown PPIs (Figure 4.9, Table 4.7) were identified. This data provide novel insights into the interactions of many mitochondrial proteins with soluble, peripheral and integral membrane proteins represented (Figure 4.10B).

Furthermore, 83 % of the interacting proteins identified have previously described sub-compartment localizations and 17 % previously ambiguous or undefined (Figure 4.10C). The distribution of the sub-compartment localizations of the proteins involved in the identified PPIs appears to make biological sense (Figure 4.10A) [161]. The most commonly observed sub-compartment localization pairs were between inner-membrane proteins (81 PPIs), inner-membrane and matrix proteins (52 PPIs), and matrix proteins (51 PPIs). PPIs with protein localizations that would preclude interaction were observed infrequently or not at all (e.g., 6 outer-membrane to matrix PPIs were observed, no inter-membrane space to matrix were observed).

4.3.7 Structural validation of crosslinks on existing structural models

I assessed the validity of the crosslinking results by mapping the identified crosslinks to existing structural models of complexes involved in the mitochondrial electron transport chain available in the Protein Data Bank (PDB) database, i.e., 3CX5, 6HU9, 6CP3, and 6B8H. I also charted the observed Cα–Cα distance distributions versus distances of random possible links in each of these complexes. Figure 4.11 shows the mapping of the identified inter- and intra-protein crosslinks to these four (super) complexes from the membrane high centrifugation fractions along with the histogram of the distances. The mapping shows that the majority of crosslinks have Cα–Cα distances below the 38 Åmaximum distance threshold for the CBDPS crosslinker. When mapping the crosslink identifications to the available PDB model of yeast mitochondrial ATP synthase (PDB: 6B8H), shown in Figure 4.11D, in which two ATP synthase monomers oriented in a V-shape with an angle of 86°, there was some concern about possible false identification of the long crosslinks between the two complexes ranging from 100 to 325 Å (labeled with red arrows in Figure 4.11D).

Table 4.6: A comparison of recent mitochondria mass spectrometry-based crosslinking studies.

Reference	This work	Schweppe et al. "Mitochondrial protein interactome elucidated by chemical cross-linking mass spectrometry." Proceedings of the National Academy of Sciences 114.7 (2017): 1732-1737.	Liu et al. "The interactome of intact mitochondria by cross-linking mass spectrometry provides evidence for coexisting respiratory supercomplexes." Molecular & Cellular Proteomics 17.2 (2018): 216-232.
Organism	Yeast	Mouse	Mouse
Material	Isolated mitochondria	Isolated mitochondria	Isolated mitochondria
Crosslinker	CBDPS	BDP-NHP	DSSO
Number of LC-MS runs	55 LC-MS runs	72 LC-MS runs 11 biological replicates run in technical duplicate	42 LC-MS runs for native condition crosslinking data 21 SCX fractions for 2 biological replicates
Soluble-protein & membrane-protein fractionation	Yes	No	No
SCX fractionation	Yes	Yes	Yes
Affinity enrichment	Yes	Yes	No
Isotopic-coding	Yes	No	No
Acquisition method types	TopN	ReACT (PMID:23413883)	TopN
Platform MS	Orbitrap Q Exactive HF	Velos-FTICR (custom-build)	Orbitrap Fusion
Non-redundant CL pairs	Inter: 751 Intra: 9521	Inter+Intra: 2427	Inter+Intra: 3322
Proteins Involved	Inter: 251 (unambiguous) Intra: 784 (unambiguous) Total: 811 (unambiguous)	Inter+Intra: 327	Inter+Intra: 359
PPI's	Inter: 338 (unambiguous) Intra: 784 (unambiguous)	459	Total: 885 Intra: 276 Inter: 609 (Not reported in manuscript, counted from supplementary materials)
FDR	2 %	1.91 %	2 %

Table 4.7: Identified protein-protein interactions with highest number of PSMs

Protein A	Gene A	Protein B	Gene B	Total Number of PSMs	Previously reported in IntAct	Previously reported in SGD	Number of unique peptide-peptide IDs	Number of PSMs in sp60 fraction (Soluble)	Number of PSMs in sp61 fraction (Membrane low centrifugation)	Number of PSMs in sp62 fraction (Membrane high centrifugation)
P40961	PHB1	P50085	PHB2	191	Yes	Yes	13	5	94	92
P18238	AAC3	P18239	PET9	157	Yes	No	18	0	117	40
P07256	COR1	P07257	QCR2	113	Yes	Yes	13	47	33	33
P07251	ATP1	P09457	ATP5	107	Yes	Yes	18	29	39	39
P12695	LAT1	P32473	PDB1	91	Yes	Yes	20	29	12	50
P81449	TIM11	P81451	ATP19	73	No	No	15	3	32	38
P00830	ATP2	P09457	ATP5	72	Yes	Yes	10	27	19	26
P28241	IDH2	P28834	IDH1	72	Yes	Yes	5	44	11	17
P09624	LPD1	P12695	LAT1	67	No	No	27	7	1	59
P05626	ATP4	P30902	ATP7	64	No	Yes	9	4	29	31
P53312	LSC2	P53598	LSC1	64	Yes	Yes	4	28	16	20
P21801	SDH2	Q00711	SDH1	54	Yes	Yes	7	7	20	27
P00830	ATP2	P07251	ATP1	51	Yes	Yes	9	17	16	18
P12695	LAT1	P16387	PDA1	46	Yes	Yes	13	9	4	33
P07253	CBP6	P21560	CBP3	41	Yes	Yes	13	0	27	14
P19262	KGD2	P20967	KGD1	40	Yes	Yes	9	19	3	18
P05626	ATP4	P07251	ATP1	38	No	Yes	6	2	22	14
P09624	LPD1	P16451	PDX1	35	Yes	Yes	10	7	2	26
P19414	ACO1	Q12497	FMP16	34	No	No	4	19	3	12
P16547	OM45	P40215	NDE1	33	No	No	11	0	21	12
P21306	ATP15	P38077	ATP3	31	Yes	Yes	4	10	8	13
P39925	AFG3	P40341	YTA12	30	Yes	Yes	7	0	21	9
P07342	ILV2	P25605	ILV6	28	Yes	Yes	7	12	8	8
P12695	LAT1	P16451	PDX1	27	No	No	10	2	0	25
P16387	PDA1	P32473	PDB1	27	Yes	Yes	8	5	4	18
P0CS90	SSC1	P39987	ECM10	26	No	No	7	7	10	9
P30902	ATP7	Q06405	ATP17	25	Yes	Yes	4	0	8	17
P40496	RSM25	Q03799	MRPS8	25	No	No	2	17	1	7
P05626	ATP4	P09457	ATP5	22	No	Yes	6	2	11	9
P04710	AAC1	P18239	PET9	21	Yes	Yes	4	0	18	3
P09624	LPD1	P19955	YMR31	20	Yes	Yes	8	11	1	8
P00830	ATP2	P05626	ATP4	19	No	Yes	3	7	9	3
P25349	YCP4	Q12335	PST2	19	Yes	Yes	3	2	16	1
P00044	CYC1	P16547	OM45	16	No	No	4	0	5	11
P19955	YMR31	P20967	KGD1	16	Yes	Yes	5	9	2	5
P05626	ATP4	Q12233	ATP20	15	No	Yes	1	2	7	6
P07143	CYT1	P40215	NDE1	15	No	No	1	0	10	5
P33421	SDH3	Q00711	SDH1	15	No	No	1	4	7	4
P53252	PIL1	Q12230	LSP1	15	Yes	Yes	3	0	9	6
P07251	ATP1	P0CS90	SSC1	14	No	No	3	6	4	4

However, when four yeast mitochondrial ATP synthase dimers are adjacent side by side, the membrane becomes flatter resulting in parallel monomers organized side-by-side with a distance of 130 Å between their rotational axis [170]. With an average diameter of 100 Å, one can calculate an expected distance of around 30 Å between the crosslinked residues (instead of 100–325 Å), which is in the range of the CBDPS crosslinker. Unfortunately, there is no PDB structure with that formation of parallel monomers to map the crosslinks to. Figure 4.12 shows all identified crosslinks mapped to PDB structures of yeast mitochondrial electron transport chain complexes and supercomplexes for all sample pre-fractions.

4.4 Discussion

Chemical crosslinking combined with mass spectrometry is a valuable method for attaining structural information about proteins and identifying protein-protein interactions. Investigations on low complexity systems, for example purified proteins or protein complexes, are now becoming routine. However, applying this technique to complex systems, for example organelles or cells, still presents a variety of challenges. For example, these challenges involve technical aspects such as overcoming the inherently low abundance of crosslinked peptides which leads to limited detection in MS^1 and concomitantly limited MS^2 acquisition when using a typical shotgun proteomics workflow. These challenges also extend to various bioinformatic aspects, which include not only the efficient and confident identification of crosslinked peptides from MS^2 spectra, but also exploiting additional crosslinking-specific aspects of the acquired data, including MS^1 features, MS^2 features, and meta-PSM features, in order to further improve identification of crosslinked peptides. To overcome these challenges, I show here that by utilizing an enrichable, isotopically-labeled, MS-

cleavable crosslinking reagent, targeted MS^2 acquisition strategy, and a software pipeline designed to integrate CL-specific information it is possible to improve the detection, acquisition, and identification of crosslinker-modified peptides and improve analysis of complex whole-proteome systems. This improved method was applied *in-organello* to isolated yeast mitochondria, and has allowed the detection of protein-protein interactions involving a sixth of the mitochondrial proteome. Moreover, 71.7 % of these identified interactions comprise interactions not reported in the EMBL-EBI IntAct Molecular Interaction Database [168, 169], while when comparing with Saccharomyces Genome Database (SGD) [171]—which is more extensively annotated—61 % identified interactions were not reported. However, it is important to mention that the annotation of interactions in all databases are not complete and lag behind the literature, so for example interactions related to Pyruvate Dehydrogenase (PDH) or Succinate Dehydrogenase (SDH) complexes are well known and identified in the data collected here, but do not appear either IntAct or SGD. A validation of the identified crosslinks by mapping to existing structural models of complexes involved in the mitochondrial electron transport chain available from PDB showed good agreement. In all four (super) complexes used, the Cα–Cα distance distributions agreed to with the expectation of the used chemical crosslinker, i.e., distances of 38 Å and less. There is no PDB model available for yeast mitochondrial ATP synthase with four or more monomers. Mapping to the available one dimer (PDB: 6B8H) produces misleading results suggesting inter-monomer crosslinks of 100 to 325 Å in length. However, it has been shown previously [170, 172] that when four or more dimers are adjacent side by side, the membrane flattens resulting in monomers organized in parallel side by side with a distance of 130 Å between their rotational axis. In this arrangement, and with a calculated distance of around 30 Å, the identified inter-complex crosslinks (red arrows Figure 4.11D) are plausible. There

is a membrane embedded portion for each of the ETC complexes shown in Figure 4.11. A small number of crosslinks from the low centrifugation fraction map with longer than expected distance for CBDPS and across this membrane boundary (see Figure 4.12 for PDB ID 3CX5 Sp61 fraction). These may be false positive CL-PSM identifications (as the FDR acceptance threshold was set to 2% some are expected). All other crosslinks shown in the figure appear to form linkages between regions that exist on the same side of this boundary - which is what we would expect to see. The presented *ex vivo* crosslinking analytical approach is suitable for proteome-wide applications, and provides a technical foundation that will yield insights into condition-specific protein conformations, protein-protein interactions, system-wide protein function or dysfunction, and diseases.

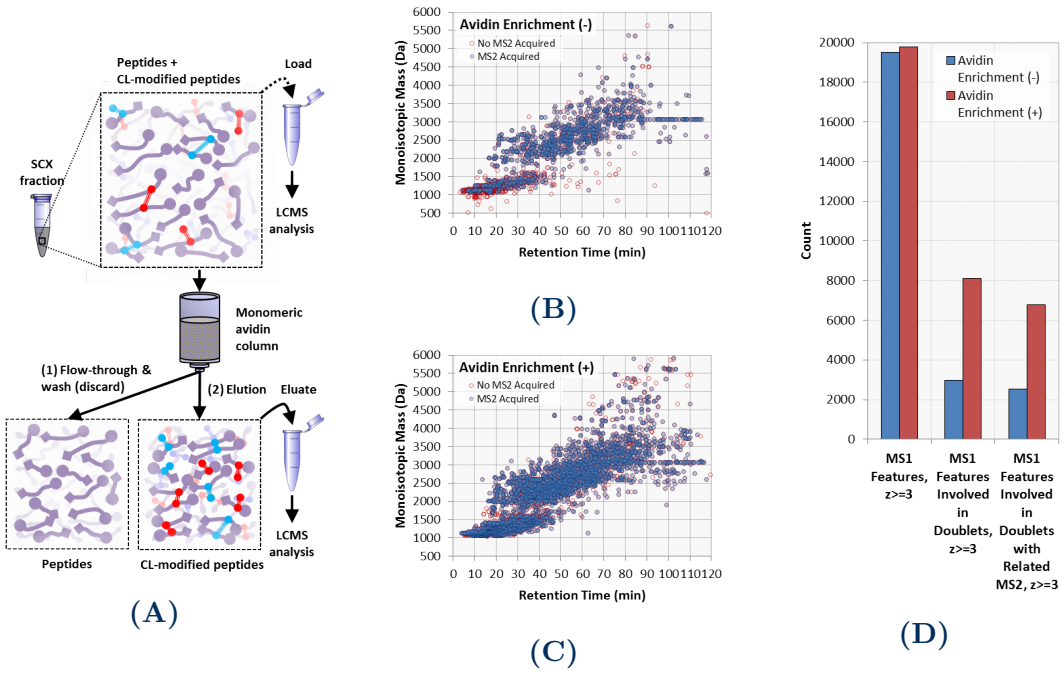


Figure 4.2: Affinity enrichment improves detection of crosslinker-modified peptides.

(A) Diagram of the affinity-tag-based enrichment strategy. An SCX fraction containing a mixture of peptides and crosslinker-modified peptides is loaded onto a monomeric avidin column. Peptides that do not contain crosslinker are discarded in the flow-through and wash fractions while those that do are retained. These retained crosslinker-modified peptides are eluted from the column and collected (eluate) for subsequent LC-MS analysis. (B, C) A portion of the SCX fraction prior to enrichment (load) may also be saved for LC-MS analysis to assess the improvement in crosslinker-modified species detected as shown. (D) A comparison of $\Delta\,8.0502\,\mathrm{Da}$ doublet features found in MS^1 for samples without (load) and with (eluate) enrichment shows approximately 2.7 times as many doublet features with enrichment.

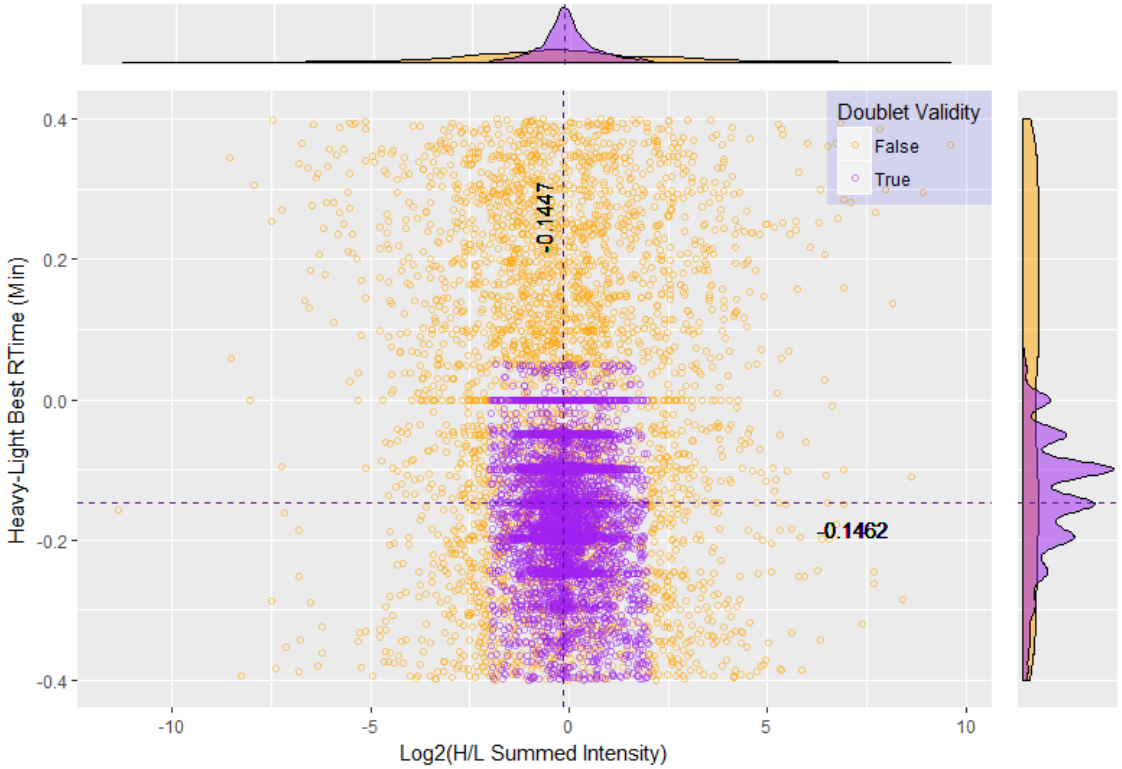


Figure 4.3: MS^1 Δ 8.0502 Da doublet features identified in Krönik output. Criteria for classifying an MS^1 feature from Krönik output as a true doublet was that the mass difference between the light and heavy monoisotopic peaks for the pair of MS^1 features are 8.050213952 Da $\pm$ 0.01 Da, that the heavy-isotopic peak has a maximum intensity that occurs at a retention time that is between -0.4 min and 0.05 min of the maximum intensity observed for the light-isotopic peak, and that the maximum intensities observed for both light and heavy-isotopic peaks are each greater than or equal to 25000 intensity units. The doublet identification and true/false classification is illustrated here using data from a single Lumos LC-MS run. The median $\log_2$(H/L) ratio for the summed isotopic-partner intensities for those MS^1 doublet features classified as "true" was 0 ± 2. A multimodal frequency distribution in the retention time H $-$ L is observed for these features and is hypothesized to be related to the to the TopN spacing.

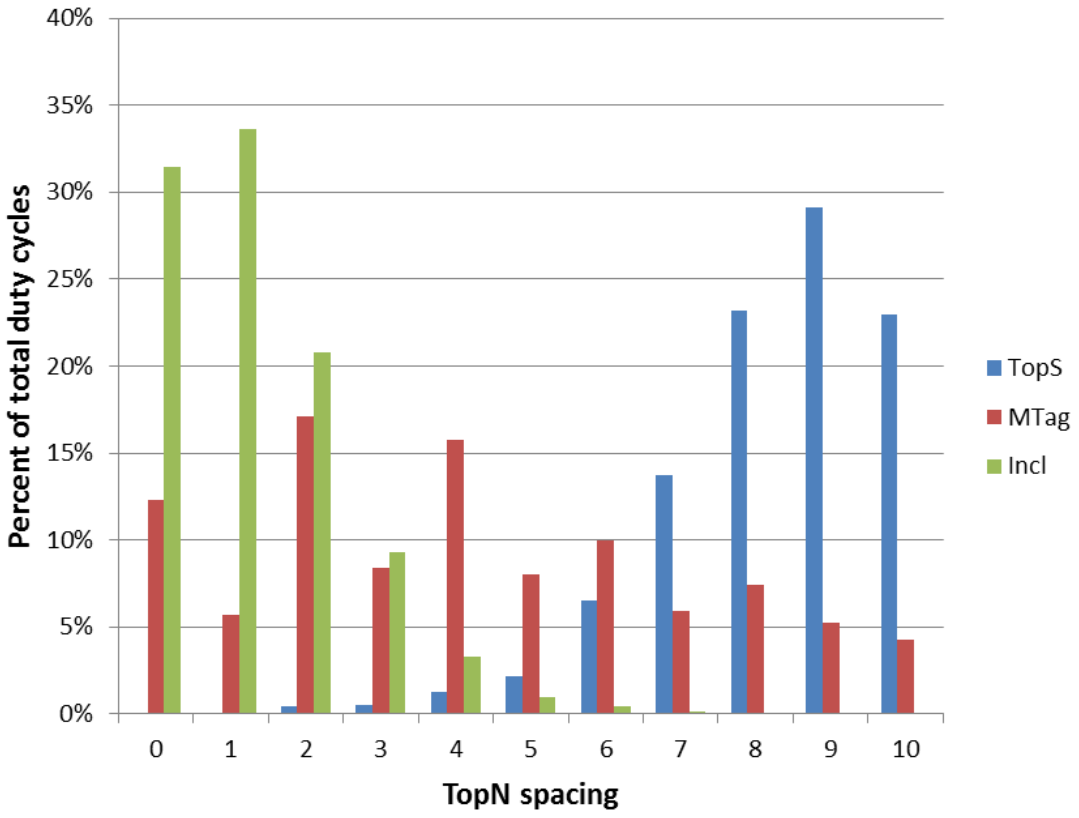


Figure 4.4: The percent of total duty cycles between retention time 20 min and 80 min with N MS^2 scans (TopN spacing) for soluble pre-fraction SCX fraction #16 is represented for each acquisition method (TopS, MTag, or Incl).

The duty cycle is frequently reaching the cycle time limits with TopS method (here limited by TopN spacing = 10 or time = 3 sec) during what is expected to be the portion of the LC-MS analysis that is most abundant in crosslinker-modified peptides (retention times between 20 and 80 minutes). With the use of a CL-specific targeted acquisition methods (i.e., MTag or Incl), the duty cycle reaches the cycle limit more infrequently than with the untargeted acquisition method (i.e., TopS).

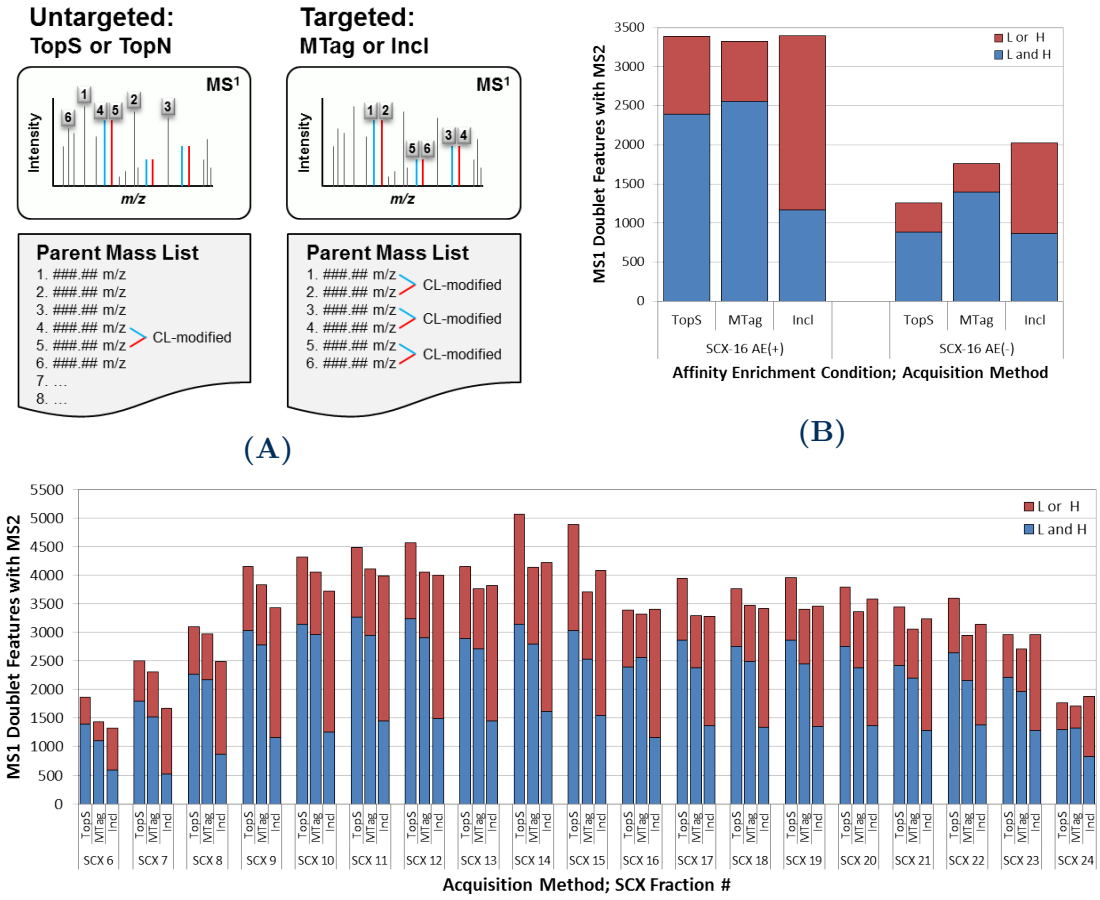


Figure 4.5: Targeted acquisition improves the coverage of crosslinker-modified peptides.

(A) Diagram of untargeted and targeted acquisition methods. Both method types have precursors selected for MS/MS acquisition in order of MS^1 signal intensity, but with a targeted method, the precursors must also be part of a $\Delta\,8.0502\,\mathrm{Da}$ doublet to be selected. (B) A comparison of the number of $\Delta\,8.0502\,\mathrm{Da}$ doublet features found in the MS^1 spectrum of soluble pre-fraction SCX fraction #16, which had corresponding MS^2 spectra in the untargeted (TopS) and targeted (MTag, and Incl) acquisition methods, revealed that the MS^2 spectra of a larger number of crosslinker-modified precursors were acquired when targeted methods were used on sample fractions that had not been affinity enriched than on those fractions that had been affinity enriched. (C) A comparison of the number of $\Delta\,8.0502\,\mathrm{Da}$ doublet features found in all SCX fractions that had been affinity enriched showed no benefit of targeted methods over the untargeted method for crosslinker-modified precursor acquisition. Data shown here is solely from the soluble pre-fraction.

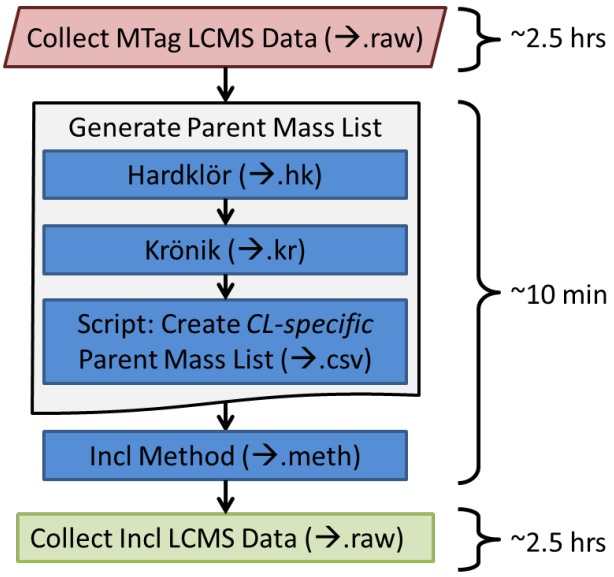


Figure 4.6: Schematic of method and approximate time required for acquisition of MTag and Incl datasets for a sample fraction.

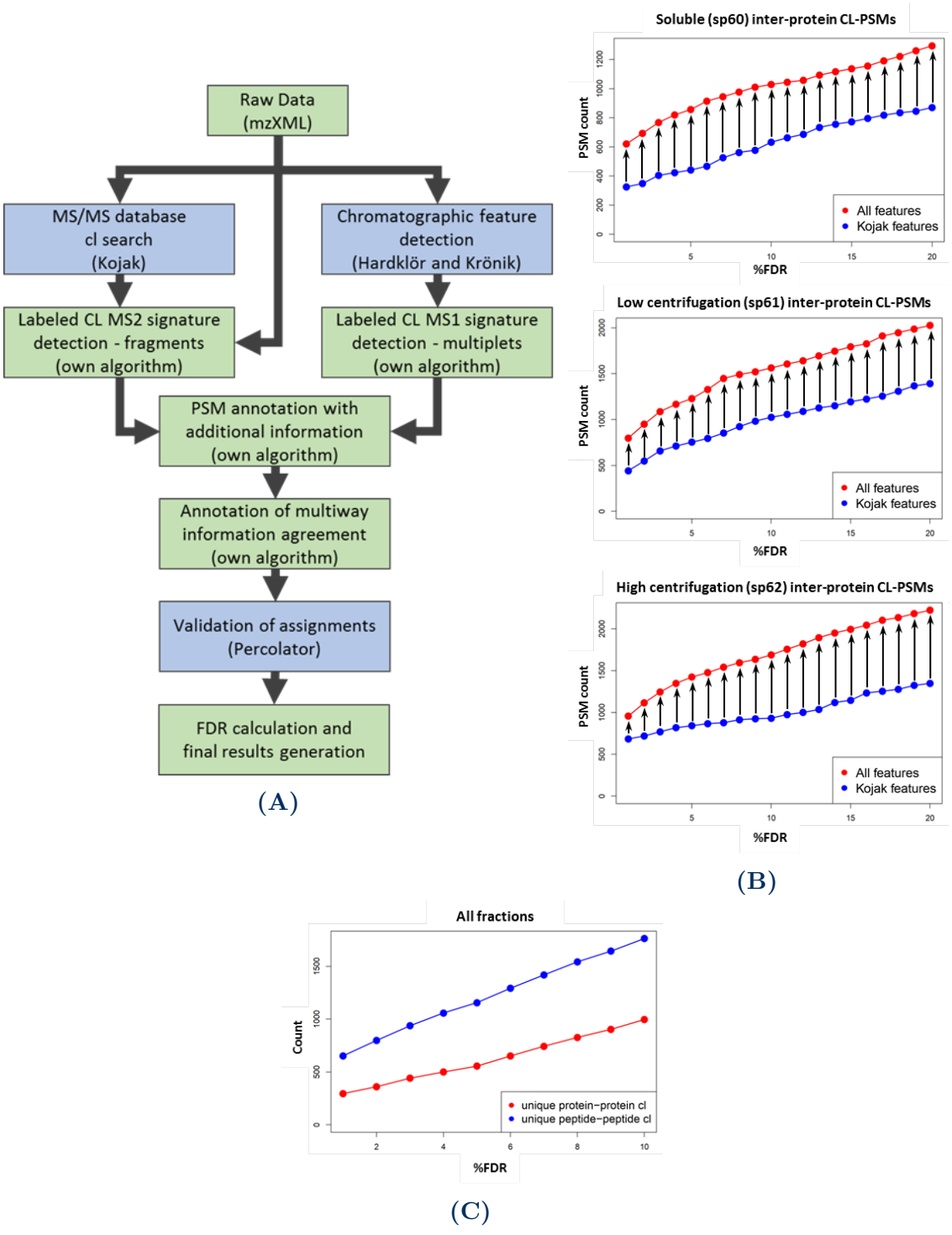
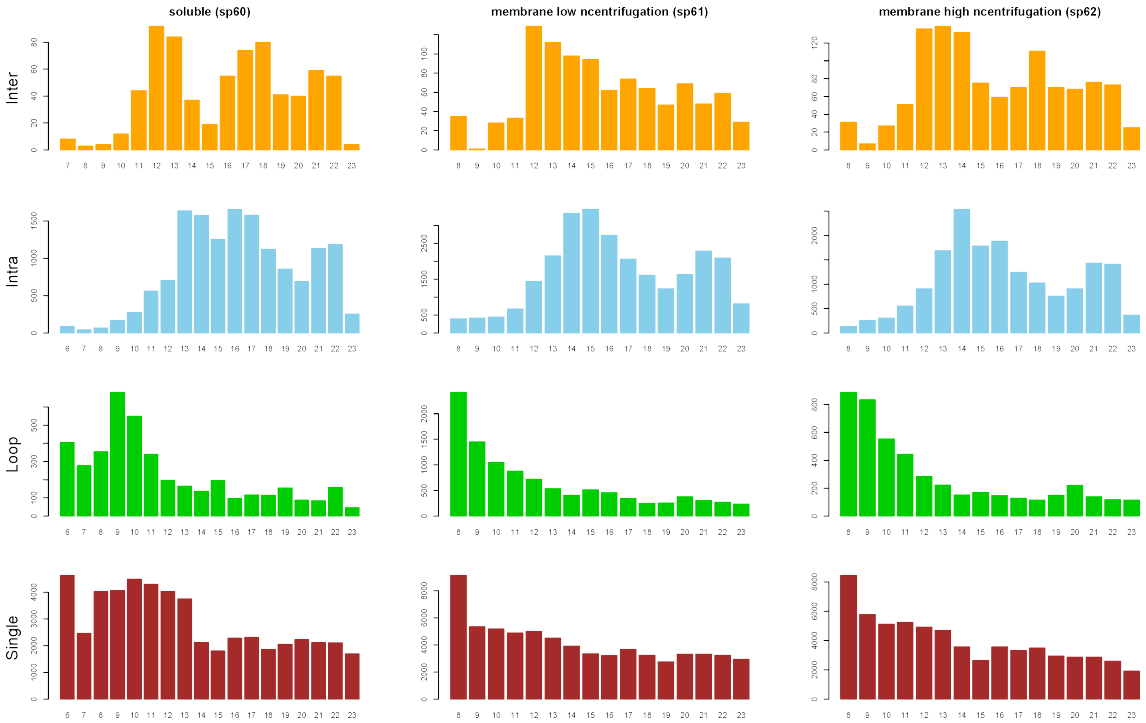
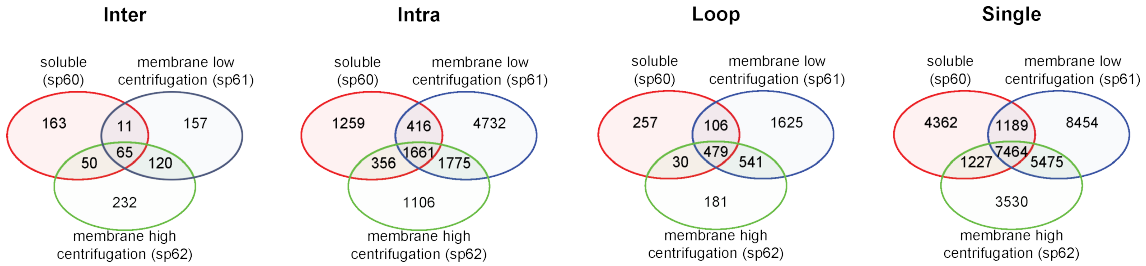


Figure 4.7: Crosslinker-specific mass spectrum features improve cross-linker-modified peptides identification. *(caption continues on next page)*

Figure 4.7: *(caption continued from previous page)*
(A) MS data is passed into a software pipeline that generates PSMs, extracts MS^1 feature information, adds additional CL-specific feature information to each PSM, executes PSM validation, and returns validated PSMs. (B) The number of identified inter-protein CL-PSMs as a function of %FDR are shown for PSM validation outputs from Percolator training with input using either the original set of Kojak PSM validation features, or the full set of PSM validation features from the data analysis pipeline. An increase in identified CL-PSMs was observed across all %FDR levels (0 %–20 % shown). (C) The total number of unique protein-protein interactions and unique residue-residue crosslink identifications in all datasets combined is shown as a function of %FDR.



(A)



(B)

Figure 4.8: Overview of the identifications.

(A) The total number of (crosslinked-)peptide-spectrum-matches in each centrifugation fraction – three columns, and each SCX fraction – in numbers below each barplot. (B) The overlap in identifications between the three centrifugation fractions. Crosslinks are divided into four types: inter- and intra-protein crosslinks, as well as loop and single peptide identifications.

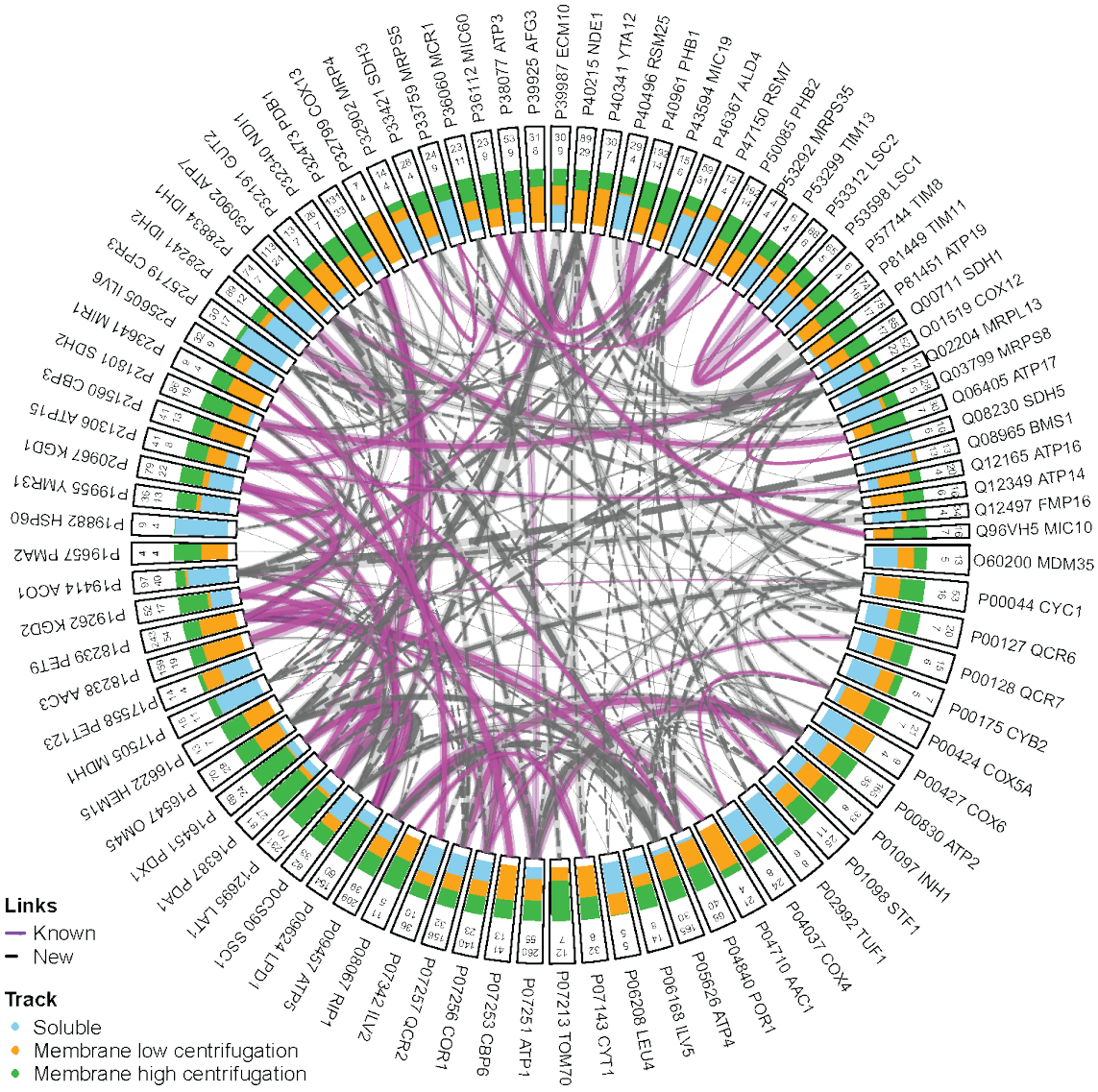
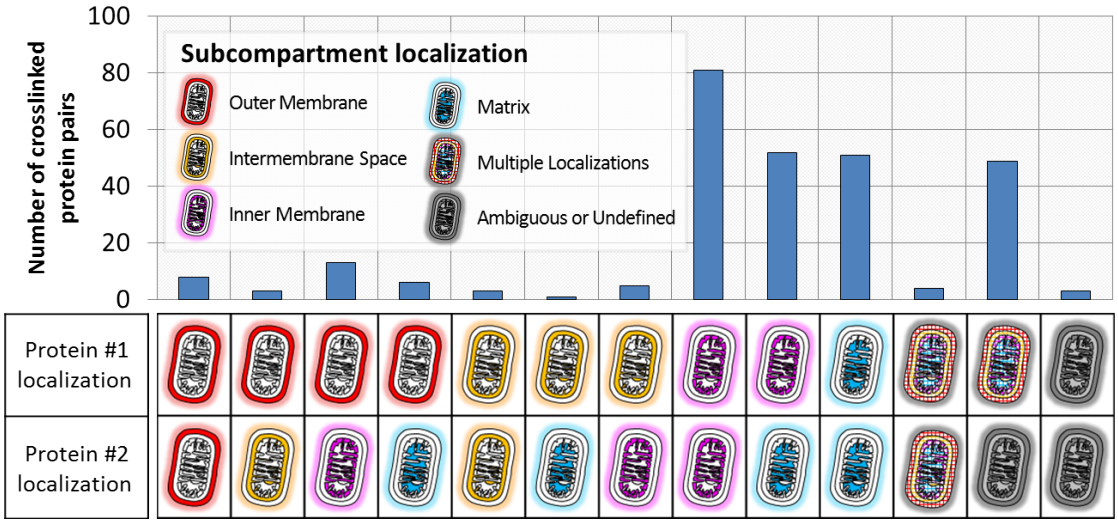


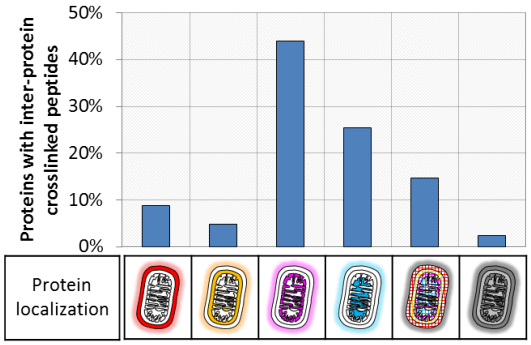
Figure 4.9: Circle diagram of the protein-protein interaction network analysis and sub-compartment localization of the identified crosslinks.

Unique inter-protein crosslinks identified at 2 % FDR are represented for those proteins with a minimum of 4 unique residue-residue crosslinks. Classification of a protein interaction (edges) as "known" or "new" was based on the EBI database of known yeast mitochondrial PPIs (retrieved: 2017-12-14) [169]. Starting from the outside, each node is labeled with the UniProt accession number followed by the gene name. The next number represents the total number of PSMs associated with that protein, followed by the number of unique residue-residue crosslinks associated with that protein. *(caption continues on next page)*

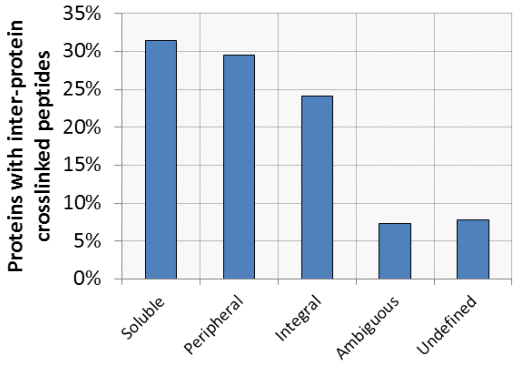
Figure 4.9: *(caption continued from previous page)*
The green, orange, and blue bars inside each rectangle indicate in which sample pre-fractions the respective protein was identified. The width of the edges (i.e., the lines connecting each node) represents the proportional number of all PSMs associated with the respective nodes with the number of validated PSMs between two proteins represented in semi-transparent highlighting and the number of unique residue-residue pairs represented by solid lines.



(A)



(B)



(C)

Figure 4.10: Protein-protein interaction network analysis and sub-compartment localization of the identified crosslinks.

(A) Distribution of sub-compartment localizations (based on Vögtle et al. [161]) for pairs of unique protein-protein interactions identified in this study (FDR 2 %). (B) Distribution of the protein classification for all proteins identified in inter-protein crosslinks. (C) Distribution of sub-compartment localizations for proteins with identified inter-protein crosslinks (FDR 2 %).

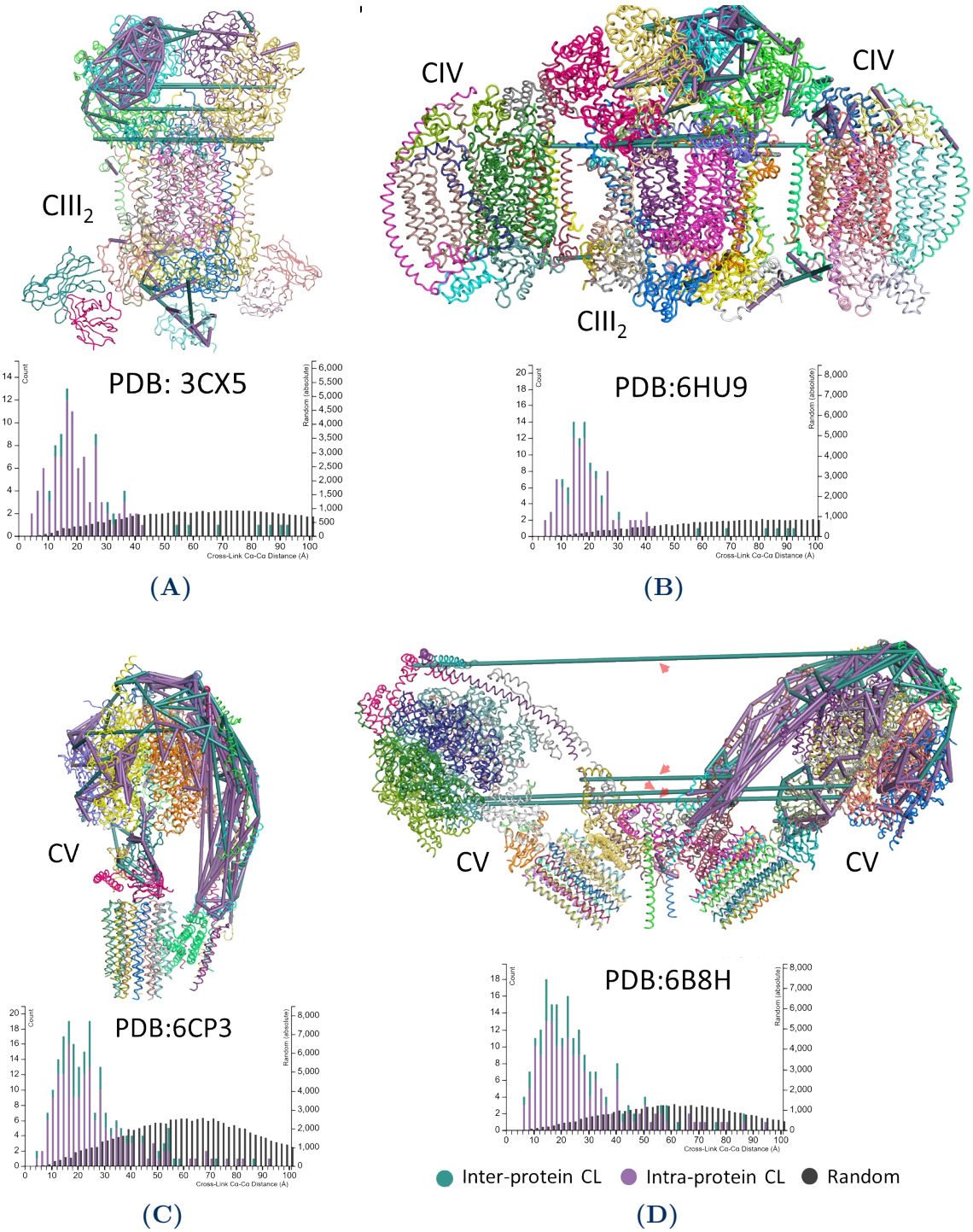


Figure 4.11: Identified crosslinks mapped to PDB structures of yeast mitochondrial electron transport chain complexes and super-complexes. *(caption continues on next page)*

Figure 4.11: *(caption continued from previous page)*
(A) Mapping of identified crosslinks to complex III_2 (PDB ID: 3CX5), (B) complex V (PDB ID: 6CP3), (C) respiratory super-complex III_2IV_2 (PDB ID: 6HU9), and (D) to complex V dimer (PDB ID: 6B8H). All panels are accompanied with a histogram of observed Cα-Cα distance distributions versus distances of random possible links. Inter-protein crosslinks are shown as green lines and intra-protein crosslinks are shown as purple lines. In cases in which a crosslink may be drawn multiple times (e.g., in each monomer of a homodimer) only the shortest constraint is shown. Red arrows in panel (D) indicate long crosslink distances ranging from 100 to 325 Å, however in alternative arrangements of two or more complex V dimers, the monomers are adjacent (side-by-side) with a distance of approximately 130 Å between their rotational axis [168]. In this arrangement an expected distances closer to 38 Å between the crosslinked sites may exist, however, PDB structures for this arrangement are not available.

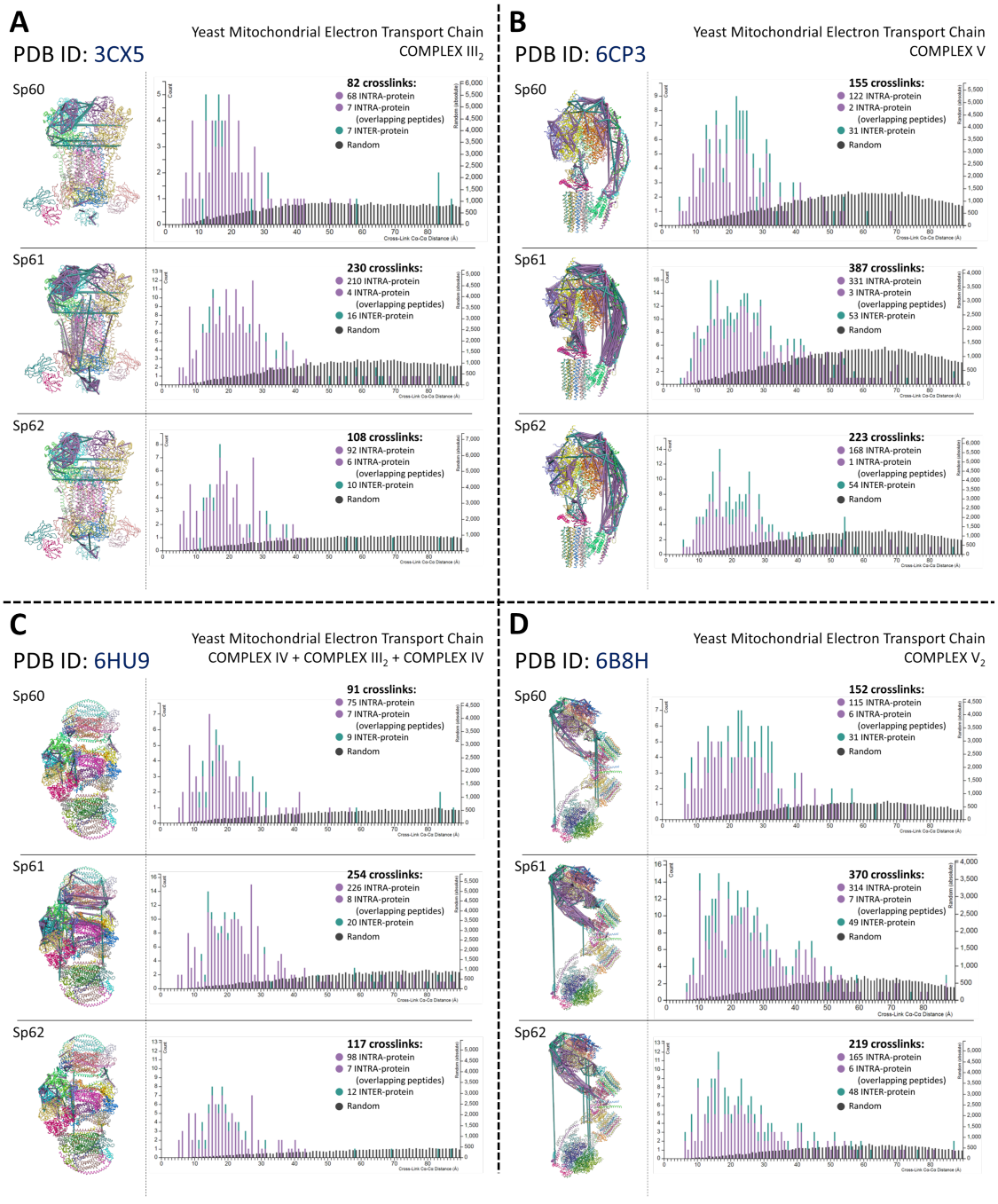


Figure 4.12: Identified crosslinks mapped to PDB structures of yeast mitochondrial electron transport chain complexes and super-complexes for all sample pre-fractions. *(caption continues on next page)*

Figure 4.12: *(caption continued from previous page)*
(A) Mapping of identified crosslinks to complex III_2 (PDB ID: 3CX5), (B) complex V (PDB ID: 6CP3), (C) respiratory super-complex III_2IV_2 (PDB ID: 6HU9), and (D) to complex V dimer (PDB ID: 6B8H). Panels are accompanied with a histogram of observed Cα-Cα distance distributions versus distances of random possible links. Inter-protein crosslinks are shown as green lines and intra-protein crosslinks are shown as purple lines. In cases in which a crosslink may be drawn multiple times (e.g., in each monomer of a homodimer) only the shortest constraint is shown.

CHAPTER

FIVE

GENERAL DISCUSSION

5.1 Summary

My overarching research objective throughout this dissertation has been to explore, develop, and apply MS-based structural proteomic techniques, primarily crosslinking mass spectrometry (CL-MS), toward understanding protein structure and protein-protein interaction (PPI) networks. I began in Chapter 2 by describing a method for identifying naturally occurring crosslinks, in the form of endogenous disulfide bonds, using MS. I then described an application of CL-MS in which crosslinks are synthetically introduced to a protein and, following MS analysis, information about the covalently linked residue-residue pairs is utilized in molecular dynamics simulations to create a *de novo* structural model of the intrinsically disordered protein tau in Chapter 3. I concluded in Chapter 4 by describing a further extension of the CL-MS technique toward elucidation of a large interactome in the form of isolated yeast mitochondria.

Altogether, this dissertation describes new and generally applicable approaches for structural studies of proteins and protein-protein interactions in sample contexts ranging from isolated proteins to intact organelles. The methods also provide new insights into the structures and structural dynamics of a variety of proteins and protein systems.

5.1.1 Chapter 2: Disulfide bonds, non-specific digestion, and higher-order CL-MS

Objectives and contributions The specific objectives achieved in Chapter 2 were to develop an improved method for determining endogenous crosslinks with a bottom-up proteomics analysis. This was accomplished using an algorithm that allowed for second-order crosslink analysis (i.e., three inter-linked peptides) and crosslinker-cleavage (i.e., fragmentation at the crosslinking group between crosslinked peptides) fragment ion matching. Also, alternative proteases were shown to improve coverage of disulfide linkages as well as increasing the amount of corroborating spectral evidence for individual disulfide linkage identifications. Using this approach I was able to identify the complete known disulfide bond arrangements for a variety of well characterized proteins: bovine serum albumin, bovine pancreatic RNase S, human fibrinogen (α, β, and γ chains), human insulin (α and β chains), and chicken lysozyme. In all cases, analysis of only tryptic first-order crosslinks did not yield the complete set of known disulfide bonds. Complete determination, however, was achieved by including the results from non-specific proteinase K digest analysis, as well as the second-order crosslinking analysis. The treatment of disulfide bonds as CID-cleavable crosslinks was key in distinguishing true-positives from false-positives. Proteinase K, on average, yielded more first-order disulfide linkages than trypsin, and, in many cases, these linkages were also confirmed multiple times due to existence of

a "family" of related proteinase-K-cleaved peptides. In cases where the cysteines were adjacent or were separated by less than 4 residues, the second-order analysis was the key to determining the disulfide bonding arrangement.

Chapter 2 reflections and future directions:

Widening support for higher-order crosslinking analysis A number of aspects in this work presented in Chapter 2 could be further explored and expanded upon. For instance, virtually all current mainstream CL-MS identification algorithms are only equipped to consider peptides, zero-order (i.e., peptides with intra-peptide crosslinks), and first-order (i.e., two inter-linked peptides) crosslinks. Incorporating the capability of second-order or higher-order crosslinking consideration into these, or future algorithms could improve the depth of structural information that is accessible in any given CL-MS analysis. Recently, a similar idea has been explored in which a tetra-reactive crosslinking reagent, with the capability of simultaneously crosslinking four proximal sites and resulting in four inter-linked peptides following protein digest, was developed and demonstrated [64]. Such an approach can yield highly specific distance constraints and makes possible the unambiguous identification of protein complex interfaces that involve greater than two proteins.

Improvement of FDR estimation procedure An idiosyncratic heuristic false-discovery rate (FDR) estimation was applied to the identified disulfide crosslinks of Chapter 2. In retrospect, and later described in Chapters 3 and 4, a more rigorous, statistically-based FDR estimation approach would have been more appropriate in discriminating real from spurious crosslink identifications, especially in such an expanded search space as occurs with non-specific protease digestion or the consideration of three inter-linked peptides per peptide-spectrum match (PSM). The same approach as was used in the later chapters could be applied to the disulfide

crosslinks and also second-order disulfide crosslinks with no special considerations. Additionally, taking this updated approach would eliminate the time-intensive, labour-intensive, capricious, and error-prone necessity of manual inspection of each MS spectra in order to make crosslink identifications. This would allow the technique to be more widely adopted without requiring experienced mass spectrometrists. Mass spectral features specific to second-order crosslinks could also be determined and included into the FDR estimation approach similar to what was developed and described in Chapter 4 in which a semi-supervised machine learning algorithm was utilized together with several crosslinker-specific spectral features in order to increase confident crosslink identification rates.

Algorithm improvement The algorithm which was used to generate the PSMs for the disulfide crosslinks was exhaustive (i.e., considered and scored every possible PSM that fit within a specified precursor ion ppm mass tolerance) making this approach feasible only with small protein sequence database sizes and specific proteases (i.e., trypsin) or with a further constrained (i.e., incomplete or fragmented) protein sequence database when non-specific digest is used. An improved implementation of this algorithm that could enable much larger sequence databases and non-specific protein digests to be considered at same time as increasing both crosslink identification specificity and sensitivity could involve utilization of crosslinker-cleavage (here, disulfide bond cleavage) information at the time of PSM generation and scoring [173, 59]. To my knowledge, an implementation of such an algorithm that considers higher-order crosslinks does not yet exist.

Alternative dissociation methods Although CID was the particular dissociation method used to fragment the disulfide-linked peptides, alternative dissociation methods may provide superior fragmentation efficiency and richer fragment ion

spectra. Strategies utilizing multiple fragmentation methods, such as electron transfer dissociation (ETD) or ultraviolet photodissociation (UVPD), to improve the analysis of crosslinked peptides have been reported [63, 174] and application of such an approach would be immediately applicable to the disulfide bond determination method I have described in Chapter 2. As there are potentially many more fragments that can result from higher-order crosslinked peptide precursor ions, use of such alternative dissociation methods may significantly increase the amount of information present in fragment ion spectra which could lead to improved identification rates of these types of crosslinks especially.

5.1.2 Chapter 3: Tau, short-distance CL-MS, and computational structural modeling

Objectives and contributions The main objective of Chapter 3 was to structurally characterize the intrinsically disordered protein tau in its native monomeric form by combining structural proteomic experimental data with computational methods. To accomplish this, a panel of short-distance crosslinkers was used which yielded several crosslink identifications that were subsequently used in guiding a discrete molecular dynamics simulation algorithm to model the monomeric tau structure. The resulting models suggested some degree of tertiary structure in tau. These models were corroborated with additional structural proteomic data and also compared to existing structural knowledge about tau in which some agreement was found in regions where transient secondary structure had been reported. This structural description of tau may serve as a basis for hypothesis generation for future studies into pathological tau misfolding and aggregation.

Chapter 3 reflections and future directions:

"High-density" crosslinking and online ion mobility separation Although models for tau were generated using the set of constraints identified here, the set by no means represents a complete set of what crosslinks might be identified with additional and repeated short-range SDA CL-MS analyses. In a paper from Belsom et al. [175] it was shown that in similar experiments with SDA crosslinking of human serum albumin that less than half of the unique residue-residue constraints identified in a set of 87 technical replicates were discovered by the 13^{th} analysis. The tau monomer SDA constraints presented in Chapter 3 derive from 6 individual analyses and therefore could be expected to increase significantly with more experiments. Conducting this experiment with the aim of achieving high-density crosslinking (i.e., many constraints per residue) could improve structural modelling outcomes [176].

Additionally, the heterobifunctional and semi-non-specific nature of SDA crosslinking, with amine-reactivity on one half and non-specific photoreactivity on the other, invariably leads to a mixture of structurally isomeric crosslinked peptides. For example, a particular pair of crosslinked peptides may only differ at the position in which they are covalently linked via the crosslinker. This is especially likely to occur in SDA crosslinks with the same lysine residue on one peptide of the pair being found linked to a set of adjacent residues on the second peptide. Because these structural isomers are, by definition, isobaric (i.e., identical in mass), if not separated prior to mass analysis, chimeric MS^2 fragment ion spectra (i.e., spectra with fragments originating from more than one distinct molecular species) occur. Some chromatographic separation of these isomeric species does occur with the reversed-phase liquid chromatography (LC) set-up that is used online with typical LC-MS analyses, however, due to the high-degree of chemical and structural similarity this separation may not be sufficient to avoid acquisition of chimeric MS^2 spectra.

Recently, MS instruments with high analytical performance together with additional capabilities for online separation of analytes in an ion mobility dimension have come to market [177].

By taking advantage of the ion mobility separation the precursor ions of structural isomers can be better isolated prior to MS^2 analysis. Indeed, these instruments have recently been successfully utilized for analysis in homobifunctional CL-MS experiments [178]. Use of these instruments will likely prove even more beneficial for mitigating acquisition of chimeric MS^2 for heterobifunctional SDA crosslinking analyses which can be employed in a high-density crosslinking fashion.

Experimental controls and comparison conditions Future improvements to this experiment could include additional control conditions in which the tau protein is crosslinked in a denaturing environment in order to tease apart which crosslinks may occur spuriously from those crosslinks which represent constraints for residues that do indeed occur in the native in solution structure of tau, however transiently. Additionally, such an experiment may be carried out in a quantitative crosslinking (qCL) fashion similar to what I have reported in recent papers involving protein chaperone conformational changes upon client binding [179, 180]. Here, depending upon the degree to which a crosslink is differentially observed between two comparison conditions (e.g., denaturing versus non-denaturing) a bias parameter may be assigned to the residue-residue distance constraint to weight its relative importance in the computational modelling procedure. This may result in modelling output that more accurately reflects what conformations, or conformational ensembles, are preferred amongst a set of distinct protein folds.

5.1.3 Chapter 4: Improving the detection, acquisition, and identification of crosslinks in large-scale CL-MS analyses

Objectives and contributions Finally, the primary objective of Chapter 4 was to develop and demonstrate a method that could address challenges inherent with the MS analysis of synthetically formed crosslinks in a complex system, yeast mitochondria. This was accomplished by taking advantage of several characteristics that were designed into the CBDPS crosslinking reagent [82]. By taking advantage of affinity enrichment, isotopic-coding, and MS-cleavability improvements were observed in crosslink detection, MS^2 acquisition, and crosslink identification. A novel data analysis pipeline enabling this was also developed and described.

Chapter 4 reflections and future directions:

Online ion mobility crosslink separation and feature information Since developing this approach a number of new technologies have been developed which may further improve this method. As mentioned above (see 5.1.2) MS instrumentation with online ion mobility separation capabilities has come to market and been successfully used in CL-MS analyses. The approach described in Chapter 4 could be performed using this instrumentation and would conceivably benefit in the same ways as described above. Additionally, analytical information afforded by the ion mobility separation of precursor ions could supplement the spectral features (see Table 4.5) used in the semi-supervised machine-learning-based FDR procedure to better discriminate true crosslink identifications from false.

Senstivity improvement with "BoxCar" full scan mode Another recently developed technology that could be taken advantage of to improve the approach, with

no other changes necessary, would be the use of a "BoxCar" full scan (i.e., MS^1) mode for Orbitrap instruments. This new scan mode effectively extends the dynamic range of the instrument on the MS^1 level by greater than one order of magnitude [181]. This is accomplished by filling the C-trap with ions in multiple ion transmission steps each consisting of interspaced mass selection windows prior to analyzing all collected ions in a single Orbitrap scan. This dramatically improves the dynamic range over the standard full scan mode essentially by creating a more "democratic" representation of ions across the entire m/z range due to high abundance ions no longer being able to dominate the C-trap's limited charge capacity. This technical advance in improved sensitivity is particularly auspicious for CL-MS experiments due to the fact that crosslinked peptides are often of very low relative abundance when compared with co-eluting non-crosslinked peptides. It should be noted that a penalty in ion injection times is paid when using the BoxCar scan mode [182]. This means that duty cycle times will increase and the total count of MS^2 scans acquired is likely to drop. In a TopN or TopS type acquisition method there may be limited or no benefit to crosslinked peptide precursor acquisition in this case. In fact, when using a TopS acquisition method I routinely found the Orbitrap instrument with the fastest scan rate available at the time (Orbitrap Fusion™ Lumos™ Tribrid™ Mass Spectrometer) hitting duty cycle time limits even with standard full scan mode settings (see Figure 4.4). However, if one were to use a targeted method for MS^2 acquisition such as the MTag or Incl methods I described in Chapter 4, then the benefit of the improved sensitivity could be better utilized. By implementing this novel scan mode it is likely that an immediate improvement in the number of crosslinked peptides will be observed in Orbitrap CL-MS analyses.

Multidimensional false-discovery rates The standard way of reporting an FDR for any given CL-MS analysis has been, and seemingly continues to be at present time,

solely on the level of PSMs. Although this is very important to ensure the veracity of a set of peptide or crosslink identifications, FDR estimates for other, arguably more biologically important measures, could also be calculated and reported. For example, Fischer et al. [183] have shown that estimation of error rates at the levels PSMs, peptide pairs, residue-residue linkage pairs, and protein-protein interaction pairs are distinctly important and when considered together (and numerically explored) can reveal a more sensitive outcome for a given CL-MS analysis. For proteome-wide CL-MS analyses in which the results of key importance are the residue-residue or protein-protein pairs that indicate spatial proximity and biological interaction it would be most beneficial to those concerned with the biological implications of the data to have FDR estimates on these terms. Reporting of these multidimensional FDR metrics will be an important consideration for future development of any CL-MS data analysis pipeline.

Quantitative large-scale CL-MS Beyond probing complex biological contexts to generate a static picture of the PPI landscape as demonstrated in Chapter 4, performing such CL-MS analyses with an added quantitative dimension represents the next evolution for the technique. Quantifying crosslinks at a proteome-wide scale will allow for protein conformational and interaction dynamics to be probed in response to any number of biologically interesting perturbations (e.g., drug treatment, disease status, and genetic manipulations). Various approaches to performing qCL experiments have been demonstrated in recent years however, lack of bioinformatic tools or access to specialized hardware has been cited as an existing barrier to more widespread adoption [184]. Further development of CL-MS data analysis pipelines should consider implementing support for such analyses or generate output that may be readily integrated into existing software dedicated to the analysis of the quantitative aspects of these experiments (e.g., XiQ [185], xTract [186], Skyline [187],

or MaxQuant [188, 189]).

5.2 CL-MS evolution and horizons

As can be seen throughout the work presented in this dissertation, the analysis and reporting of CL-MS has evolved significantly. Going from manually interrogated crosslink identifications, siloed datasets (i.e., no open public access to underlying raw MS data), and idiosyncratically determined heuristics for FDR estimation in Chapter 2 to fully automated and open source identification of crosslinks [58], publicly deposited [126, 128] and available datasets, and using an open source algorithm [65] for statistically-based FDR estimation in Chapters 3 and 4. Indeed, the field of CL-MS is still relatively new but it is now reaching a level of maturity where harmonization and standardization efforts are being discussed throughout the community of researchers who employ structural proteomic techniques and have strong interests in the continued refinement of the CL-MS technique in particular [39, 190]. These efforts will become increasingly important and necessary as CL-MS is included into an increasing number of studies employing integrative structural biology approaches [191, 192] and as proteome-wide quantitative CL-MS approaches are poised to reveal deep insights into the molecular details of PPI networks in health and disease [49, 193, 194, 195].

BIBLIOGRAPHY

[1] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, April 1953.

[2] F. H. Crick. On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163, 1958.

[3] Francis Crick. Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563, August 1970.

[4] Valerie C. Wasinger, Stuart J. Cordwell, Anne Cerpa-Poljak, Jun X. Yan, Andrew A. Gooley, Marc R. Wilkins, Mark W. Duncan, Ray Harris, Keith L. Williams, and Ian Humphery-Smith. Progress with gene-product mapping of the Mollicutes:Mycoplasma genitalium. *Electrophoresis*, 16(1):1090–1094, 1995.

[5] G. J. Mulder. Ueber die Zusammensetzung einiger thierischen Substanzen. *Journal für Praktische Chemie*, 16(1):129–152, 1839.

[6] Chunaram Choudhary and Matthias Mann. Decoding signalling networks by mass spectrometry-based proteomics. *Nature Reviews Molecular Cell Biology*, 11(6):427–439, June 2010.

[7] L. Pauling and R. B. Corey. Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets. *Proceedings of the National Academy of Sciences*, 37(11):729–740, November 1951.

[8] D. Eisenberg. The discovery of the α-helix and β-sheet, the principal structural features of proteins. *Proceedings of the National Academy of Sciences*, 100(20):11207–11210, September 2003.

[9] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, December 1983.

[10] Dmitrij Frishman and Patrick Argos. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Genetics*, 23(4):566–579, December 1995.

[11] Jane S. Richardson. Schematic drawings of protein structures. In *Methods in Enzymology*, volume 115, pages 359–380. Elsevier, 1985.

[12] Warren L DeLano. Pymol: An open-source molecular graphics tool. *CCP4 Newsletter on protein crystallography*, 40(1):82–92, 2002.

[13] A. Fiser and A. Sali. ModLoop: Automated modeling of loops in protein structures. *Bioinformatics*, 19(18):2500–2501, December 2003.

[14] Leslie Regad, Juliette Martin, Gregory Nuel, and Anne-Claude Camproux. Mining protein loops using a structural alphabet and statistical exceptionality. *BMC Bioinformatics*, 11(1):75, December 2010.

[15] I. M.A. Nooren. NEW EMBO MEMBER'S REVIEW: Diversity of protein-protein interactions. *The EMBO Journal*, 22(14):3486–3492, July 2003.

[16] James R. Perkins, Ilhem Diboun, Benoit H. Dessailly, Jon G. Lees, and Christine Orengo. Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. *Structure*, 18(10):1233–1243, October 2010.

[17] Yipeng Wang and Eckhard Mandelkow. Tau in physiology and pathology. *Nature Reviews Neuroscience*, 17(1):22–35, January 2016.

[18] Michel Goedert and Maria Grazia Spillantini. Propagation of Tau aggregates. *Molecular Brain*, 10(1):18, December 2017.

[19] J.J. Thomson. XXVI. *Rays of positive electricity*. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 21(122):225–249, February 1911.

[20] J.J. Thomson. XIX. *Further experiments on positive rays*. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 24(140):209–253, August 1912.

[21] RG Cooks and AL Rockwood. The Thomson-A suggested unit for mass spectroscopists. *Rapid Communications in Mass Spectrometry*, 5(2):93–93, 1991.

[22] Matthias Wilm and Matthias Mann. Analytical Properties of the Nanoelectrospray Ion Source. *Analytical Chemistry*, 68(1):1–8, January 1996.

[23] I Kaltashov and R Abzalimov. Do Ionic Charges in ESI MS Provide Useful Information on Macromolecular Structure? *Journal of the American Society for Mass Spectrometry*, 19(9):1239–1246, September 2008.

[24] Michaela Scigelova, Martin Hornshaw, Anastassios Giannakopulos, and Alexander Makarov. Fourier Transform Mass Spectrometry. *Molecular & Cellular Proteomics*, 10(7):M111.009431, July 2011.

[25] Jesper V Olsen, Boris Macek, Oliver Lange, Alexander Makarov, Stevan Horning, and Matthias Mann. Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods*, 4(9):709–712, September 2007.

[26] Lekha Sleno and Dietrich A. Volmer. Ion activation methods for tandem mass spectrometry. *Journal of Mass Spectrometry*, 39(10):1091–1112, October 2004.

[27] P. Roepstorff and J. Fohlman. Letter to the editors. *Biological Mass Spectrometry*, 11(11):601–601, November 1984.

[28] David L. Tabb, Lori L. Smith, Linda A. Breci, Vicki H. Wysocki, Dayin Lin, and John R. Yates. Statistical Characterization of Ion Trap Tandem Mass Spectra from Doubly Charged Tryptic Peptides. *Analytical Chemistry*, 75(5):1155–1163, March 2003.

[29] Gary L. Glish and David J. Burinsky. Hybrid mass spectrometers for tandem mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 19(2):161–172, February 2008.

[30] Jesper V. Olsen, Lyris M. F. de Godoy, Guoqing Li, Boris Macek, Peter Mortensen, Reinhold Pesch, Alexander Makarov, Oliver Lange, Stevan Horning, and Matthias Mann. Parts per Million Mass Accuracy on an Orbitrap Mass Spectrometer via Lock Mass Injection into a C-trap. *Molecular & Cellular Proteomics*, 4(12):2010–2021, December 2005.

[31] Yaoyang Zhang, Bryan R. Fonslow, Bing Shan, Moon-Chang Baek, and John R. Yates. Protein analysis by shotgun/bottom-up proteomics. *Chemical Reviews*, 113(4):2343–2394, April 2013.

[32] Alicia L Richards, Alexander S Hebert, Arne Ulbrich, Derek J Bailey, Emma E Coughlin, Michael S Westphall, and Joshua J Coon. One-hour proteome analysis in yeast. *Nature Protocols*, 10(5):701–714, May 2015.

[33] Adelinda Yee, Keith Pardee, Dinesh Christendat, Alexei Savchenko, Aled M. Edwards, and Cheryl H. Arrowsmith. Structural Proteomics: Toward High-Throughput Structural Biology as a Tool in Functional Genomics. *Accounts of Chemical Research*, 36(3):183–189, March 2003.

[34] Rebecca Beveridge and Antonio N. Calabrese. Structural Proteomics Methods to Interrogate the Conformations and Dynamics of Intrinsically Disordered Proteins. *Frontiers in Chemistry*, 9:603639, March 2021.

[35] Christopher Clegg and Donal Hayes. Identification of Neighbouring Proteins in the Ribosomes of Escherichia coli. A Topographical Study with the Cross-Linking Reagent Dimethyl Suberimidate. *European Journal of Biochemistry*, 42(1):21–28, February 1974.

[36] Tung-Tien Sun, Alex Bollen, Lawrence Kahan, and Robert R. Traut. Topography of ribosomal proteins of the Escherichia coli 30S subunit as studied with reversible cross-linking reagent methyl 4-mercaptobutyrimidate. *Biochemistry*, 13(11):2334–2340, May 1974.

[37] E H Manting, C van der Does, and A J Driessen. In vivo cross-linking of the SecA and SecY subunits of the Escherichia coli preprotein translocase. *Journal of bacteriology*, 179(18):5699–5704, 1997.

[38] Éva Kurucz, István Andó, Máté Sümegi, Harald Hölzl, Barbara Kapelari, Wolfgang Baumeister, and Andor Udvardy. Assembly of the Drosophila 26 S proteasome is accompanied by extensive subunit rearrangements. *Biochemical Journal*, 365(2):527–536, July 2002.

[39] Alexander Leitner, Alexandre M.J.J. Bonvin, Christoph H. Borchers, Robert J. Chalkley, Julia Chamot-Rooke, Colin W. Combe, Jürgen Cox, Meng-Qiu Dong, Lutz Fischer, Michael Götze, Fabio C. Gozzo, Albert J.R. Heck, Michael R. Hoopmann, Lan Huang, Yasushi Ishihama, Andrew R. Jones, Nir Kalisman, Oliver Kohlbacher, Karl Mechtler, Robert L. Moritz, Eugen Netz, Petr Novak, Evgeniy Petrotchenko, Andrej Sali, Richard A. Scheltema, Carla Schmidt, David Schriemer, Andrea Sinz, Frank Sobott, Florian Stengel, Konstantinos Thalassinos, Henning Urlaub, Rosa Viner, Juan A. Vizcaíno, Marc R. Wilkins, and Juri Rappsilber. Toward Increased Reliability, Transparency, and Accessibility in Cross-linking Mass Spectrometry. *Structure*, 28(11):1259–1268, November 2020.

[40] Roberto Vanacore, Amy-Joan L. Ham, Markus Voehler, Charles R. Sanders, Thomas P. Conrads, Timothy D. Veenstra, K. Barry Sharpless, Philip E. Dawson, and Billy G. Hudson. A Sulfilimine Bond Identified in Collagen IV. *Science*, 325(5945):1230–1234, September 2009.

[41] A. L. Fidler, R. M. Vanacore, S. V. Chetyrkin, V. K. Pedchenko, G. Bhave, V. P. Yin, C. L. Stothers, K. L. Rose, W. H. McDonald, T. A. Clark, D.-B. Borza, R. E. Steele, M. T. Ivy, The Aspirnauts, J. K. Hudson, and B. G. Hudson. A unique covalent bond in basement membrane is a primordial innovation for tissue evolution. *Proceedings of the National Academy of Sciences*, 111(1):331–336, January 2014.

[42] Mitsuru Haniu, Linda O. Narhi, Tsutomu Arakawa, Steve Elliott, and Michael F. Rohde. Recombinant human erythropoietin (rHuEPO): Cross-linking with

disuccinimidyl esters and identification of the interfacing domains in EPO. *Protein Science*, 2(9):1441–1451, September 1993.

[43] M. M. Young, N. Tang, J. C. Hempel, C. M. Oshiro, E. W. Taylor, I. D. Kuntz, B. W. Gibson, and G. Dollinger. High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proceedings of the National Academy of Sciences*, 97(11):5802–5806, May 2000.

[44] Juri Rappsilber, Symeon Siniossoglou, Eduard C. Hurt, and Matthias Mann. A Generic Strategy To Analyze the Spatial Organization of Multi-Protein Complexes by Cross-Linking and Mass Spectrometry. *Analytical Chemistry*, 72(2):267–275, January 2000.

[45] Keiryn L. Bennett, Martin Kussmann, Marie Mikkelsen, Peter Roepstorff, Per Björk, Magdalena Godzwon, and Poul Sörensen. Chemical cross-linking with thiol-cleavable reagents combined with differential mass spectrometric peptide mapping-A novel approach to assess intermolecular protein contacts. *Protein Science*, 9(8):1503–1518, 2000.

[46] Evgeniy V. Petrotchenko, Lars C. Pedersen, Christoph H. Borchers, Kenneth B. Tomer, and Masahiko Negishi. The dimerization motif of cytosolic sulfotransferases. *FEBS Letters*, 490(1-2):39–43, February 2001.

[47] A. Sinz and K. Wang. Mapping protein interfaces with a fluorogenic cross-linker and mass spectrometry: Application to nebulin-calmodulin complexes. *Biochemistry*, 40(26):7903–7913, July 2001.

[48] Thomas Taverner, Nathan E. Hall, Richard A. J. O'Hair, and Richard J. Simpson. Characterization of an Antagonist Interleukin-6 Dimer by Stable Isotope Labeling, Cross-linking, and Mass Spectrometry. *Journal of Biological Chemistry*, 277(48):46487–46492, November 2002.

[49] Juan D. Chavez, Chi Fung Lee, Arianne Caudal, Andrew Keller, Rong Tian, and James E. Bruce. Chemical Crosslinking Mass Spectrometry Analysis of Protein Conformations and Supercomplexes in Heart Tissue. *Cell Systems*, 6(1):136–141.e5, January 2018.

[50] Michelle Trester-Zedlitz, Katsuhiko Kamada, Stephen K. Burley, David Fenyö, Brian T. Chait, and Tom W. Muir. A Modular Cross-Linking Approach for Exploring Protein Interactions. *Journal of the American Chemical Society*, 125(9):2416–2425, March 2003.

[51] Feixia Chu, Sami Mahrus, Charles S. Craik, and Alma L. Burlingame. Isotope-Coded and Affinity-Tagged Cross-Linking (ICATXL): An Efficient Strategy to Probe Protein Interaction Surfaces. *Journal of the American Chemical Society*, 128(32):10362–10363, August 2006.

[52] Alexander Leitner, Roland Reischl, Thomas Walzthoeni, Franz Herzog, Stefan Bohn, Friedrich Förster, and Ruedi Aebersold. Expanding the Chemical Cross-Linking Toolbox by the Use of Multiple Proteases and Enrichment by Size Exclusion Chromatography. *Molecular & Cellular Proteomics*, 11(3):M111.014126, 2012.

[53] Oliver Rinner, Jan Seebacher, Thomas Walzthoeni, Lukas N Mueller, Martin Beck, Alexander Schmidt, Markus Mueller, and Ruedi Aebersold. Identification of cross-linked peptides from large sequence databases. *Nature Methods*, 5(4):315–318, April 2008.

[54] Verena Tinnefeld, A. Saskia Venne, Albert Sickmann, and René P. Zahedi. Enrichment of Cross-Linked Peptides Using Charge-Based Fractional Diagonal Chromatography (ChaFRADIC). *Journal of Proteome Research*, 16(2):459–469, 2017.

[55] Şule Yılmaz, Genet A. Shiferaw, Josep Rayo, Anastassios Economou, Lennart Martens, and Elien Vandermarliere. Cross-linked peptide identification: A computational forest of algorithms. *Mass Spectrometry Reviews*, 37(6):738–749, November 2018.

[56] Evgeniy V. Petrotchenko and Christoph H. Borchers. ICC-CLASS: isotopically-coded cleavable crosslinking analysis software suite. *BMC Bioinformatics*, 11(1):64, December 2010.

[57] Evgeniy V. Petrotchenko, Karl A.T. Makepeace, and Christoph H. Borchers. DXMSMS Match Program for Automated Analysis of LC-MS/MS Data Obtained Using Isotopically Coded CID-Cleavable Cross-Linking Reagents: DXMSMS Match for Analysis of LC-MS/MS Data. In Alex Bateman, William R. Pearson, Lincoln D. Stein, Gary D. Stormo, and John R. Yates, editors, *Current Protocols in Bioinformatics*, pages 8.18.1–8.18.19. John Wiley & Sons, Inc., Hoboken, NJ, USA, December 2014.

[58] Michael R. Hoopmann, Alex Zelter, Richard S. Johnson, Michael Riffle, Michael J. MacCoss, Trisha N. Davis, and Robert L. Moritz. Kojak: Efficient Analysis of Chemically Cross-Linked Protein Complexes. *Journal of Proteome Research*, 14(5):2190–2198, May 2015.

[59] Fan Liu, Dirk T S Rijkers, Harm Post, and Albert J R Heck. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nature Methods*, 12(12):1179–1184, 2015.

[60] Jared P. Mohr, Poorna Perumalla, Juan D. Chavez, Jimmy K. Eng, and James E. Bruce. Mango: A General Tool for Collision Induced Dissociation-Cleavable Cross-Linked Peptide Identification. *Analytical Chemistry*, 90(10):6028–6034, 2018.

[61] Xiaoting Tang and James E. Bruce. A new cross-linking strategy: Protein interaction reporter (PIR) technology for protein–protein interaction studies. *Molecular BioSystems*, 6(6):939, 2010.

[62] Chad R. Weisbrod, Juan D. Chavez, Jimmy K. Eng, Li Yang, Chunxiang Zheng, and James E. Bruce. In Vivo Protein Interaction Network Identified with a Novel Real-Time Cross-Linked Peptide Identification Strategy. *Journal of Proteome Research*, 12(4):1569–1579, 2013.

[63] Fan Liu, Philip Lössl, Richard Scheltema, Rosa Viner, and Albert J. R. Heck. Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. *Nature Communications*, 8(1):15473, August 2017.

[64] Jared P. Mohr, Juan D. Chavez, and James E. Bruce. Multidimensional Cross-Linking with a Tetra-Reactive Cross-Linker. In *Proceedings of the 67th ASMS Conference on Mass Spectrometry and Allied Topics, June 3-7, 2019*, Atlanta, Georgia, June 2019.

[65] Lukas Käll, Jesse D Canterbury, Jason Weston, William Stafford Noble, and Michael J MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4(11):923–925, November 2007.

[66] D. R. Müller, P. Schindler, H. Towbin, U. Wirth, H. Voshol, S. Hoving, and M. O. Steinmetz. Isotope-Tagged Cross-Linking Reagents. A New Tool in Mass Spectrometric Protein Interaction Analysis. *Analytical Chemistry*, 73(9):1927–1934, May 2001.

[67] Nicholas I. Brodie, Konstantin I. Popov, Evgeniy V. Petrotchenko, Nikolay V. Dokholyan, and Christoph H. Borchers. Solving protein structures using short-distance cross-linking constraints as a guide for discrete molecular dynamics simulations. *Science Advances*, 3(7):e1700479, July 2017.

[68] Nicholas I. Brodie, Konstantin I. Popov, Evgeniy V. Petrotchenko, Nikolay V. Dokholyan, and Christoph H. Borchers. Conformational ensemble of native α-synuclein in solution as determined by short-distance crosslinking constraint-guided discrete molecular dynamics simulations. *PLOS Computational Biology*, 15(3):e1006859, March 2019.

[69] Nikolay V. Dokholyan. Experimentally-driven protein structure modeling. *Journal of Proteomics*, 220:103777, May 2020.

[70] Jason J. Serpa, Konstantin I. Popov, Evgeniy V. Petrotchenko, Nikolay V. Dokholyan, and Christoph H. Borchers. Structure of prion B-oligomers as determined by short-distance crosslinking constraint-guided discrete molecular dynamics simulations. *PROTEOMICS*, 21(21-22):2000298, November 2021.

[71] Marco D Mukrasch, Stefan Bibow, Jegannath Korukottu, Sadasivam Jeganathan, Jacek Biernat, Christian Griesinger, Eckhard Mandelkow, and Markus Zweckstetter. Structural Polymorphism of 441-Residue Tau at Single Residue Resolution. *PLoS Biology*, 7(2):e1000034, February 2009.

[72] Shaolong Zhu, Agnesa Shala, Alexandr Bezginov, Adnan Sljoka, Gerald Audette, and Derek J. Wilson. Hyperphosphorylation of intrinsically disordered tau protein induces an amyloidogenic shift in its conformational ensemble. *PloS One*, 10(3):e0120416, 2015.

[73] Miriam S. Goyder, Fabien Rebeaud, Marc E. Pfeifer, and Franka Kálmán. Strategies in mass spectrometry for the assignment of Cys-Cys disulfide connectivities in proteins. *Expert Review of Proteomics*, 10(5):489–501, October 2013.

[74] Jeffrey J. Gorman, Tristan P. Wallis, and James J. Pitt. Protein disulfide bond determination by mass spectrometry. *Mass Spectrometry Reviews*, 21(3):183–216, May 2002.

[75] Dariusz J. Janecki and Jennifer F. Nemeth. Application of MALDI TOF/TOF mass spectrometry and collision-induced dissociation for the identification of disulfide-bonded peptides. *Journal of Mass Spectrometry*, 46(7):677–688, July 2011.

[76] Stefan Schürch, Johann Schaller, Urs Kämpfer, Lucia Kuhn-Nentwig, and Wolfgang Nentwig. Neurotoxic Peptides in the Multicomponent Venom of the Spider Cupiennius Salei Part II. Elucidation of the Disulphide-Bridge Pattern of the Neurotoxic Peptide CSTX-9 by Tandem Mass Spectrometry. *CHIMIA International Journal for Chemistry*, 55(12):1063–1066, 2001.

[77] Seonhwa Choi, Jaeho Jeong, Seungjin Na, Hyo Sun Lee, Hwa-Young Kim, Kong-Joo Lee, and Eunok Paek. New Algorithm for the Identification of Intact Disulfide Linkages Based on Fragmentation Characteristics in Tandem Mass Spectra. *Journal of Proteome Research*, 9(1):626–635, January 2010.

[78] Evgeniy V. Petrotchenko, Jason J. Serpa, Darryl B. Hardie, Mark Berjanskii, Bow P. Suriyamongkol, David S. Wishart, and Christoph H. Borchers. Use of Proteinase K Nonspecific Digestion for Selective and Comprehensive Identification of Interpeptide Cross-links: Application to Prion Proteins. *Molecular & Cellular Proteomics*, 11(7):M111.013524, July 2012.

[79] Christine C. Wu, Michael J. MacCoss, Kathryn E. Howell, and John R. Yates. A method for the comprehensive proteomic analysis of membrane proteins. *Nature Biotechnology*, 21(5):532–538, May 2003.

[80] J. Hwa, J. Klein-Seetharaman, and H. G. Khorana. Structure and function in rhodopsin: Mass spectrometric identification of the abnormal intradiscal disulfide bond in misfolded retinitis pigmentosa mutants. *Proceedings of the National Academy of Sciences*, 98(9):4872–4876, April 2001.

[81] M. O. Glocker, B. Arbogast, R. Milley, C. Cowgill, and M. L. Deinzer. Disulfide linkages in the in vitro refolded intermediates of recombinant human macrophage-colony-stimulating factor: analysis of the sulfhydryl alkylation of free cysteine residues by fast-atom bombardment mass spectrometry. *Proceedings of the National Academy of Sciences*, 91(13):5868–5872, June 1994.

[82] Evgeniy V. Petrotchenko, Jason J. Serpa, and Christoph H. Borchers. An Isotopically Coded CID-cleavable Biotinylated Cross-linker for Structural Proteomics. *Molecular & Cellular Proteomics*, 10(2):M110.001420, February 2011.

[83] Evgeniy V. Petrotchenko and Christoph H. Borchers. Application of a fast sorting algorithm to the assignment of mass spectrometric cross-linking data. *PROTEOMICS*, 14(17-18):1987–1989, September 2014.

[84] Evgeniy V. Petrotchenko, Karl A. T. Makepeace, Jason J. Serpa, and Christoph H. Borchers. Analysis of protein structure by cross-linking combined with mass spectrometry. In Daniel Martins-de Souza, editor, *Shotgun proteomics: methods and protocols*, pages 447–463. Springer New York, New York, NY, 2014.

[85] V. K. Brown. Essays in Toxicology, Volume 4. *Biochemical Society Transactions*, 2(6):1389–1390, December 1974.

[86] Justin M. Kollman, Leela Pandi, Michael R. Sawaya, Marcia Riley, and Russell F. Doolittle. Crystal Structure of Human Fibrinogen. *Biochemistry*, 48(18):3877–3886, May 2009.

[87] Hongcheng Liu and Kimberly May. Disulfide bond structures of IgG molecules: Structural variations, chemical modifications and possible impacts to stability and biological function. *mAbs*, 4(1):17–23, January 2012.

[88] P. E. Oyer, S. Cho, J. D. Peterson, and D. F. Steiner. Studies on human proinsulin. Isolation and amino acid sequence of the human pancreatic C-peptide. *J. Biol. Chem.*, 246(5):1375–1386, 1971.

[89] R. E. Canfield and A. K. Liu. The Disulfide Bonds Of Egg White Lysozyme (Muramidase). *J. Biol. Chem.*, 240:1997–2002, 1965.

[90] D. G. Smyth, W. H. Stein, and S. Moore. The sequence of amino acid residues in bovine pancreatic ribonuclease: revisions and confirmations. *J. Biol. Chem.*, 238:227–34, 1963.

[91] Costel C. Darie, Martin L. Biniossek, Luca Jovine, Eveline S. Litscher, and Paul M. Wassarman. Structural Characterization of Fish Egg Vitelline Envelope Proteins by Mass Spectrometry †. *Biochemistry*, 43(23):7459–7478, June 2004.

[92] Cindy Beharry, Leah S. Cohen, Jing Di, Kawsar Ibrahim, Susan Briffa-Mirabella, and Alejandra del C. Alonso. Tau-induced neurodegeneration: mechanisms and targets. *Neuroscience Bulletin*, 30(2):346–358, April 2014.

[93] Bess Frost, Jürgen Götz, and Mel B. Feany. Connecting the dots between tau dysfunction and neurodegeneration. *Trends in Cell Biology*, 25(1):46–53, January 2015.

[94] Dominic M. Walsh, Igor Klyubin, Julia V. Fadeeva, William K. Cullen, Roger Anwyl, Michael S. Wolfe, Michael J. Rowan, and Dennis J. Selkoe. Naturally secreted oligomers of amyloid β protein potently inhibit hippocampal long-term potentiation in vivo. *Nature*, 416(6880):535–539, April 2002.

[95] Florence Clavaguera, Jürgen Hench, Michel Goedert, and Markus Tolnay. Invited review: Prion-like transmission and spreading of tau pathology: Prion-like transmission and spreading of tau pathology. *Neuropathology and Applied Neurobiology*, 41(1):47–58, February 2015.

[96] Renaud La Joie, Adrienne V. Visani, Suzanne L. Baker, Jesse A. Brown, Viktoriya Bourakova, Jungho Cha, Kiran Chaudhary, Lauren Edwards, Leonardo Iaccarino, Mustafa Janabi, Orit H. Lesman-Segev, Zachary A. Miller, David C. Perry, James P. O'Neil, Julie Pham, Julio C. Rojas, Howard J. Rosen, William W. Seeley, Richard M. Tsai, Bruce L. Miller, William J. Jagust, and Gil D. Rabinovici. Prospective longitudinal atrophy in Alzheimer's disease correlates with the intensity and topography of baseline tau-PET. *Science Translational Medicine*, 12(524):eaau5732, January 2020.

[97] Gayathri Ramachandran and Jayant B. Udgaonkar. Mechanistic Studies Unravel the Complexity Inherent in Tau Aggregation Leading to Alzheimer's Disease and the Tauopathies. *Biochemistry*, 52(24):4107–4126, June 2013.

[98] Anthony W. P. Fitzpatrick, Benjamin Falcon, Shaoda He, Alexey G. Murzin, Garib Murshudov, Holly J. Garringer, R. Anthony Crowther, Bernardino Ghetti, Michel Goedert, and Sjors H. W. Scheres. Cryo-EM structures of tau filaments from Alzheimer's disease. *Nature*, 547(7662):185–190, July 2017.

[99] Benjamin Falcon, Wenjuan Zhang, Alexey G. Murzin, Garib Murshudov, Holly J. Garringer, Ruben Vidal, R. Anthony Crowther, Bernardino Ghetti, Sjors H. W. Scheres, and Michel Goedert. Structures of filaments from Pick's disease reveal a novel tau protein fold. *Nature*, 561(7721):137–140, September 2018.

[100] Sjors HW Scheres, Wenjuan Zhang, Benjamin Falcon, and Michel Goedert. Cryo-EM structures of tau filaments. *Current Opinion in Structural Biology*, 64:17–25, October 2020.

[101] Sofia Lövestam, Fujiet Adrian Koh, Bart van Knippenberg, Abhay Kotecha, Alexey G Murzin, Michel Goedert, and Sjors HW Scheres. Assembly of recombinant tau into filaments identical to those of Alzheimer's disease and chronic traumatic encephalopathy. *eLife*, 11:e76494, March 2022.

[102] Abhinav Nath, Maria Sammalkorpi, David C. DeWitt, Adam J. Trexler, Shana Elbaum-Garfinkle, Corey S. O'Hern, and Elizabeth Rhoades. The Conformational Ensembles of α-Synuclein and Tau: Combining Single-Molecule FRET and Simulations. *Biophysical Journal*, 103(9):1940–1949, November 2012.

[103] Sadasivam Jeganathan, Martin von Bergen, Henrik Brutlach, Heinz-Jürgen Steinhoff, and Eckhard Mandelkow. Global Hairpin Folding of Tau in Solution. *Biochemistry*, 45(7):2283–2293, February 2006.

[104] Efstratios Mylonas, Antje Hascher, Pau Bernadó, Martin Blackledge, Eckhard Mandelkow, and Dmitri I. Svergun. Domain Conformation of Tau Protein Studied by Solution Small-Angle X-ray Scattering †. *Biochemistry*, 47(39):10345–10353, September 2008.

[105] Nicole L. Zabik, Matthew M. Imhof, and Sanela Martic-Milne. Structural evaluations of tau protein conformation: methodologies and approaches. *Biochemistry and Cell Biology*, 95(3):338–349, June 2017.

[106] R. Kaptein, E.R.P. Zuiderweg, R.M. Scheek, R. Boelens, and W.F. van Gunsteren. A protein structure from nuclear magnetic resonance data. *Journal of Molecular Biology*, 182(1):179–182, March 1985.

[107] Andrew E. Torda, Ruud M. Scheek, and Wilfred F. van Gunsteren. Time-averaged nuclear overhauser effect distance restraints applied to tendamistat. *Journal of Molecular Biology*, 214(1):223–235, July 1990.

[108] Emanuel Peter and Jiří Černý. A Hybrid Hamiltonian for the Accelerated Sampling along Experimental Restraints. *International Journal of Molecular Sciences*, 20(2):370, January 2019.

[109] Yiwen Chen, Feng Ding, and Nikolay V. Dokholyan. Fidelity of the Protein Structure Reconstruction from Inter-Residue Proximity Constraints. *The Journal of Physical Chemistry B*, 111(25):7432–7438, June 2007.

[110] Massimiliano Bonomi, Gabriella T. Heller, Carlo Camilloni, and Michele Vendruscolo. Principles of protein structural ensemble determination. *Current Opinion in Structural Biology*, 42:106–116, February 2017.

[111] Feng Ding and Nikolay V. Dokholyan. Emergence of Protein Fold Families through Rational Design. *PLoS Computational Biology*, 2(7):e85, 2006.

[112] Feng Ding, Douglas Tsao, Huifen Nie, and Nikolay V. Dokholyan. Ab Initio Folding of Proteins with All-Atom Discrete Molecular Dynamics. *Structure*, 16(7):1010–1018, July 2008.

[113] David Shirvanyants, Feng Ding, Douglas Tsao, Srinivas Ramachandran, and Nikolay V. Dokholyan. Discrete Molecular Dynamics: An Efficient And Versatile Simulation Method For Fine Protein Characterization. *The Journal of Physical Chemistry B*, 116(29):8375–8382, July 2012.

[114] Elizabeth A. Proctor, Feng Ding, and Nikolay V. Dokholyan. Discrete Molecular Dynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(1):80–92, January 2011.

[115] Feng Ding, Yoshiaki Furukawa, Nobuyuki Nukina, and Nikolay V. Dokholyan. Local Unfolding of Cu, Zn Superoxide Dismutase Monomer Determines the Morphology of Fibrillar Aggregates. *Journal of Molecular Biology*, 421(4-5):548–560, August 2012.

[116] Richard D.S. Dixon, Yiwen Chen, Feng Ding, Sagar D. Khare, Kirk C. Prutzman, Michael D. Schaller, Sharon L. Campbell, and Nikolay V. Dokholyan. New Insights into FAK Signaling and Localization Based on Detection of a FAT Domain Folding Intermediate. *Structure*, 12(12):2161–2171, December 2004.

[117] Dániel Szöllősi, Tamás Horváth, Kyou-Hoon Han, Nikolay V. Dokholyan, Péter Tompa, Lajos Kalmár, and Tamás Hegedűs. Discrete Molecular Dynamics Can Predict Helical Prestructured Motifs in Disordered Proteins. *PLoS ONE*, 9(4):e95795, April 2014.

[118] Andrej Shevchenko, Henrik Tomas, Jan Havli, Jesper V Olsen, and Matthias Mann. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nature Protocols*, 1(6):2856–2860, December 2006.

[119] Yuko Okamoto. Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. *Journal of Molecular Graphics and Modelling*, 22(5):425–439, May 2004.

[120] Shuangye Yin, Lada Biedermannova, Jiri Vondrasek, and Nikolay V. Dokholyan. MedusaScore: An Accurate Force Field-Based Scoring Function for Virtual Drug Screening. *Journal of Chemical Information and Modeling*, 48(8):1656–1662, August 2008.

[121] Sander Pronk, Szilárd Páll, Roland Schulz, Per Larsson, Pär Bjelkmar, Rossen Apostolov, Michael R. Shirts, Jeremy C. Smith, Peter M. Kasson, David van der Spoel, Berk Hess, and Erik Lindahl. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, April 2013.

[122] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, March 2008.

[123] Robert B. Best, Xiao Zhu, Jihyun Shim, Pedro E. M. Lopes, Jeetain Mittal, Michael Feig, and Alexander D. MacKerell. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ_1 and χ_2 Dihedral Angles. *Journal of Chemical Theory and Computation*, 8(9):3257–3273, September 2012.

[124] Brinda Vallat, Benjamin Webb, John D. Westbrook, Andrej Sali, and Helen M. Berman. Development of a Prototype System for Archiving Integrative/Hybrid Structure Models of Biological Macromolecules. *Structure*, 26(6):894–904.e2, June 2018.

[125] Brinda Vallat, Benjamin Webb, Maryam Fayazi, Serban Voinea, Hongsuda Tangmunarunkit, Sai J. Ganesan, Catherine L. Lawson, John D. Westbrook, Carl Kesselman, Andrej Sali, and Helen M. Berman. New system for archiving integrative structures. *Acta Crystallographica Section D Structural Biology*, 77(12):1486–1496, December 2021.

[126] Eric W. Deutsch, Attila Csordas, Zhi Sun, Andrew Jarnuczak, Yasset Perez-Riverol, Tobias Ternent, David S. Campbell, Manuel Bernal-Llinares, Shujiro Okuda, Shin Kawano, Robert L. Moritz, Jeremy J. Carver, Mingxun Wang, Yasushi Ishihama, Nuno Bandeira, Henning Hermjakob, and Juan Antonio Vizcaíno. The ProteomeXchange consortium in 2017: Supporting the cultural change in proteomics public data deposition. *Nucleic Acids Research*, 45(D1):D1100–D1106, January 2017.

[127] Yasset Perez-Riverol, Attila Csordas, Jingwen Bai, Manuel Bernal-Llinares, Suresh Hewapathirana, Deepti J Kundu, Avinash Inuganti, Johannes Griss, Gerhard Mayer, Martin Eisenacher, Enrique Pérez, Julian Uszkoreit, Julianus Pfeuffer, Timo Sachsenberg, Şule Yılmaz, Shivani Tiwary, Jürgen Cox, Enrique Audain, Mathias Walzer, Andrew F Jarnuczak, Tobias Ternent, Alvis Brazma, and Juan Antonio Vizcaíno. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Research*, 47(D1):D442–D450, January 2019.

[128] Yasset Perez-Riverol, Jingwen Bai, Chakradhar Bandla, David García-Seisdedos, Suresh Hewapathirana, Selvakumar Kamatchinathan, Deepti J Kundu, Ananth Prakash, Anika Frericks-Zipper, Martin Eisenacher, Mathias Walzer, Shengbo Wang, Alvis Brazma, and Juan Antonio Vizcaíno. The PRIDE database resources in 2022: A hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research*, 50(D1):D543–D552, January 2022.

[129] Nicholas I. Brodie, Karl A.T. Makepeace, Evgeniy V. Petrotchenko, and Christoph H. Borchers. Isotopically-coded short-range hetero-bifunctional photo-reactive crosslinkers for studying protein structure. *Journal of Proteomics*, 118:12–20, April 2015.

[130] Hilda Mirbaha, Dailu Chen, Olga A Morazova, Kiersten M Ruff, Apurwa M Sharma, Xiaohua Liu, Mohammad Goodarzi, Rohit V Pappu, David W Colby, Hamid Mirzaei, Lukasz A Joachimiak, and Marc I Diamond. Inert and seed-competent tau monomers suggest structural origins of aggregation. *eLife*, 7:e36584, July 2018.

[131] Kris Pauwels, Pierre Lebrun, and Peter Tompa. To be disordered or not to be disordered: is that still a question for proteins in the cell? *Cellular and Molecular Life Sciences*, 74(17):3185–3204, September 2017.

[132] András Micsonai, Frank Wien, Éva Bulyáki, Judit Kun, Éva Moussong, Young-Ho Lee, Yuji Goto, Matthieu Réfrégiers, and József Kardos. BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra. *Nucleic Acids Research*, 46(W1):W315–W322, July 2018.

[133] Daniel S. Ziemianowicz, Ryan Bomgarden, Chris Etienne, and David C. Schriemer. Amino Acid Insertion Frequencies Arising from Photoproducts Generated Using Aliphatic Diazirines. *Journal of The American Society for Mass Spectrometry*, 28(10):2011–2021, October 2017.

[134] Feng Ding, Ramesh K. Jha, and Nikolay V. Dokholyan. Scaling Behavior and Structure of Denatured Proteins. *Structure*, 13(7):1047–1054, July 2005.

[135] Waltraud Mair, Jan Muntel, Katharina Tepper, Shaojun Tang, Jacek Biernat, William W. Seeley, Kenneth S. Kosik, Eckhard Mandelkow, Hanno Steen, and Judith A. Steen. FLEXITau: Quantifying Post-translational Modifications of Tau Protein in Vitro and in Human Disease. *Analytical Chemistry*, 88(7):3704–3714, April 2016.

[136] Carol V. Robinson, Andrej Sali, and Wolfgang Baumeister. The molecular sociology of the cell. *Nature*, 450(7172):973–982, 2007.

[137] Andrej Sali, Robert Glaeser, Thomas Earnest, and Wolfgang Baumeister. From words to literature in structural proteomics. *Nature*, 422(6928):216–225, 2003.

[138] C. Plaschka, L. Larivière, L. Wenzeck, M. Seizl, M. Hemann, D. Tegunov, E. V. Petrotchenko, C. H. Borchers, W. Baumeister, F. Herzog, E. Villa, and P. Cramer. Architecture of the RNA polymerase II–Mediator core initiation complex. *Nature*, 518(7539):376–380, 2015.

[139] Robyn M. Kaake, Xiaorong Wang, Anthony Burke, Clinton Yu, Wynne Kandur, Yingying Yang, Eric J. Novtisky, Tonya Second, Jicheng Duan, Athit Kao, Shenheng Guan, Danielle Vellucci, Scott D. Rychnovsky, and Lan Huang. A New in vivo Cross-linking Mass Spectrometry Platform to Define Protein–Protein Interactions in Living Cells. *Molecular & Cellular Proteomics*, 13(12):3533–3543, 2014.

[140] Clinton Yu and Lan Huang. Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology. *Analytical Chemistry*, 90(1):144–165, 2018.

[141] Andrea Sinz. Investigation of protein–protein interactions in living cells by chemical crosslinking and mass spectrometry. *Analytical and Bioanalytical Chemistry*, 397(8):3433–3440, 2010.

[142] Alexander Leitner, Marco Faini, Florian Stengel, and Ruedi Aebersold. Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. *Trends in Biochemical Sciences*, 41(1):20–32, 2016.

[143] Evgeniy V. Petrotchenko and Christoph H. Borchers. Crosslinking combined with mass spectrometry for structural proteomics. *Mass Spectrometry Reviews*, 29(6):862–876, 2010.

[144] James E. Bruce. In vivo protein complex topologies: Sights through a cross-linking lens. *PROTEOMICS*, 12(10):1565–1575, 2012.

[145] Juri Rappsilber. The beginning of a beautiful friendship: Cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *Journal of Structural Biology*, 173(3):530–540, 2011.

[146] Fan Liu and Albert JR Heck. Interrogating the architecture of protein assemblies and protein interaction networks by cross-linking mass spectrometry. *Current Opinion in Structural Biology*, 35:100–108, 2015.

[147] Marco Y. Hein, Nina C. Hubner, Ina Poser, Jürgen Cox, Nagarjuna Nagaraj, Yusuke Toyoda, Igor A. Gak, Ina Weisswange, Jörg Mansfeld, Frank Buchholz, Anthony A. Hyman, and Matthias Mann. A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell*, 163(3):712–723, 2015.

[148] Erik J. Soderblom and Michael B. Goshe. Collision-Induced Dissociative Chemical Cross-Linking Reagents and Methodology: Applications to Protein Structural Characterization Using Tandem Mass Spectrometry Analysis. *Analytical Chemistry*, 78(23):8059–8068, 2006.

[149] Frank Dreiocker, Mathias Q. Müller, Andrea Sinz, and Mathias Schäfer. Collision-induced dissociative chemical cross-linking reagent for protein structure characterization: Applied Edman chemistry in the gas phase. *Journal of Mass Spectrometry*, 45(2):178–189, 2010.

[150] Mathias Q. Müller, Frank Dreiocker, Christian H. Ihling, Mathias Schäfer, and Andrea Sinz. Cleavable Cross-Linker for Protein Structure Analysis: Reliable Identification of Cross-Linking Products by Tandem MS. *Analytical Chemistry*, 82(16):6958–6968, 2010.

[151] Athit Kao, Chi-li Chiu, Danielle Vellucci, Yingying Yang, Vishal R. Patel, Shenheng Guan, Arlo Randall, Pierre Baldi, Scott D. Rychnovsky, and Lan Huang. Development of a Novel Cross-linking Strategy for Fast and Accurate Identification of Cross-linked Peptides of Protein Complexes. *Molecular & Cellular Proteomics*, 10(1):M110.002212, 2011.

[152] Andrea Sinz. Divide and conquer: Cleavable cross-linkers to study protein conformation and protein–protein interactions. *Analytical and Bioanalytical Chemistry*, 409(1):33–44, 2017.

[153] Devin K. Schweppe, Juan D. Chavez, Chi Fung Lee, Arianne Caudal, Shane E. Kruse, Rudy Stuppard, David J. Marcinek, Gerald S. Shadel, Rong Tian, and James E. Bruce. Mitochondrial protein interactome elucidated by chemical cross-linking mass spectrometry. *Proceedings of the National Academy of Sciences*, 114(7):1732–1737, 2017.

[154] Fan Liu, Philip Lössl, Beverley M. Rabbitts, Robert S. Balaban, and Albert J. R. Heck. The interactome of intact mitochondria by cross-linking mass spectrometry provides evidence for coexisting respiratory supercomplexes. *Molecular & Cellular Proteomics*, 17(2):216–232, 2018.

[155] Chris Meisinger, Nikolaus Pfanner, and Kaye N. Truscott. *Isolation of Yeast Mitochondria*, volume 313, pages 033–040. Humana Press, 2005.

[156] A. Sickmann, J. Reinders, Y. Wagner, C. Joppich, R. Zahedi, H. E. Meyer, B. Schonfisch, I. Perschil, A. Chacinska, B. Guiard, P. Rehling, N. Pfanner, and C. Meisinger. The proteome of Saccharomyces cerevisiae mitochondria. *Proceedings of the National Academy of Sciences*, 100(23):13207–13212, 2003.

[157] Jacek R Wiśniewski, Alexandre Zougman, Nagarjuna Nagaraj, and Matthias Mann. Universal sample preparation method for proteome analysis. *Nature Methods*, 6(5):359–362, 2009.

[158] Michael R. Hoopmann, Gregory L. Finney, and Michael J. MacCoss. High-Speed Data Reduction, Feature Detection, and MS/MS Spectrum Quality Assessment of Shotgun Proteomics Data Sets Using High-Resolution Mass Spectrometry. *Analytical Chemistry*, 79(15):5620–5632, 2007.

[159] Michael R. Hoopmann, Michael J. MacCoss, and Robert L. Moritz. Identification of Peptide Features in Precursor Spectra Using Hardklör and Krönik. In Andreas D. Baxevanis, Gregory A. Petsko, Lincoln D. Stein, and Gary D. Stormo, editors, *Current Protocols in Bioinformatics*, page bi1318s37. John Wiley & Sons, Inc., 2012.

[160] Darren Kessner, Matt Chambers, Robert Burke, David Agus, and Parag Mallick. ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics*, 24(21):2534–2536, 2008.

[161] F.-Nora Vögtle, Julia M. Burkhart, Humberto Gonczarowska-Jorge, Cansu Kücükköse, Asli Aras Taskin, Dominik Kopczynski, Robert Ahrends, Dirk Mossmann, Albert Sickmann, René P. Zahedi, and Chris Meisinger. Landscape of submitochondrial protein distribution. *Nature Communications*, 8(1):290, 2017.

[162] Karl A. T. Makepeace, Yassene Mohammed, Elena L. Rudashevskaya, Evgeniy V. Petrotchenko, F.-Nora Vögtle, Chris Meisinger, Albert Sickmann, and Christoph H. Borchers. Improving Identification of In-organello Protein-Protein Interactions Using an Affinity-enrichable, Isotopically Coded, and Mass Spectrometry-cleavable Chemical Crosslinker. *Molecular & Cellular Proteomics*, 19(4):624–639, April 2020.

[163] Martin J. Graham, Colin Combe, Lars Kolbowski, and Juri Rappsilber. xiView: A common platform for the downstream analysis of Crosslinking Mass Spectrometry data. *bioRxiv*, 2019.

[164] Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8, 2015.

[165] Romy Fritzsche, Christian H. Ihling, Michael Götze, and Andrea Sinz. Optimizing the enrichment of cross-linked products for mass spectrometric protein analysis: Enrichment of cross-linked products for MS protein analysis. *Rapid Communications in Mass Spectrometry*, 26(6):653–658, 2012.

[166] Joerg Reinders, René P. Zahedi, Nikolaus Pfanner, Chris Meisinger, and Albert Sickmann. Toward the Complete Yeast Mitochondrial Proteome: Multidimensional Separation Techniques for Mitochondrial Proteomics. *Journal of Proteome Research*, 5(7):1543–1554, 2006.

[167] Young M. Park, Silvano Squizzato, Nicola Buso, Tamer Gur, and Rodrigo Lopez. The EBI search engine: EBI search as a service—making biological data accessible for all. *Nucleic Acids Research*, 45(W1):W545–W549, 2017.

[168] Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H. Campbell, Gayatri Chavali, Carol Chen, Noemi del Toro, Margaret Duesbury, Marine Dumousseau, Eugenia Galeota, Ursula Hinz, Marta Iannuccelli, Sruthi Jagannathan, Rafael Jimenez, Jyoti Khadake, Astrid Lagreid, Luana Licata, Ruth C. Lovering, Birgit Meldal, Anna N. Melidoni, Mila Milagros, Daniele Peluso, Livia Perfetto, Pablo Porras, Arathi Raghunath, Sylvie Ricard-Blum, Bernd Roechert, Andre Stutz, Michael Tognolli, Kim van Roey, Gianni Cesareni, and Henning Hermjakob. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(D1):D358–D363, 2014.

[169] H. Hermjakob. IntAct: An open source molecular interaction database. *Nucleic Acids Research*, 32(90001):452D–455, 2004.

[170] Claudio Anselmi, Karen M. Davies, and José D. Faraldo-Gómez. Mitochondrial ATP synthase dimers spontaneously associate due to a long-range membrane-induced force. *The Journal of General Physiology*, 150(5):763–770, 2018.

[171] J. Michael Cherry, Catherine Ball, Shuai Weng, Gail Juvik, Rita Schmidt, Caroline Adler, Barbara Dunn, Selina Dwight, Linda Riles, Robert K. Mortimer, and David Botstein. Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, 387(S6632):67–73, 1997.

[172] K. M. Davies, C. Anselmi, I. Wittig, J. D. Faraldo-Gomez, and W. Kuhlbrandt. Structure of the yeast F1Fo-ATP synthase dimer and its role in shaping the mitochondrial cristae. *Proceedings of the National Academy of Sciences*, 109(34):13602–13607, 2012.

[173] Fan Liu, Bas van Breukelen, and Albert J.R. Heck. Facilitating Protein Disulfide Mapping by a Combination of Pepsin Digestion, Electron Transfer Higher Energy Dissociation (EThcD), and a Dedicated Search Algorithm SlinkS. *Molecular & Cellular Proteomics*, 13(10):2776–2786, October 2014.

[174] Michael B. Cammarata, Luis A. Macias, Jake Rosenberg, Alexander Bolufer, and Jennifer S. Brodbelt. Expanding the Scope of Cross-Link Identifications by Incorporating Collisional Activated Dissociation and Ultraviolet Photodissociation Methods. *Analytical Chemistry*, 90(11):6385–6389, June 2018.

[175] Adam Belsom, Michael Schneider, Lutz Fischer, Oliver Brock, and Juri Rappsilber. Serum Albumin Domain Structures in Human Blood Serum by Mass Spectrometry and Computational Biology. *Molecular & Cellular Proteomics*, 15(3):1105–1116, March 2016.

[176] Michael Schneider, Adam Belsom, and Juri Rappsilber. Protein Tertiary Structure by Crosslinking/Mass Spectrometry. *Trends in Biochemical Sciences*, 43(3):157–169, March 2018.

[177] Florian Meier, Andreas-David Brunner, Scarlet Koch, Heiner Koch, Markus Lubeck, Michael Krause, Niels Goedecke, Jens Decker, Thomas Kosinski, Melvin A. Park, Nicolai Bache, Ole Hoerning, Jürgen Cox, Oliver Räther, and Matthias Mann. Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. *Molecular & Cellular Proteomics*, 17(12):2534–2545, December 2018.

[178] Christian H. Ihling, Lolita Piersimoni, Marc Kipping, and Andrea Sinz. Cross-Linking/Mass Spectrometry Combined with Ion Mobility on a timsTOF Pro Instrument for Structural Proteomics. *Analytical Chemistry*, 93(33):11442–11450, August 2021.

[179] Filipa Teixeira, Eric Tse, Helena Castro, Karl A. T. Makepeace, Ben A. Meinen, Christoph H. Borchers, Leslie B. Poole, James C. Bardwell, Ana M. Tomás, Daniel R. Southworth, and Ursula Jakob. Chaperone activation and client binding of a 2-cysteine peroxiredoxin. *Nature Communications*, 10(1):659, December 2019.

[180] Karl A.T. Makepeace, Nicholas I. Brodie, Konstantin I. Popov, Geoff Gudavicius, Christopher J. Nelson, Evgeniy V. Petrotchenko, Nikolay V. Dokholyan, and Christoph H. Borchers. Ligand-induced disorder-to-order transitions characterized by structural proteomics and molecular dynamics simulations. *Journal of Proteomics*, 211:103544, January 2020.

[181] Florian Meier, Philipp E. Geyer, Sebastian Virreira Winter, Juergen Cox, and Matthias Mann. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nature Methods*, 15(6):440–448, June 2018.

[182] Conor Jenkins and Ben Orsburn. BoxCar Assisted MS Fragmentation (BAMF). Preprint, Biochemistry, December 2019.

[183] Lutz Fischer and Juri Rappsilber. Quirks of Error Estimation in Cross-Linking/Mass Spectrometry. *Analytical Chemistry*, 89(7):3829–3833, April 2017.

[184] Lolita Piersimoni, Panagiotis L. Kastritis, Christian Arlt, and Andrea Sinz. Cross-Linking Mass Spectrometry for Investigating Protein Conformations and Protein–Protein Interactions-A Method for All Seasons. *Chemical Reviews*, 122(8):7500–7531, April 2022.

[185] Lutz Fischer, Zhuo Angel Chen, and Juri Rappsilber. Quantitative cross-linking/mass spectrometry using isotope-labelled cross-linkers. *Journal of Proteomics*, 88:120–128, August 2013.

[186] Thomas Walzthoeni, Lukasz A Joachimiak, George Rosenberger, Hannes L Röst, Lars Malmström, Alexander Leitner, Judith Frydman, and Ruedi Aebersold. xTract: Software for characterizing conformational changes of protein complexes by quantitative cross-linking mass spectrometry. *Nature Methods*, 12(12):1185–1190, December 2015.

[187] Brendan MacLean, Daniela M. Tomazela, Nicholas Shulman, Matthew Chambers, Gregory L. Finney, Barbara Frewen, Randall Kern, David L. Tabb, Daniel C. Liebler, and Michael J. MacCoss. Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*, 26(7):966–968, April 2010.

[188] Jürgen Cox and Matthias Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372, December 2008.

[189] Zhuo A. Chen, Lutz Fischer, Jürgen Cox, and Juri Rappsilber. Quantitative Cross-linking/Mass Spectrometry Using Isotope-labeled Cross-linkers and MaxQuant. *Molecular & Cellular Proteomics*, 15(8):2769–2778, August 2016.

[190] Claudio Iacobucci, Christine Piotrowski, Ruedi Aebersold, Bruno C. Amaral, Philip Andrews, Katja Bernfur, Christoph Borchers, Nicolas I. Brodie, James E. Bruce, Yong Cao, Stéphane Chaignepain, Juan D. Chavez, Stéphane Claverol, Jürgen Cox, Trisha Davis, Gianluca Degliesposti, Meng-Qiu Dong, Nufar Edinger, Cecilia Emanuelsson, Marina Gay, Michael Götze, Francisco Gomes-Neto, Fabio C. Gozzo, Craig Gutierrez, Caroline Haupt, Albert J. R. Heck, Franz Herzog, Lan Huang, Michael R. Hoopmann, Nir Kalisman, Oleg Klykov, Zdeněk Kukačka, Fan Liu, Michael J. MacCoss, Karl Mechtler, Ravit Mesika, Robert L. Moritz, Nagarjuna Nagaraj, Victor Nesati, Ana G. C. Neves-Ferreira, Robert Ninnis, Petr Novák, Francis J. O'Reilly, Matthias Pelzing, Evgeniy Petrotchenko, Lolita Piersimoni, Manolo Plasencia, Tara Pukala, Kasper D. Rand, Juri Rappsilber, Dana Reichmann, Carolin Sailer, Chris P. Sarnowski, Richard A. Scheltema, Carla Schmidt, David C. Schriemer, Yi Shi, J. Mark Skehel, Moriya Slavin, Frank Sobott, Victor Solis-Mezarino, Heike Stephanowitz, Florian Stengel, Christian E. Stieger, Esben Trabjerg, Michael Trnka, Marta Vilaseca, Rosa Viner, Yufei Xiang, Sule Yilmaz, Alex Zelter, Daniel Ziemianowicz, Alexander Leitner, and Andrea Sinz. First Community-Wide, Comparative Cross-Linking Mass Spectrometry Study. *Analytical Chemistry*, 91(11):6953–6961, June 2019.

[191] Andrej Sali, Helen M. Berman, Torsten Schwede, Jill Trewhella, Gerard Kleywegt, Stephen K. Burley, John Markley, Haruki Nakamura, Paul Adams, Alexandre M.J.J. Bonvin, Wah Chiu, Matteo Dal Peraro, Frank Di Maio, Thomas E. Ferrin, Kay Grünewald, Aleksandras Gutmanas, Richard Henderson, Gerhard Hummer, Kenji Iwasaki, Graham Johnson, Catherine L. Lawson,

Jens Meiler, Marc A. Marti-Renom, Gaetano T. Montelione, Michael Nilges, Ruth Nussinov, Ardan Patwardhan, Juri Rappsilber, Randy J. Read, Helen Saibil, Gunnar F. Schröder, Charles D. Schwieters, Claus A.M. Seidel, Dmitri Svergun, Maya Topf, Eldon L. Ulrich, Sameer Velankar, and John D. Westbrook. Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure*, 23(7):1156–1167, July 2015.

[192] Helen M. Berman, Paul D. Adams, Alexandre A. Bonvin, Stephen K. Burley, Bridget Carragher, Wah Chiu, Frank DiMaio, Thomas E. Ferrin, Margaret J. Gabanyi, Thomas D. Goddard, Patrick R. Griffin, Juergen Haas, Christian A. Hanke, Jeffrey C. Hoch, Gerhard Hummer, Genji Kurisu, Catherine L. Lawson, Alexander Leitner, John L. Markley, Jens Meiler, Gaetano T. Montelione, George N. Phillips, Thomas Prisner, Juri Rappsilber, David C. Schriemer, Torsten Schwede, Claus A.M. Seidel, Timothy S. Strutzenberg, Dmitri I. Svergun, Emad Tajkhorshid, Jill Trewhella, Brinda Vallat, Sameer Velankar, Geerten W. Vuister, Benjamin Webb, John D. Westbrook, Kate L. White, and Andrej Sali. Federating Structural Models and Data: Outcomes from A Workshop on Archiving Integrative Structures. *Structure*, 27(12):1745–1759, December 2019.

[193] Juan D. Chavez, Andrew Keller, Bo Zhou, Rong Tian, and James E. Bruce. Cellular Interactome Dynamics during Paclitaxel Treatment. *Cell Reports*, 29(8):2371–2383.e5, November 2019.

[194] David F. Burke, Patrick Bryant, Inigo Barrio-Hernandez, Danish Memon, Gabriele Pozzati, Aditi Shenoy, Wensi Zhu, Alistair S Dunham, Pascal Albanese, Andrew Keller, Richard A. Scheltema, James E. Bruce, Alexander Leitner, Petras Kundrotas, Pedro Beltrao, and Arne Elofsson. Towards a structurally resolved human protein interaction network. Preprint, Bioinformatics, November 2021.

[195] Helisa H. Wippel, Juan D. Chavez, Andrew D. Keller, and James E. Bruce. Multiplexed Isobaric Quantitative Cross-Linking Reveals Drug-Induced Interactome Changes in Breast Cancer Cells. *Analytical Chemistry*, 94(6):2713–2722, February 2022.